Computational Methods of Information Geometry with Real-Time Applications in Audio Signal Processing

Arnaud Dessein

THÈSE DE DOCTORAT DE
L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École Doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

M. Arnaud DESSEIN

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Méthodes Computationnelles en Géométrie de l'Information et Applications Temps Réel au Traitement du Signal Audio

Computational Methods of Information Geometry with Real-Time Applications in Audio Signal Processing

soutenue le 13 décembre 2012

devant le jury composé de :

M. Gérard Assayag	Directeur
M. Arshia Cont	Encadrant
M. Francis Bach	Rapporteur
M. Frank Nielsen	Rapporteur
M. Roland Badeau	Examinateur
M. Silvère Bonnabel	Examinateur
M. Jean-Luc Zarader	Examinateur

Thèse préparée dans l'Équipe-Projet MuTant (INRIA)
Accueillie par l'Équipe Représentations Musicales
UMR 9912 STMS (IRCAM, CNRS, UPMC)
1, place Igor Stravinsky
75004 Paris, France

Et financée principalement par l'UPMC (EDITE)
Avec le soutien d'INRIA (Paris-Rocquencourt)
Ainsi que du CNRS (JST-CNRS ICT)

À Sylvie et Hubert, mes parents, pour tout ce qu'ils m'ont donné.
À Ève, ma compagne, pour son soutien, sa patience, et son amour.

Résumé

Cette thèse propose des méthodes computationnelles nouvelles en géométrie de l'information, avec des applications temps réel au traitement du signal audio. Dans ce contexte, nous traitons en parallèle les problèmes applicatifs de la segmentation audio en temps réel, et de la transcription de musique polyphonique en temps réel. Nous abordons ces applications par le développement respectif de cadres théoriques pour la détection séquentielle de ruptures dans les familles exponentielles, et pour la factorisation en matrices non négatives avec des divergences convexes-concaves. D'une part, la détection séquentielle de ruptures est étudiée par l'intermédiaire de la géométrie de l'information dualement plate liée aux familles exponentielles. Nous développons notamment un cadre statistique générique et unificateur, reposant sur des tests d'hypothèses multiples à l'aide de rapports de vraisemblance généralisés exacts. Nous appliquons ce cadre à la conception d'un système modulaire pour la segmentation audio temps réel avec des types de signaux et de critères d'homogénéité arbitraires. Le système proposé contrôle le flux d'information audio au fur et à mesure qu'il se déroule dans le temps pour détecter des changements. D'autre part, nous étudions la factorisation en matrices non négatives avec des divergences convexes-concaves sur l'espace des mesures discrètes positives. En particulier, nous formulons un cadre d'optimisation générique et unificateur pour la factorisation en matrices non négatives, utilisant des bornes variationnelles par le biais de fonctions auxiliaires. Nous mettons ce cadre à profit en concevant un système temps réel de transcription de musique polyphonique avec un contrôle explicite du compromis fréquentiel pendant l'analyse. Le système développé décompose le signal musical arrivant au cours du temps sur un dictionnaire de modèles spectraux de notes. Ces contributions apportent des pistes de réflexion et des perspectives de recherche intéressantes dans le domaine du traitement du signal audio, et plus généralement de l'apprentissage automatique et du traitement du signal, dans le champ relativement jeune mais néanmoins fécond de la géométrie de l'information computationnelle.

Mots-clés : méthodes computationnelles, géométrie de l'information, applications temps réel, traitement du signal audio, détection de ruptures, familles exponentielles, factorisation en matrices non négatives, divergences convexes-concaves, segmentation audio, transcription de musique polyphonique.

Abstract

This thesis proposes novel computational methods of information geometry with real-time applications in audio signal processing. In this context, we address in parallel the applicative problems of real-time audio segmentation, and of real-time polyphonic music transcription. This is achieved by developing theoretical frameworks respectively for sequential change detection with exponential families, and for non-negative matrix factorization with convex-concave divergences. On the one hand, sequential change detection is studied in the light of the dually flat information geometry of exponential families. We notably develop a generic and unifying statistical framework relying on multiple hypothesis testing with decision rules based on exact generalized likelihood ratios. This is applied to devise a modular system for real-time audio segmentation with arbitrary types of signals and of homogeneity criteria. The proposed system controls the information rate of the audio stream as it unfolds in time to detect changes. On the other hand, non-negative matrix factorization is investigated by the way of convex-concave divergences on the space of discrete positive measures. In particular, we formulate a generic and unifying optimization framework for non-negative matrix factorization based on variational bounding with auxiliary functions. This is employed to design a real-time system for polyphonic music transcription with an explicit control on the frequency compromise during the analysis. The developed system decomposes the music signal as it arrives in time onto a dictionary of note spectral templates. These contributions provide interesting insights and directions for future research in the realm of audio signal processing, and more generally of machine learning and signal processing, in the relatively young but nonetheless prolific field of computational information geometry.

Keywords: computational methods, information geometry, real-time applications, audio signal processing, change detection, exponential families, non-negative matrix factorization, convex-concave divergences, audio segmentation, polyphonic music transcription.

Contents

List of Tables

List of Figures

Preface

It was September 2009 when I undertook the present work as a contract doctoral student at the Université Pierre et Marie Curie (UPMC), within the Institut de Recherche et Coordination Acoustique/Musique (IRCAM) in Paris. I entered there for my very first steps into scientific research in May 2008, when I joined the Sound Perception and Design Team to work on the recognition and identification of environmental sound events from a perceptual viewpoint. I am indebted to Guillaume Lemaitre who proposed this project at that time and introduced me to the world of research. Since then, I have known I was made for research, and I have started to train and gain experience in the domain.

I followed the Master's program Acoustique, Traitement du signal, et Informatique, Appliqués à la Musique (ATIAM), in collaboration between the IRCAM and the UPMC. I acknowledge all teachers and students I met during the program, and in particular my friends, Carlos Agon, Philippe Esling, Benoit Fuentes, and Manuel Moussallam, for what has been a salutary experience where I learned again how to enjoy learning after almost three years of resignation into intellectual idleness. I also warmly thank Arshia Cont and Guillaume Lemaitre for supervising me during the Master's thesis on the automatic recognition of overlapping sound sources, within the Real-Time Musical Interactions Team.

During the whole scholar year, my motivation towards research never stopped growing, and I eventually started the Doctorate in the Music Representations Team, along the prospective path of studying whether the theoretical field of information geometry would reveal relevant and beneficial for applications within the realm of audio signal processing. I am so grateful to my supervisors, Gérard Assayag and Arshia Cont, for placing their trust in me for this ambitious and multidisciplinary project. I also sincerely thank all members of the thesis jury, Francis Bach, Frank Nielsen, Roland Badeau, Silvère Bonnabel, and Jean-Luc Zarader, who accepted to evaluate my work and opened insightful discussions about it.

Today, after three intensive but enthusiastic years, there are many people I would like to acknowledge for their appearance on my research path, should they be colleagues, friends, or the both. To begin with, I warmly thank Philippe Codognet who hosted me for an exchange program at the Japanese-French Laboratory for Informatics (JFLI), within the University of Tokyo, and all people I met there, in particular, Philippe Esling and Florian Richoux for the nice moments we shared. The couple of months I spent at the JFLI have been one of the most enriching experiences in my life, both from professional and personal perspectives. I am also obliged to Shunichi Amari for his invitation to present my modest work at the Riken Brain Science Institute (BSI) in Tokyo.

Back in France, I would also like to acknowledge all speakers and members, and in particular the co-organizers, Frédéric Barbaresco, Arshia Cont, and Frank Nielsen, who participated in the Léon Brillouin seminar series on geometric science of information, that we organized at the IRCAM together with Thales and the Laboratoire d'Informatique de l'École Polytechnique (LIX). I do hope this is just the beginning of a long and fruitful collaboration.

I would further like to acknowledge all people I met at the IRCAM over the past few years. I notably thank the scientific director, Hugues Vinet, for hosting me there before and during the thesis. I am also grateful to Chunghsin Yeh and Axel Röbel, who kindly helped me for the evaluations of the algorithms on polyphonic music transcription and onset detection. I also thank, again for some of them, Carlos Agon, Moreno Andreatta, Gérard Assayag, Jean Bresson, Georges Bloch, and Laurent Bonnasse-Gahot, from the Musical Representation Team, as well as the engendered freaks, Arshia Cont, Jean-Louis Giavitto, and Florent Jacquemard, from the MuTant Project-Team, for the warm welcome and the passionate discussions we had. I must not forget the doctoral students of the team, Louis Bigo, José Echeveste, Philippe Esling, Jérémie Garcia, Benjamin Lévy, John Mandereau, Jérôme Nika, for their contagious good mood and passion.

There are many other people whom I appreciated for various reasons, including their common kindness, namely, Sara Adhitya, Yun-Kang Ahn, Grégory Beller, Joël Bensoam, Frédéric Bevilacqua, Tommasso Bianco, Éric Boyer, Baptiste Caramiaux, Grégoire Carpentier, Carmine Emanuele Cella, Frédéric Cornu, Pauline Eveno, Sarah Fdili Alaoui, Jules Françoise, Thomas Goepfer, Karim Haddad, Thomas Hélie, Thomas Hézard, Olivier Houix, Florian Kaiser, Mathieu Lagrange, Ianis Lallemand, Carlo Laurenzi, Marco Liuni, Fivos Maniatakos, Nicolas Misdariis, Yuki Mitsufuji, Markus Noisternig, Gilbert Nouno, Nicolas Obin, Olivier Pasquet, Geoffroy Peters, Norbert Schnell, Diemo Schwarz, Patrick Susini, Olivier Warusfel, and Bruno Zamborlin. I wish them all the best for the future and hope we will see again soon.

Last but not least, my hearted thanks go to my family, and more particularly to Sylvie, Hubert, and Ève, for their support, their patience, and simply their love.

Arnaud Dessein,
December 2012.

> There is geometry in the humming of the strings, there is music in the spacing of the spheres.
>
> Pythagoras

Introduction

This thesis aims at providing novel computational methods within the statistical framework of information geometry. We notably develop schemes for sequential change detection with exponential families and for non-negative matrix factorization with convex-concave divergences. Our primary motivations come from the context of audio signal processing where we apply these schemes to devise systems for real-time audio segmentation and for real-time polyphonic music transcription. The proposed methods, however, also fit naturally in the more general contexts of machine learning and signal processing. In the sequel, we frame the context of this thesis by introducing some background on information geometry from both theoretical and computational perspectives. We also position the outcomes of the thesis in this context. We notably outline the directions and main contributions of the present work, and describe the related publications and communications that arose out of this work.

Context of the thesis

To begin with, we introduce some background on information geometry. We first discuss the historical context of information geometry to frame this relatively recent mathematical theory. We then discuss more recent developments of the field of information geometry towards computational perspectives.

From information geometry theory

In general terms, *information geometry* is a field of mathematics that studies the theory of statistics, by using concepts of differential geometry such as smooth manifolds, and of information theory such as statistical divergences. Historically, information geometry emerged from the idea that many parametric statistical models of probability distributions possess a natural and intrinsic geometrical structure of differential manifold. Studying statistical inference in such structures ensures that the results of inference are invariant under the arbitrary choice of a parametrization for the family. Moreover, several statistical constructs can be interpreted in relation to geometrical concepts, which often provides interesting insights.

The founding work in information geometry is attributed to Rao [1945] who emphasized the importance to consider statistical inference from an intrinsic viewpoint, and notably proposed a structure of Riemannian manifold for certain parametric families with a metric defined by the Fisher information matrix. Efron [1975] first clarified the relations between the statistical notion of efficiency in asymptotic theory of estimation, and the geometrical concept of curvature for one-parameter statistical models. This was further pursued by Eguchi [1983] who introduced the notion of divergence, or contrast function, on statistical manifolds and discussed its relations with the efficiency of certain estimators for curved exponential families.

In the meantime, Chentsov [1982] provided a formal mathematical framework for information geometry using the language of category theory, in which he introduced the family of affine α-connections, discussed the duality of these affine connections with respect to the Fisher information metric, and proved the uniqueness of this metric and of these connections for statistical manifolds on finite sample spaces, when requiring invariance under Markov morphisms.[1] Independently, Amari [1982] studied the α-connections and α-divergences in link with asymptotic theory of estimation, and Nagaoka and Amari [1982] elucidated the duality of the α-connections and of the α-divergences by proposing a general theory of dually flat spaces.

Since then, several research directions have been investigated to extend these geometrical structures. For example, Eguchi [1985, 1992] developed the information geometry of divergences, by showing that any divergence on a statistical manifold induces a canonical torsion-free dualistic structure in terms of a Riemannian metric and of a pair of dual symmetric affine connections. Conversely, it was also shown by Matumoto [1993] that any torsion-free dualistic structure on a statistical manifold can be induced by a statistical divergence. Certain divergences have received a lot of attention in this context, notably Csiszár divergences whose geometry was thoroughly studied by Vos [1991]. Bregman divergences also revealed deep interests, in connection with exponential families of distributions and with dually flat structures, as put in perspective by Amari and Cichocki [2010]. Other divergences were also introduced by Zhang [2004] who elucidated a more general framework of duality.

Complementary directions were investigated by Barndorff-Nielsen [1986, 1987], Barndorff-Nielsen and Jupp [1997], who considered other Riemannian metrics than the expected Fisher information metric, by introducing the observed Fisher information metric, and by developing a general theory of yokes on statistical manifolds. Alternatives were also studied, such as the preferred point geometry of Critchley et al. [1993]. On a different perspective, Pistone and Sempi [1995], Gibilisco and Pistone [1998], Cena and Pistone [2007], extended the parametric finite-dimensional information geometry modeled on Euclidean spaces, by considering non-parametric infinite-dimensional statistical families modeled on Orlicz spaces.

Today many theoretical and applicative research works enlightened the relevance of studying statistics and its applications in various domains by the way of information geometry. This stimulated the creation of a large community with various interests in fields such as mathematics, physics, machine learning, signal processing, engineering science, which led to the maturity of information geometry and to its modern formulation in the seminal book of Amari and Nagaoka [2000]. For a good starting point, the early books of Amari [1985] and Amari et al. [1987] provide a solid theoretical basis and historical insights into the development of the field. For complementary treatments, we also refer to the books of Murray and Rice [1993], Kass and Vos [1997] and Arwini and Dodson [2008].

[1]With this respect, the study of information geometry differs from Klein's Erlangen program and related approaches to geometry, by replacing group transformations, or actions, with statistical transformations, or reductions, though the notion of symmetry, or invariance, remains central.

To computational information geometry

The research field of *computational information geometry* gathers a broad community around the development and application of computational methods that rely on theoretical constructs from information geometry. This community notably intersects the communities of machine learning and of signal processing. In particular, many techniques from machine learning and signal processing rely on the use of statistical models or distance functions to analyze and process the data. It is therefore a natural approach to elaborate computational methods based on information geometry, from the perspective of statistical manifolds or information metrics and divergences, and from the interplay between these notions.

Several authors have undertaken this approach with various purposes, such as studying theoretical aspects of Boltzmann machines [Amari et al. 1992], neural networks [Amari 1995], natural gradient learning [Amari 1998], robust estimation through minimization of divergences [Basu et al. 1998, Eguchi and Kano 2001], mean-field approximation [Tanaka 2000], hierarchies of probability distributions [Amari 2001], turbo codes [Ikeda et al. 2004], diffusion kernels [Lafferty and Lebanon 2005], prior selection [Snoussi 2005]. The information-geometric approach has also proved beneficial in a variety of applications such as data clustering and mining with α-divergences [Hero et al. 2002, Schwander and Nielsen 2011], data embedding and dimensionality reduction with the Fisher information [Carter et al. 2009, 2011], shape analysis with information metrics [Peter and Rangarajan 2006, 2009], blind source separation with independent component analysis in the space of estimating functions [Amari and Cardoso 1997, Amari 1999], or with robust estimation based on minimization of divergences [Mihoko and Eguchi 2002, Eguchi 2009].

In this context, certain divergences have been employed extensively. This is in particular the case of Bregman divergences and of their extensions, because of their links with convex optimization through convex duality, and with statistical exponential families through dually flat spaces. These divergences have notably been used to develop novel computational methods, often generalizing standard algorithms and schemes to a vast family of distance measures or related statistical models. Famous examples include the generalization of principal component analysis to exponential families based on the minimization of Bregman divergences [Collins et al. 2002], the extension of hard and soft clustering with consideration of k-means and expectation-maximization within a unifying framework for exponential families and Bregman divergences [Banerjee et al. 2005], and the extension of least squares and normality assumptions in regression to generalized linear models with exponential families [Wainwright and Jordan 2008].

These divergences have also been employed in a variety of techniques such as boosting methods [Murata et al. 2004] and their relations to weighted clustering [Nock and Nielsen 2006], clustering with approximation guarantees [Nock et al. 2008], surrogates for learning [Nock and Nielsen 2009], matrix factorizations [Dhillon and Sra 2006, Dhillon and Tropp 2008], low-rank kernel learning [Kulis et al. 2009], mixture learning [Fujimoto and Murata 2007], simplification and hierarchical representations of mixtures of exponential families [Garcia and Nielsen 2010], contextual re-ranking [Schwander and Nielsen 2010], shape retrieval [Liu et al. 2010, 2012], distance learning for semi-supervised clustering [Wu et al. 2012].

They have also proved relevant for the generalization of several standard computational geometry algorithms that were originally designed for the Euclidean distance, including nearest neighbor search [Cayton 2008, Nielsen et al. 2009a,b], range search [Cayton 2009], centroid computation [Nielsen and Nock 2009b, Nielsen and Boltz 2011, Vemuri et al. 2011], smallest enclosing balls [Nock and Nielsen 2005, Nielsen and Nock 2005, 2008, 2009a], Voronoi diagrams [Nielsen et al. 2007, Boissonnat et al. 2010, Nielsen and Nock 2011].

More generally, Bregman divergences, and other information divergences including Csiszár divergences, have revealed of key importance in statistical approaches to machine learning and signal processing. This had already been put in perspective in the early paper of Basseville [1989]. The literature on these issues has considerably expanded in the recent years and an up-to-date and thorough review is presented by Basseville [2013].

Outcomes of the thesis

Hereafter, we outline the directions of this thesis and summarize our main contributions in the realm of computational information geometry and signal processing. We also describe the publications and communications that arose out of this work.

Outline and main contributions

In the present work, we propose two independent algorithmic schemes that fall within the framework of computational information geometry. Although these methods naturally fit within the general domains of machine learning and signal processing, our initial motivations actually arise from two problems in audio signal processing, that of audio segmentation and that of polyphonic music transcription. Furthermore, we are deeply concerned with online machine listening, and we seek to design real-time systems to solve the two mentioned problems.

In this context, we address the problem of real-time audio segmentation by introducing novel computational methods for sequential change detection with exponential families. Concerning real-time polyphonic music transcription, we develop novel schemes for non-negative matrix factorization with convex-concave divergences. As discussed above, the two proposed algorithmic schemes are nonetheless of independent interest and directly applicable in other areas of statistical machine learning and signal processing. Therefore, the main body of this manuscript is organized into two parts, reporting respectively the theoretical contributions of the computational methods developed on the one hand, and the applicative contributions of these methods to audio signal processing on the other hand. The outline of the thesis is shown in Figure 1 and can be discussed as follows.

In Chapter 1, we introduce the theoretical preliminaries on information geometry that are necessary to the developments of Chapter 2 and of Chapter 3. The chapter is further divided into two parallel sections corresponding to the mathematical constructs required respectively for the two subsequent and independent chapters. We first present important results on exponential families of probability distributions

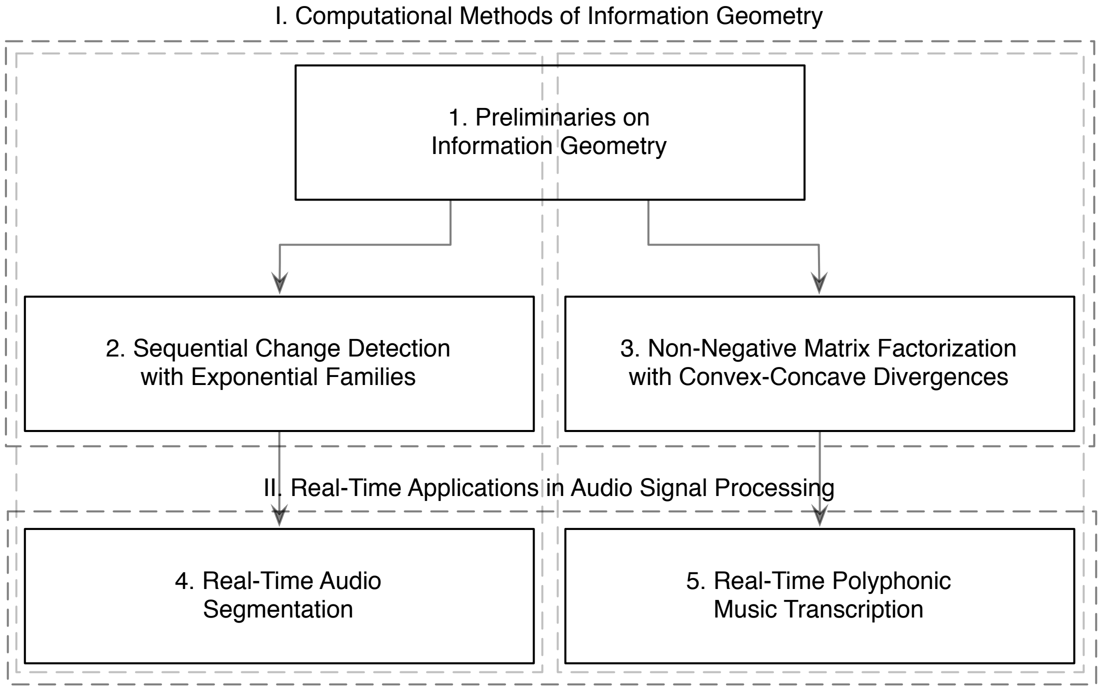

Figure 1.: Outline of the thesis. The thesis is organized into two main parts, that report in parallel the two developed computational methods of information geometry on the one hand, and their respective use for real-time applications in audio signal processing on the other hand.

in relation to convex duality and dually flat information geometry. These results are employed in Chapter 2 to develop computational methods for sequential change detection with exponential families. We then focus on introducing relevant notions about separable information divergences on the space of discrete positive measures, including Csiszár divergences, Bregman divergences and their generalizations through Jeffreys-Bregman and Jensen-Bregman divergences, as well as α-divergences or β-divergences and their generalization through skew (α, β, λ)-divergences. This is employed in Chapter 3 to elaborate computational methods for non-negative matrix factorization with convex-concave divergences.

In Chapter 2, we elaborate on the novel computational methods for sequential change detection with exponential families. To the best of our knowledge, it is the first time that the celebrated problem of change detection is investigated in the light of information geometry. We follow a standard approach where change detection is considered as a statistical decision problem with multiple hypotheses and is solved using generalized likelihood ratio test statistics. A major drawback of previous work in this context is to consider only known parameters before change, or to approximate the exact statistics when these parameters are actually unknown. This is addressed by introducing exact generalized likelihood ratios with arbitrary estimators, and by expanding them for exponential families. By showing tight links between the computation of these statistics and of maximum likelihood estimates, we derive a generic scheme for change detection with exponential families under common scenarios with known or unknown parameters and arbitrary estimators. We also interpret this scheme within the dually flat information geometry of exponential families, hence providing both statistical and geometrical intuitions to the problem, and bridging the gap between statistical and distance-based approaches to change detection. The scheme is finally revisited through convex duality, leading to an attractive scheme with closed-form sequential updates for the exact generalized likelihood ratio statistics, when both the parameters before and after change are unknown and are estimated by maximum likelihood. This scheme is applied in Chapter 4 to devise a general and unifying system for real-time audio segmentation.

In Chapter 3, we elaborate on the novel computational methods developed for non-negative matrix factorization with convex-concave divergences. We notably formulate a generic and unifying framework for non-negative matrix factorization with convex-concave divergences. This framework encompasses many common information divergences, such as Csiszár divergences, certain Bregman divergences, and in particular all α-divergences and β-divergences. A general optimization scheme is developed based on variational bounding with surrogate auxiliary functions for almost arbitrary convex-concave divergences. Monotonically decreasing updates are then obtained by minimizing the auxiliary function. The proposed framework also permits to consider symmetrized and skew divergences for the cost function. In particular, the generic updates are specialized to provide updates for Csiszár divergences, certain skew Jeffreys-Bregman divergences, and skew Jensen-Bregman divergences. This leads to several known multiplicative updates as well as novel multiplicative updates for α-divergences, β-divergences, and their symmetrized or skew versions. These results are also generalized by considering the family of skew (α, β, λ)-divergences. This is applied in Chapter 5 to design a real-time system for polyphonic music transcription.

In Chapter 4, we investigate the problem of audio segmentation. We notably devise a generic and unifying framework for real-time audio segmentation, based on the methods for sequential change detection with exponential families developed in Chapter 2. A major drawback of previous works in the context of audio segmentation, is that they consider specific signals and homogeneity criteria, or assume normality of the data distribution. Other issues arise from the potential computational complexity and non-causality of the schemes. The proposed system explicitly addresses these issues by controlling the information rate of the audio stream to detect changes in real time. The framework also bridges the gap between statistical and distance-based approaches to segmentation through the dually flat geometry of exponential families. We notably clarify the relations between various standard approaches to audio segmentation, and show how they can be unified and generalized in the proposed framework. Various applications are showcased to illustrate the generality of the framework, and a quantitative evaluation is performed for musical onset detection to demonstrate how the proposed approach can leverage modeling in complex problems.

In Chapter 5, we investigate the problem of polyphonic music transcription. We notably elaborate a real-time system for polyphonic music transcription by employing the computational methods for non-negative matrix factorization with convex-concave divergences developed in Chapter 3. We consider a supervised setup based on non-negative decomposition, where the music signal arrives in real time to the system and is projected onto a dictionary of note spectral templates that are learned offline prior to the decomposition. An important drawback of existing approaches in this context is the lack of controls on the decomposition. This is addressed by using the parametric family of (α, β)-divergences, and by explicitly interpreting their relevancy as a way to control the frequency compromise in the decomposition. The proposed system is evaluated through a methodological series of experiments, and is shown to outperform two state-of-the-art offline systems while maintaining low computational costs that are suitable to real-time constraints.

We conclude the manuscript with a general discussion and draw perspectives for future work. It is our hope that the presented contributions will bring interesting insights and directions for future research in the realm of audio signal processing, and more generally of machine learning and signal processing, in the relatively young but nonetheless prolific field of computational information geometry.

Related publications and communications

The present thesis has led to several publications and communications. This includes one book chapter, one journal article, three conference articles, two conference abstracts, one conference tutorial, and seven invited talks. These publications and communications are detailed below.

The work on sequential change detection with exponential families and its applications to real-time audio segmentation was first presented as an article with oral presentation at a national peer-reviewed conference:

> A. Dessein and A. Cont. Segmentation statistique de flux audio en temps-réel dans le cadre de la géométrie de l'information. In *23e Colloque du Groupe de Recherche et d'Études du Traitement du Signal (GRETSI)*, Bordeaux, France, September 2011a. [Dessein and Cont 2011a].

In this early version, we considered change detection in exponential families as a purely geometrical problem where the observations fall into Bregman balls onto the underlying statistical manifold, and are segmented according to the ball radii. This was applied to audio change detection based on spectral distributions for segmentation of polyphonic music into note slices. The ideas developed in this paper then matured as formulated in the thesis, that is, as an information-geometric problem where the statistical tests based on generalized likelihood ratio statistics are shown to be intrinsically linked with maximum likelihood estimation and with the canonical Kullback-Leibler divergences between the estimated distributions, or Bregman divergences on their parameters, in the different change hypotheses. This leveraged other audio applications where the previous approach failed, such as silence and activity segmentation based on energy features, as well as segmentation into speech and music, or into different speakers, based on timbral characteristics. Parts of this work have been accepted as an article in an international peer-reviewed journal:

> A. Dessein and A. Cont. An information-geometric approach to real-time audio segmentation. *IEEE Signal Processing Letters*, 20(4):331–334, April 2013a. [Dessein and Cont 2013a].

A complementary paper has also been accepted as an article with oral presentation in an international peer-reviewed conference with edited book proceedings:

> A. Dessein and A. Cont. Online change detection in exponential families with unknown parameters before and after change. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information: First International Conference, GSI 2013, Paris, France, August 28-30, 2013, Proceedings*, volume 8085 of *Lecture Notes in Computer Science*, pages 633–640. Springer, Berlin/Heidelberg, Germany, 2013b. [Dessein and Cont 2013b].

While the former article mostly focuses on audio processing applications with just the necessary theory to develop a real-time scheme for audio segmentation, the latter is rather targeted to the community of information geometry, and introduces more theoretical aspects as well as other applications to two well-known time series of the literature in geophysics and finance.

Concerning non-negative matrix factorization with convex-concave divergences and its applications to real-time polyphonic music transcription, our contributions were first presented as an article with poster presentation at an international peer-reviewed conference:

> A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 489–494, Utrecht, Netherlands, August 2010a. [Dessein et al. 2010a].

In this preliminary work, we employed the β-divergences with no discussion about convergence guarantees. We also presented this work as an extended abstract with code submission at an evaluation campaign in the reference international competition for tasks in music information retrieval:

> A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *6th Music Information Retrieval Evaluation eXchange (MIREX)*, Utrecht, Netherlands, August 2010b. [Dessein et al. 2010b].

During this contest, our system achieved results comparable to the state-of-the-art by finishing second out of six systems for the task of note transcription in polyphonic music, while actually being the only real-time system in competition. The system later evolved to include convergence guarantees for critical safety, in particular to ensure the monotonic decrease of the cost function along the iterations. This extension has been accepted as a peer-reviewed book chapter in a collective work by world-leading researchers on matrix information geometry:

> A. Dessein, A. Cont, and G. Lemaitre. Real-time detection of overlapping sound events with non-negative matrix factorization. In F. Nielsen and R. Bhatia, editors, *Matrix Information Geometry*, chapter 14, pages 341–371. Springer, Berlin/Heidelberg, Germany, 2013. [Dessein et al. 2013].

In this work, we also explored alternatives to the β-divergences by using a convex Euclidean cost function with sparsity penalty via non-negative sparse coding and ℓ_1-norm regularization. In addition, we considered other tasks of sound detection with overlapping sources, namely drum transcription and environmental sound recognition. These parallel developments are however not described in the thesis for the sake of conciseness.

During the three years of the thesis, we also investigated several other applications in audio signal processing and music information retrieval using computational methods of information geometry. In particular, we employed existing algorithms for parameter estimation in mixture models based on exponential families, as well as clustering and proximity search with Bregman divergences, for sample applications in audio quantization, indexing, structuring and querying. These applications were however not developed as thoroughly as that of audio segmentation and polyphonic music transcription, and are at present more at a proof-of-concept level. We therefore chose not to discuss them in this thesis and to focus instead on the more accomplished work, with both theoretical and practical contributions, that is described above. Nonetheless, our researches on information geometry for audio signal processing have been presented in their general framework at several scientific venues. We notably presented a tutorial at an international peer-reviewed conference:

> A. Dessein and A. Cont. Applications of information geometry to audio signal processing. In *14th International Conference on Digital Audio Effects (DAFx)*, Paris, France, September 2011b. [Dessein and Cont 2011b].

We also had an abstract with poster presentation at a non-peer-reviewed national conference for doctoral students:

> A. Dessein. La géométrie de l'information pour l'analyse et le traitement des flux audio. In *5es Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio (JJCAAS)*, Marseille, France, November 2009b. [Dessein 2009b].

Last but not least, we presented our researches orally in various invited talks:

> A. Dessein. Introduction to the tools of information geometry for the manipulation of audio streams. In *Séminaire Mathématiques, Musique et relations avec d'autres disciplines (MaMuX)*, Institut de Recherche et Coordination Acoustique/Musique, Paris, France, October 2009a. [Dessein 2009a].

> A. Dessein. Some applications of non-negative matrix factorization and of information geometry in audio signal processing. In *Research Seminar*, RIKEN Brain Science Institute, Tokyo, Japan, October 2010a. [Dessein 2010a].

> A. Dessein. Information geometry for real-time processing of audio signals. In *Research Seminar*, Japanese-French Laboratory for Informatics, Tokyo, Japan, November 2010b. [Dessein 2010b].

A. Cont and A. Dessein. Applications de la géométrie de l'information au traitement des flux audio temps réel. In *5e Biennale Française des Mathématiques Appliquées, Congrès de la Société de Mathématiques Appliquées et Industrielles (SMAI)*, Guidel, France, May 2011. [Cont and Dessein 2011].

A. Dessein. Music information geometry. In *Journée interdisciplinaire Mathématiques et Musique (JMM)*, Institut de Recherche Mathématique Avancée, Strasbourg, France, April 2011. [Dessein 2011].

A. Dessein and A. Cont. Applications of information geometry to audio signal processing. In *Indo-French Workshop on Matrix Information Geometry (MIG)*, pages 16–17, École Polytechnique, Palaiseau, France, February 2011c. [Dessein and Cont 2011c].

A. Dessein. Computational information geometry for audio signal processing. In *Symposium on Optimal Transport and Information Geometry*, Institut Henri Poincaré, Paris, France, February 2012. [Dessein 2012].

Part I.

Computational Methods of Information Geometry

1. Preliminaries on Information Geometry

This chapter presents the theoretical preliminaries on information geometry that are required for the elaboration of the computational methods proposed in the present work. We first introduce some prerequisites on exponential families of probability distributions, in relation to convex duality and dually flat information geometry. These notions are used in Chapter 2 to develop computational methods for sequential change detection with exponential families. We then focus on defining notions about separable information divergences on the space of discrete positive measures. This is employed in Chapter 3 to elaborate computational methods for non-negative matrix factorization with convex-concave divergences.

1.1. Exponential families of probability distributions

In this section, we introduce preliminaries on exponential families of probability distributions. We first define basic notions on standard and general exponential families. We then present first properties of these families, including reduction of general families to minimal standard families. We also discuss some results from convex duality for minimal steep standard families, and for maximum likelihood estimation when the family is also full. We finally interpret these notions within the framework of dually flat information geometry.

1.1.1. Basic notions

Exponential families are general parametric families of probability distributions that were introduced by Fisher [1934], Darmois [1935], Koopman [1936], Pitman [1936]. These families encompass a large class of statistical models that are commonly used in the realm of statistics and its applications, including the Bernoulli, Dirichlet, Gaussian, Laplace, Pareto, Poisson, Rayleigh, Von Mises-Fisher, Weibull, Wishart, log-normal, exponential, beta, gamma, geometric, binomial, negative binomial, categorical, multinomial models, among others.[1] Moreover, the class of exponential families is stable under various statistical constructs such as truncated and censored models, marginals, conditionals through linear projections, joint distributions of independent variables and in particular i.i.d. samples, among others. In this context,

[1]To be precise, some of these models actually need a restriction of their original parameter space to be considered as exponential families. We also notice that some statistical models are not exponential families, such as the uniform distributions because they do not share the same support, or the Cauchy distributions because they do not have finite moments.

employing exponential families not only permits the unification and generalization of the problems considered, but also often contributes to a deeper understanding of the problem structures. The theory of exponential families has become wide and we only expose here the main notions and results needed in the present work. For more theoretical background, we redirect to the early article of Chentsov [1966], and to the dedicated books of Barndorff-Nielsen [1978] and Brown [1986] which contain proofs of the results stated here. Complementary information is provided in the more general books of Lehmann and Casella [1998], Lehmann and Romano [2005]. Before defining general exponential families, we first introduce the useful notion of standard exponential family.

Definition 1.1. A *standard exponential family* is a parametric statistical model $\{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$ on the Borel subsets of $\mathbb{R}^m$, which is dominated by a σ-finite measure μ, and whose respective probability densities $p_{\boldsymbol{\theta}}$ with respect to μ can be expressed for any $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^m$ as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \lambda(\boldsymbol{\theta})^{-1} \exp(\boldsymbol{\theta}^\top \mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathbb{R}^m \enspace , \tag{1.1}$$

where $\lambda \colon \Theta \to \mathbb{R}_+^*$. The parameter $\boldsymbol{\theta}$ is then called *canonical parameter* or *natural parameter*, $\mathbf{x}$ is called *canonical observation* or *sufficient observation*, and λ is called *partition function* or *normalizer*.

Remark 1.1. We assume implicitly that the parameter space Θ is non-empty.

Remark 1.2. The normalizer λ ensures that all probability densities $p_{\boldsymbol{\theta}}$ normalize to one, and thus verifies the following relation:

$$\lambda(\boldsymbol{\theta}) = \int_{\mathbb{R}^m} \exp(\boldsymbol{\theta}^\top \mathbf{x}) \, \mu(\mathrm{d}\mathbf{x}) \enspace . \tag{1.2}$$

We see from the latter remark that the parameter space Θ is not necessarily maximal, in the sense that the above integral may be finite for other values of $\boldsymbol{\theta}$. As a result, the normalizer λ can be extended to determine probability densities $p_{\boldsymbol{\theta}}$ for these values of $\boldsymbol{\theta} \in \mathbb{R}^m \setminus \Theta$. This leads naturally to the following definitions.

Definition 1.2. The *natural parameter space* $\mathcal{N}$ is the set defined as follows:

$$\mathcal{N} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^m \colon \int_{\mathbb{R}^m} \exp(\boldsymbol{\theta}^\top \mathbf{x}) \, \mu(\mathrm{d}\mathbf{x}) < +\infty \right\} \enspace . \tag{1.3}$$

Remark 1.3. The above integral is always positive since μ cannot be null. The normalizer can therefore define probability densities for any $\theta \in \mathcal{N}$.

Remark 1.4. By construction, Θ is a subset of $\mathcal{N}$, and $\mathcal{N}$ is the maximal parameter space onto which the family can be extended in the sense discussed above. Nevertheless, it may happen that the added parameters $\boldsymbol{\theta} \in \mathcal{N} \setminus \Theta$ determine distributions $p_{\boldsymbol{\theta}}$ that already are in the original family when the parametrization is not one-to-one.

Definition 1.3. A standard exponential family is *full* if $\Theta = \mathcal{N}$.

Other important classes of standard exponential families can be defined depending on the properties of the natural parameter space.

Definition 1.4. A standard exponential family is *regular* if $\mathcal{N} = \operatorname{int} \mathcal{N}$.

Definition 1.5. A standard exponential family is *minimal* if $\dim \mathcal{N} = \dim \mathcal{K} = k$, where $\mathcal{K}$ is the convex support of the dominating measure μ.

Remark 1.5. Minimality thus avoids dimensional degeneracy of both $\mathcal{N}$ and $\mathcal{K}$, by requiring that the parameters do not lie in a proper affine subspace of $\mathbb{R}^m$, and that the dominating measure is not concentrated on a proper affine subspace of $\mathbb{R}^m$.

We now introduce a function which reveals crucial to the study of minimal exponential families.

Definition 1.6. The *log-partition function* or *log-normalizer* ψ is the logarithm of the normalizer λ:

$$\psi(\boldsymbol{\theta}) = \log \lambda(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \Theta \ . \tag{1.4}$$

Remark 1.6. The respective probability densities $p_{\boldsymbol{\theta}}$ with respect to μ can therefore also be expressed as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp(\boldsymbol{\theta}^\top \mathbf{x} - \psi(\boldsymbol{\theta})) \ . \tag{1.5}$$

In the sequel, we often consider the normalizer λ and log-normalizer ψ extended to $\mathcal{N}$ or $\mathbb{R}^m$. It is clear that λ and ψ take respectively finite positive and finite values not only on Θ but also on $\mathcal{N}$, and they equal $+\infty$ on $\mathbb{R}^m \setminus \mathcal{N}$. We finally move on to more general exponential families. Considering standard families is not always convenient from a practical viewpoint. Indeed, many useful statistical models are not directly standard families, but their theoretical study can often be reduced to that of standard families after suitable transformations.

Definition 1.7. An *exponential family* is a parametric statistical model $\{P_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \Xi}$ on a measurable space $(\mathcal{X}, \mathcal{A})$, which is dominated by a σ-finite measure μ, and whose respective probability densities $p_{\boldsymbol{\xi}}$ with respect to μ can be expressed for any $\boldsymbol{\xi} \in \Xi$ as follows:

$$p_{\boldsymbol{\xi}}(\mathbf{x}) = C(\boldsymbol{\xi}) h(\mathbf{x}) \exp(R(\boldsymbol{\xi})^\top T(\mathbf{x})) \quad \text{for all } \mathbf{x} \in \mathcal{X} \ , \tag{1.6}$$

where $C \colon \Xi \to \mathbb{R}_+^*$, $R \colon \Xi \to \mathbb{R}^m$, $h \colon \mathcal{X} \to \mathbb{R}_+$ is Borel measurable, and $T \colon \mathcal{X} \to \mathbb{R}^m$ is Borel measurable.

Remark 1.7. We again assume implicitly that Ξ is non-empty.

Remark 1.8. A standard exponential family is obviously an exponential family.

In exponential families, the function C plays the same role as the inverse of the normalizer for standard families. The transformation R is intuitively a reparametrization of the family, while the transformation h is a modification of the dominating measure. Finally, the transformation T can be seen as a suitable reduction from a statistical viewpoint.

1.1.2. First properties

We begin with explaining how to reduce the study of general exponential families to that of minimal standard families. This is a consequence of the following proposition and theorem.

Proposition 1.1. *The function T is a sufficient statistic.*

Remark 1.9. It justifies that $\mathbf{x}$ is called sufficient observation for standard families.

Theorem 1.2. *Any exponential family can be reduced by sufficiency, reparametrization, and proper choice of a dominating measure, to a minimal standard exponential family.*

Remark 1.10. It can then be shown that two such reductions have necessarily the same dimension, and are actually related through linked affine transforms of their respective natural parameters and of their respective sufficient observations.

The study of exponential families can be reduced to that of minimal standard families. We thus focus in the sequel on minimal standard exponential families. These families inherit several useful properties from their structure. We discuss two of these properties hereafter.

Proposition 1.3. *Any minimal standard exponential family is identifiable.*

Remark 1.11. This means that the natural parametrization is one-to-one, and thus makes statistical inference about parameters relevant.

Proposition 1.4. *The normalizer λ and log-normalizer ψ are smooth on the interior $\operatorname{int}\mathcal{N}$ of the natural parameter space. Moreover, λ can be differentiated at any order $n \in \mathbb{N}$ with respect to variables $\alpha \in \{1, \ldots, k\}^n$ under the integral sign:*

$$\partial^{\alpha} \lambda(\boldsymbol{\theta}) = \int_{\mathbb{R}^m} \partial^{\alpha}_{\boldsymbol{\theta}} \exp(\boldsymbol{\theta}^\top \mathbf{x}) \, \mu(\mathrm{d}\mathbf{x}) \quad \textit{for all } \boldsymbol{\theta} \in \operatorname{int}\mathcal{N} \ . \tag{1.7}$$

Remark 1.12. For regular families, smoothness and differentiability under the integral sign hold everywhere on $\mathcal{N} = \operatorname{int}\mathcal{N}$.

Remark 1.13. An interesting consequence is that the moments of a random variable X distributed according to $p_{\boldsymbol{\theta}}$ can be obtained from the derivatives of ψ. In particular, we obtain the expectation and variance as follows:

$$E_{\boldsymbol{\theta}}(X) = \nabla\psi(\boldsymbol{\theta}) \quad , \tag{1.8}$$

$$V_{\boldsymbol{\theta}}(X) = \nabla^2\psi(\boldsymbol{\theta}) \ . \tag{1.9}$$

1.1.3. Convex duality

We now introduce notions from convex duality. We only expose the relevant application of this to minimal standard exponential families, which is just the tip of a much richer theory in convex analysis. For additional information, we redirect to the comprehensive book of Rockafellar [1970].

Definition 1.8. The *Fenchel conjugate* of an arbitrary function φ on $\mathbb{R}^m$ is the function $\varphi^\star$ defined as follows:

$$\varphi^\star(\boldsymbol{\xi}^\star) = \sup_{\boldsymbol{\xi} \in \mathbb{R}^m} \boldsymbol{\xi}^\top \boldsymbol{\xi}^\star - \varphi(\boldsymbol{\xi}) \quad \text{for all } \boldsymbol{\xi}^\star \in \mathbb{R}^m \ . \tag{1.10}$$

Proposition 1.5. *The Fenchel conjugate $\varphi^\star$ of a closed proper convex function φ is also a closed proper convex function, and we have $\varphi^{\star\star} = \varphi$.*

Definition 1.9. A proper convex function φ is *essentially smooth* if the interior $\operatorname{int}\operatorname{dom}\varphi$ of its effective domain is non-empty, if it is differentiable on $\operatorname{int}\operatorname{dom}\varphi$, and if $\lim_{n\to+\infty} \|\nabla\varphi(\boldsymbol{\xi}_n)\| = +\infty$ for any sequence of points $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots \in \operatorname{int}\operatorname{dom}\varphi$ that converges to a boundary point of $\operatorname{int}\operatorname{dom}\varphi$.

Definition 1.10. A proper convex function φ is of *Legendre type* if it is closed, essentially smooth, and strictly convex on the interior $\operatorname{int}\operatorname{dom}\varphi$ of its effective domain.

The application of convex duality to exponential families arises from the nice properties possessed by the log-normalizer.

Proposition 1.6. *The natural parameter space $\mathcal{N}$ is a convex set.*

Proposition 1.7. *The log-normalizer ψ is a closed proper strictly convex function with effective domain $\operatorname{dom}\psi = \mathcal{N}$. Moreover, its Fenchel conjugate $\phi = \psi^\star$ is a closed essentially smooth function with effective domain $\operatorname{int}\mathcal{K} \subseteq \operatorname{dom}\phi \subseteq \mathcal{K}$, and we have $\psi = \phi^\star$.*

Remark 1.14. This is the result of a more general duality between essential smoothness and essential convexity for arbitrary convex functions.

Remark 1.15. We remark that in order to have full duality between ψ and ϕ, we would need ψ to be essentially smooth, and ϕ to be strictly convex, which is not necessarily the case.

In this context, it is convenient to require stronger regularity of the exponential family in order to have a full convex duality. This can be discussed as follows.

Definition 1.11. A minimal standard exponential family is *steep* if the log-normalizer ψ is essentially smooth.

Remark 1.16. In particular, it can be shown that any regular family is actually steep.

Remark 1.17. Since the log-normalizer ψ is necessarily a closed proper strictly convex function which is differentiable on $\operatorname{int}\operatorname{dom}\psi \neq \emptyset$, essential smoothness of ψ and steepness of the family are equivalent to the assumption $\lim_{n\to+\infty} \|\nabla\psi(\boldsymbol{\theta}_n)\| = +\infty$ for any sequence of points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \in \operatorname{int}\operatorname{dom}\psi$ that converges to a boundary point of $\operatorname{int}\operatorname{dom}\psi$.

Theorem 1.8. *For any minimal steep standard exponential family, ψ and ϕ are of Legendre type. Moreover, $\nabla\psi$ defines a homeomorphism of $\operatorname{int}\operatorname{dom}\psi = \operatorname{int}\mathcal{N}$ and $\operatorname{int}\operatorname{dom}\phi = \operatorname{int}\mathcal{K}$, with inverse $(\nabla\psi)^{-1} = \nabla\phi$.*

Remark 1.18. Steepness ensures that the map $\nabla\psi$ is onto. If the family is not steep, then $\nabla\psi(\operatorname{int}\mathcal{N})$ is a proper subset of $\operatorname{int}\mathcal{K}$, so that actually $\nabla\psi$ is a homeomorphism of $\operatorname{int}\mathcal{N}$ and $\operatorname{int}\mathcal{K}$ iff the family is steep.

Remark 1.19. In particular, the theorem holds for any regular minimal standard exponential family.

This theorem shows that a minimal steep family with parameter space $\operatorname{int}\mathcal{N}$ can be reparametrized by the gradient $\nabla\psi$ of the log-normalizer, and the range of this parametrization is $\operatorname{int}\mathcal{K}$.

Definition 1.12. The *mean value parameter*, or *expectation parameter*, $\boldsymbol{\eta} \in \operatorname{int}\mathcal{K}$ is the parameter associated to the reparametrization of the natural parameter $\boldsymbol{\theta} \in \operatorname{int}\mathcal{N}$ by the gradient $\nabla\psi$ of the log-normalizer:

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \operatorname{int}\mathcal{N} \ , \tag{1.11}$$

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \nabla\phi(\boldsymbol{\eta}) \quad \text{for all } \boldsymbol{\eta} \in \operatorname{int}\mathcal{K} \ . \tag{1.12}$$

Remark 1.20. The parameter name as a mean value or expectation comes from the relation $\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(X)$.

Remark 1.21. It is convenient for regular families that the expectation parameter reparametrizes the full family.

In certain situations, such as when studying maximum likelihood estimators, this parametrization is more convenient than the natural one.

1.1.4. Maximum likelihood

We now present some general results about maximum likelihood estimation in full minimal steep standard exponential families.

Theorem 1.9. *For any full minimal steep standard exponential family, there exists a unique maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ml}}$ of $\boldsymbol{\theta}$ on $\operatorname{int}\mathcal{K}$, and it can be expressed as follows:*

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ml}}(\mathbf{x}) = \nabla\phi(\mathbf{x}) \quad \textit{for all } \mathbf{x} \in \operatorname{int}\mathcal{K} \ . \tag{1.13}$$

Moreover, if $\mathbf{x} \notin \operatorname{int}\mathcal{K}$, then no maximum likelihood estimate of $\boldsymbol{\theta}$ from $\mathbf{x}$ exists.

Remark 1.22. The theorem shows that the maximum likelihood estimator on $\operatorname{int}\mathcal{K}$ is one-to-one and can be expressed in the expectation parametrization simply as the sufficient observation:

$$\widehat{\boldsymbol{\eta}}_{\mathrm{ml}}(\mathbf{x}) = \mathbf{x} \ . \tag{1.14}$$

Remark 1.23. As a result, it is sufficient for maximum likelihood estimates to exist with probability one that $\mu(\mathcal{K} \setminus \operatorname{int}\mathcal{K}) = 0$. This is always satisfied when μ is dominated by the Lebesgue measure, but never satisfied when μ has finite support or more generally countable support and $\mathcal{K} \neq \mathbb{R}^m$.

Remark 1.24. When the family is not steep, maximum likelihood estimates also exist and are unique iff $\mathbf{x} \in \operatorname{int}\mathcal{K}$, and have the same expression as above on $\nabla\psi(\operatorname{int}\mathcal{N}) \subset \operatorname{int}\mathcal{K}$. Nonetheless, the expression cannot be determined as is when $\mathbf{x} \notin \nabla\psi(\operatorname{int}\mathcal{N})$. Moreover, the maximum likelihood estimator is not one-to-one anymore.

Remark 1.25. In a steep family, boundary points of $\mathcal{N}$ which belong to $\mathcal{N}$ do not occur among the values of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ml}}$ and are thus superfluous in this sense.

We naturally present the extension of this when considering an i.i.d. sample from the exponential family.

Corollary 1.10. *For any i.i.d. sampling model of size $n \in \mathbb{N}^*$ from a full minimal steep standard exponential family, there exists a unique maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ml}}$ of $\boldsymbol{\theta}$ on $\mathcal{K}^n_{\mathrm{ml}} = \{(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in (\mathbb{R}^m)^n : \frac{1}{n}\sum_{j=1}^n \mathbf{x}_j \in \operatorname{int}\mathcal{K}\}$, and it can be expressed as follows:*

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ml}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \nabla\phi\left(\frac{1}{n}\sum_{j=1}^n \mathbf{x}_j\right) \quad \text{for all } (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{K}^n_{\mathrm{ml}} \ . \tag{1.15}$$

Moreover, if $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \notin \mathcal{K}^n_{\mathrm{ml}}$, then no maximum likelihood estimate of $\boldsymbol{\theta}$ from $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ exists.

Remark 1.26. This is actually a direct consequence of the fact that an i.i.d. sampling model of size n from an exponential family with log-normalizer ψ, natural parameter $\boldsymbol{\theta}$ and sufficient observation $\mathbf{x}$, is also an exponential family with log-normalizer $n\psi$, natural parameter $\boldsymbol{\theta}$, and sufficient observation $\sum_{j=1}^n \mathbf{x}_j$.

Remark 1.27. The corollary shows that the maximum likelihood estimator on $\mathcal{K}^n_{\mathrm{ml}}$ can be expressed in the expectation parametrization simply as follows:

$$\widehat{\boldsymbol{\eta}}_{\mathrm{ml}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{n}\sum_{j=1}^n \mathbf{x}_j \ . \tag{1.16}$$

Remark 1.28. It appears that maximum likelihood estimates for steep families exist with probability increasing up to one as the sample size n tends to $+\infty$.

1.1.5. Dually flat geometry

The above notions are interpretable within the framework of dually flat information geometry. For the sake of conciseness, we do not introduce the mathematical constructs behind this theory, and redirect instead to the book of Amari and Nagaoka [2000]. We rather present intuitively the concepts that are relevant to the present work. In the sequel, we consider a minimal steep standard exponential family $\mathcal{P} = \{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\operatorname{int}\mathcal{N}}$ on the interior of its natural parameter space.

To sum up intuitively the basic concepts, information geometry considers a parametric statistical model as a space that locally looks like a Euclidean vector space, but that globally differs from this Euclidean vector space in general. This is the basic intuition behind viewing the statistical model as a topological manifold. On this statistical manifold, each point represents a probability distribution of the model. Moreover, the parameters of the respective distributions are their coordinates in the underlying coordinate system. The exponential family $\mathcal{P}$ being identifiable and having a non-empty connected open parameter space $\operatorname{int}\mathcal{N}$, it can be viewed as a topological manifold with a global coordinate system provided by the natural parameters $\boldsymbol{\theta}$ on $\operatorname{int}\mathcal{N}$.

We can then equip the statistical manifold with a differential structure by considering an atlas of coordinate systems that are compatible with the reference one in the sense that they are smooth reparametrizations of the model. This permits to define tangent spaces at each point of the manifold, which are intuitively linearizations of the manifold around these respective points. It permits to enhance the exponential family $\mathcal{P}$ as a differential manifold with a differential structure consisting of the smooth reparametrizations of the natural parameter $\boldsymbol{\theta} \in \operatorname{int}\mathcal{N}$, and notably includes the expectation parameter $\boldsymbol{\eta} \in \operatorname{int}\mathcal{K}$.

The statistical manifold can further be endowed with the Fisher information Riemannian metric, defined by the Fisher information matrix, and consisting of scalar products on the respective tangent spaces. It makes the model a Riemannian manifold and provides a way to compute the length of vectors in the tangent spaces. We can also compute the length of a curve joining two distributions by integrating the length of the speed vector along it. It defines an intrinsic notion of metric distance between two probability distributions on the statistical manifold by considering the metric geodesics, which are the curves that minimize the length between two points. Considering the exponential family $\mathcal{P}$, the Fisher information matrix is given by $G(\boldsymbol{\theta}) = \nabla^2\psi(\boldsymbol{\theta})$ on $\operatorname{int}\mathcal{N}$, and thus also equals the variance $V_{\boldsymbol{\theta}}(X)$. Since ψ is strictly convex, its Hessian and the Fisher information matrix $G(\boldsymbol{\theta})$ are positive-definite, hence defining a Riemannian metric g on $\mathcal{P}$ and making $(\mathcal{P}, g)$ a Riemannian manifold.

More general notions of geodesics can also be defined by introducing the affine α-connections which are dual in pairs with respect to the Fisher information metric. These connections intuitively characterize the way of passing from one tangent space to another one in its neighborhood. The affine α-geodesics are then defined as curves with a null acceleration, similarly to the straight lines in Euclidean geometry. This generalization coincides with that of metric geodesics when considering the self-dual, or metric, Levi-Civita connection. Thanks to the smoothness properties of the exponential family $\mathcal{P}$, the dual affine $\pm\alpha$-connections $\{(\nabla^{(+\alpha)}, \nabla^{(-\alpha)})\}_{\alpha\in\mathbb{R}_+}$, and corresponding dual affine $\pm\alpha$-geodesics, can be defined.

Last but not least, more general distance functions can be introduced by employing relevant information divergences that are locally compatible with both the metric and the affine connections considered. It appears that for certain families, there exist both a pair of dual affine $\pm\alpha$-connections that are flat, and a somewhat canonical pair of associated dual $\pm\alpha$-divergences. In such structures, there also exist two dual affine coordinate systems in which the respective geodesics are provided by simple line segments between the parameters. In particular, $(\mathcal{P}, g, \nabla^{(+1)}, \nabla^{(-1)})$ is a dually flat space, and the natural and expectation parameters $\{(\boldsymbol{\theta}, \operatorname{int}\mathcal{N}), (\boldsymbol{\eta}, \operatorname{int}\mathcal{K})\}$ are actually dual affine coordinate systems. Additionally, the canonical dual $\pm\alpha$-divergences are provided by the Kullback-Leibler and dual Kullback-Leibler divergences on the probability distributions, and can alternatively be computed in the respective coordinate systems with Bregman divergences, generated respectively by the log-normalizer ψ and its Fenchel conjugate ϕ, on the parameters. This dually flat geometry generalizes the standard self-dual Euclidean geometry, with two dual Bregman divergences instead of the self-dual Euclidean distance, two dual geodesics, and a generalized Pythagorean theorem.

Let us formalize the notions of divergences that we need in the present work. We first define the Kullback-Leibler divergence, introduced by Kullback and Leibler [1951], then define the Bregman divergences, introduced by Bregman [1967]. We notice that these divergences can be defined in a wider setting, but the following definitions are sufficient here.

Definition 1.13. Let $\mathcal{S}$ be a statistical model on a measurable space $(\mathcal{X}, \mathcal{A})$, which is dominated by a σ-finite measure μ. The *Kullback-Leibler* divergence D_{KL} on $\mathcal{S}$ is the function defined as follows:

$$D_{\mathrm{KL}}(P \| P') = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{p'(x)} \, \mu(\mathrm{d}x) \quad \text{for all } P, P' \in \mathcal{S} \ , \tag{1.17}$$

where p, p' are the respective probability densities of P, P' with respect to μ.

Remark 1.29. The Kullback-Leibler divergence can be defined more generally between two probability measures as soon as the first one is absolutely continuous with respect to the second one. For exponential families, the probability measures share the same support so that they are actually absolutely continuous with respect to each other.

Definition 1.14. Let φ be a convex function that is differentiable on the interior $\operatorname{int} \Xi$ of its effective domain $\operatorname{dom} \varphi = \Xi$. The *Bregman divergence* generated by φ is the function B_φ defined as follows:

$$B_\varphi(\boldsymbol{\xi} \| \boldsymbol{\xi}') = \varphi(\boldsymbol{\xi}) - \varphi(\boldsymbol{\xi}') - (\boldsymbol{\xi} - \boldsymbol{\xi}')^\top \nabla \varphi(\boldsymbol{\xi}') \quad \text{for all } \boldsymbol{\xi}, \boldsymbol{\xi}' \in \operatorname{int} \Xi \ . \tag{1.18}$$

Remark 1.30. We can extend the divergence straightforward to include any $\boldsymbol{\xi} \in \Xi$.

Finally, for exponential families, the Bregman divergences on natural and expectations parameters are linked with the Kullback-Leibler divergence on corresponding distributions.

Proposition 1.11. *For any minimal steep standard exponential family, we have the following relation:*

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}'}) = B_\psi(\boldsymbol{\theta}' \| \boldsymbol{\theta}) = B_\phi(\boldsymbol{\eta}(\boldsymbol{\theta}) \| \boldsymbol{\eta}(\boldsymbol{\theta}')) \quad \textit{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \operatorname{int} \mathcal{N} \ . \tag{1.19}$$

Remark 1.31. The presented notions and proposition can in general be extended to non-steep families. The difference is that the expectation parameter $\boldsymbol{\eta}$ lies in a proper subset of $\operatorname{int} \mathcal{K}$. From a technical viewpoint, steepness is however useful to maximum likelihood estimation, where the maximum likelihood estimates exist and are unique as soon as the average of the sufficient observations lies in $\operatorname{int} \mathcal{K}$, which happens with probability increasing up to one as the sample size grows to infinity, and are then given in expectation parameters by this average. For non-steep families, they also exist and are unique, but are given as is only when the average further lies in the interior of the range of the expectation parameter, which does not necessarily happen with probability increasing up to one as the sample size grows to infinity. Finally, certain notions, such as the generalized Pythagorean theorem, also rely on steepness to be properly constructed.

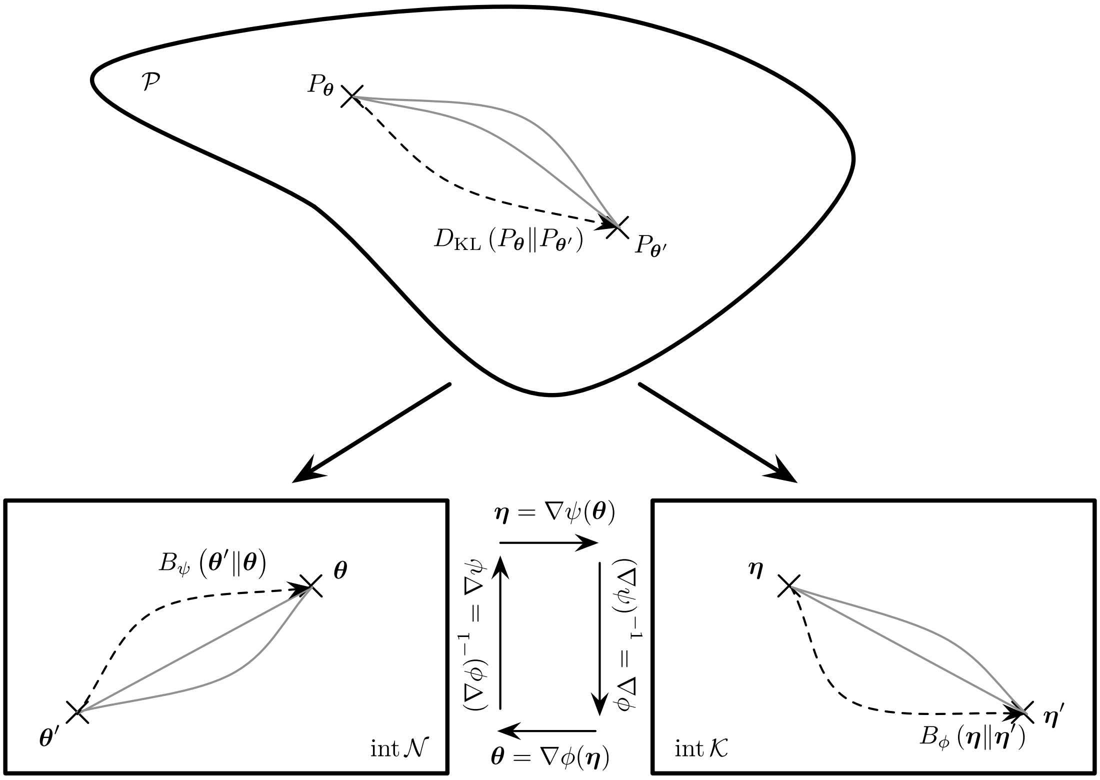

Figure 1.1.: Dually flat geometry of exponential families. The canonical Kullback-Leibler divergence between two probability distributions on the statistical manifold can be computed in the natural and expectation parameters as Bregman divergences using convex duality.

The notions formalized here and employed in the sequel are summarized in Figure 1.1. Although we do not use them explicitly in the sequel, we have also represented the two dual geodesics as gray lines on the statistical manifold and on the natural and expectation parameter spaces, for the sake of illustration.

1.2. Separable divergences on the space of discrete positive measures

In this section, we introduce preliminaries about separable divergences on the space of discrete positive measures. We begin with defining basic notions on divergences and in particular on separable divergences. We then present some well-known classes of divergences, in particular Csiszár divergences, but also Bregman divergences and their skew generalizations through Jeffreys-Bregman and Jensen-Bregman divergences. These general classes encompass famous information divergences, including the parametric families of α-divergences and β-divergences. These two parametric families can also be unified and extended with the recently proposed (α, β)-divergences. We further introduce a direct but novel generalization of them as skew (α, β, λ)-divergences through a standard skewing procedure.

1.2.1. Basic notions

We begin with introducing the central concept of divergence which generalizes the usual notion of metric distance.

Definition 1.15. A *divergence* on a set $\mathcal{Y} \subseteq \mathbb{R}^m$ is a function $D \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ such that $D(\mathbf{y}\|\mathbf{y}') \geq 0$ and $D(\mathbf{y}\|\mathbf{y}) = 0$ for all $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$.

Remark 1.32. Metric distances are usually defined by the three axioms of (i) coincidence, or identity of indiscernibles, $D(\mathbf{y}, \mathbf{y}') = 0 \Rightarrow \mathbf{y} = \mathbf{y}'$, (ii) symmetry, $D(\mathbf{y}, \mathbf{y}') = D(\mathbf{y}', \mathbf{y})$, (iii) subadditivity, or triangular inequality, $D(\mathbf{y}, \mathbf{y}'') \leq D(\mathbf{y}, \mathbf{y}') + D(\mathbf{y}', \mathbf{y}'')$. Together, these three axioms imply the separation property $D(\mathbf{y}, \mathbf{y}') \geq 0$ with equality iff $\mathbf{y} = \mathbf{y}'$. For divergences, we actually only require the two less restrictive axioms of (i) non-negativity, $D(\mathbf{y}\|\mathbf{y}') \geq 0$, (ii) identity, $D(\mathbf{y}\|\mathbf{y}) = 0$. Therefore a divergence may not be symmetric, hence the notation $D(\mathbf{y}\|\mathbf{y}')$ instead of $D(\mathbf{y}, \mathbf{y}')$, nor verify the triangular inequality. Moreover, we do not require here the identity of indiscernibles nor the separation property, which are actually equivalent for divergences as defined here, since they are not needed for the subsequent derivations to hold.

Remark 1.33. For technical convenience, we consider here divergences on the Cartesian square of $\mathcal{Y}$ so that we can compare any pair of points in $\mathcal{Y}$. Sometimes, it is possible to extend a divergence on a subset of $\mathbb{R}^m \times \mathbb{R}^m$ which is not a Cartesian square, nor even a Cartesian product, for example, on $\mathbb{R}_+^* \times \mathbb{R}_+ \cup \{(0,0)\}$ for certain scalar Csiszár divergences, or $\mathbb{R}_+ \times \mathbb{R}_+^* \cup \{(0,0)\}$ for certain scalar Bregman divergences. Such divergences are not necessarily well-behaved everywhere, notably on the boundary. Moreover, this technical requirement permits to consider symmetrization and skewing of arbitrary divergences without difficulty.

Remark 1.34. In practice, choosing a relevant divergence for a given problem is often of key importance but is not always an easy task. Several complementary approaches or general guidelines may help this decision. First of all, one may select a divergence or class of divergences using an axiomatic approach, by listing the desired properties that the relevant divergences should verify. Some common properties include symmetry, triangular inequality, additivity, invariance under sufficiency. In some cases, it is then possible to determine exactly the form of divergences that verify the desired properties. For example, an axiomatic approach to choose over common classes of distortion functions such as the Csiszár, Bregman, and Burbea-Rao divergences, has been elucidated through the information bottleneck method [Harremoës and Tishby 2007]. Another famous example is the distinction between Rényi and Tsallis entropies in statistical physics depending on extensivity or non-extensivity. Secondly, it is also possible to derive relevant divergences in link with statistical considerations. There exist strong relations between inference in certain statistical models, and optimization of associated distance functions between the model and observed data. An example which is relevant to the present work is that of maximum likelihood estimation in exponential families and its relation to minimization of the Kullback-Leibler or Bregman divergences. Other known relations include inference in power laws and Tsallis relative distances, as well as inference in Tweedie models, or in exponential dispersion and generalized linear models, and β-divergences. Such relations may guide the honest choice of a divergence if one has some intuitions on the distribution of the observations. In this context, the topology of the data, that is, if the data are discrete, continuous, complex, real, positive, negative, compact, among other properties, can provide insights on a reasonable choice for the underlying model. For instance, it is hopefully quite unlikely that the discrete and positive height of a population measured in centimeters, is modeled reliably with the ubiquitous bell-like shaped Gaussian model, though it might be a sufficient and convenient approximation for some applications. A statistical perspective may also allow the use of techniques from model selection to choose the best statistical model, and hence associated divergence, among several possible models. Thirdly, and with analogy to model selection in a statistical approach, the framework of distance or metric learning provides a data-driven approach to learn a relevant distance function directly by analyzing the data. The drawback of such methods is in general to accommodate only certain types of distances so as to formulate the learning as an optimization problem over a set of restricted and parametrized distances. For example, Mahalanobis distances are widely used in this context, where the optimization takes place over the convex cone of positive-definite matrices. From a statistical perspective, this restriction would correspond to assuming some normality properties of the data or of the noise distribution. Last but not least, expertise is in general an important quality for the practitioner to make a good choice, and rules of thumb or trial and error may often guide this choice. Several divergences are well-known in certain communities because they have already proved relevant for specific tasks. In this context, using parametric families of divergences that encompass common distance measures may be of great interest, since the user can then tweak the parameters to tune them and tailor the algorithm to the application at hand, while guaranteeing results that are at least as good as that of the usual encompassed distance measures.

Many common divergences on $\mathbb{R}^m$ can actually be computed coordinate-wise by summing the corresponding distances on the respective axes. This is the case of the information divergences considered here, and it can be discussed as follows.

Definition 1.16. A *scalar divergence* is a divergence d on a set $Y \subseteq \mathbb{R}$.

Definition 1.17. A *separable divergence* is a divergence D on a set $\mathcal{Y} = Y^m$ for some set $Y \subseteq \mathbb{R}$, generated by a given scalar divergence d on Y as follows:

$$D(\mathbf{y}\|\mathbf{y}') = \sum_{i=1}^{m} d(y_i\|y_i') \quad \text{for all } \mathbf{y}, \mathbf{y}' \in \mathcal{Y} \ . \tag{1.20}$$

A separable divergence D is completely defined by the generating scalar divergence d. In the sequel, we thus concentrate on formulating such scalar divergences.

Example 1.1. The squared Euclidean distance on $\mathbb{R}$ is probably the most common example of scalar divergence:

$$d_{\mathrm{E}}(y\|y') = (y - y')^2 \ . \tag{1.21}$$

Remark 1.35. This cannot be however extended to the more general Mahalanobis distances since the covariance matrix makes the divergence non-separable in general.

Example 1.2. The Kullback-Leibler divergence on $\mathbb{R}_+^*$ provides a well-known example of scalar divergence which is asymmetric:

$$d_{\mathrm{KL}}(y\|y') = y \log \frac{y}{y'} - y + y' \ . \tag{1.22}$$

Remark 1.36. This divergence can be extended to include all $\mathbb{R}_+$ in the first argument since the limit is always finite at zero, and to include the origin $(0,0)$ on the diagonal by setting the divergence null there.

Example 1.3. The Itakura-Saito divergence on $\mathbb{R}_+^*$ provides another famous example of asymmetric separable divergence:

$$d_{\mathrm{IS}}(y\|y') = \frac{y}{y'} - \log \frac{y}{y'} - 1 \ . \tag{1.23}$$

Remark 1.37. This divergence can be extended to include the origin $(0,0)$ on the diagonal by setting the divergence null there, but cannot be extended to include all $\mathbb{R}_+^*$ in either of the arguments by considering the limits since they are not finite.

We now define wide classes of such divergences. We restrict to separable divergences on the space of discrete positive measures seen as $\mathcal{Y} = (\mathbb{R}_+^*)^m$. Comprehensive reviews of the early and the recent literatures on more general divergence measures, some of their properties, and applications to statistical machine learning and signal processing can be found in [Csiszár 1978, 2008, Basseville 1989, 2013, Cichocki and Amari 2010].

1.2.2. Csiszár divergences

We begin with the family of Csiszár divergences, which encompasses many common distance measures, such as the Kullback-Leibler and dual Kullback-Leibler divergences, the total variation distance, the Hellinger distance, or the Pearson's and Neyman's χ^2 distances. These divergences were introduced independently by Csiszár [1963], Morimoto [1963], Ali and Silvey [1966], as generic distances between probability measures. Their general properties have also been extensively studied, and a recent account is provided by Liese and Vajda [2006]. In the context of discrete positive measures, these divergences can be discussed as follows.

Definition 1.18. Let $\varphi \colon \mathbb{R}_+^* \to \mathbb{R}$ be a differentiable convex function such that $\varphi(1) = \varphi'(1) = 0$. The *Csiszár φ-divergence* is the scalar divergence $d_\varphi^{(C)}$ defined as follows:

$$d_\varphi^{(C)}(y\|y') = y\, \varphi(y'/y) \quad \text{for all } y, y' \in \mathbb{R}_+^* \ . \tag{1.24}$$

Remark 1.38. The non-negativity is a direct consequence of the convex function φ attaining its global minimum at $\varphi(1) = 0$ since $\varphi'(1) = 0$. Moreover, the identity trivially holds since we have $\varphi(y/y) = \varphi(1) = 0$, on the diagonal.

Remark 1.39. The class of Csiszár divergences is stable under swapping the arguments since we have $d_\varphi^{(C)}(y'\|y) = d_{\varphi^*}^{(C)}(y\|y')$, where $\varphi^*(y) = y\, \varphi(1/y)$ is convex, differentiable, and such that $\varphi^*(1) = \varphi^{*\prime}(1) = 0$. Moreover, the class is also stable under the convex combination of different generator functions φ. As a result, Csiszár divergences can be symmetrized or skewed straightforward while staying in the class.

The class of Csiszár divergences notably contains the well-known parametric family of α-divergences. This parametric family encompasses distance measures such as the Kullback-Leibler and dual Kullback-Leibler divergences, the Hellinger distance, or the Pearson's and Neyman's χ^2 distances. Moreover, it actually corresponds to the intersection of Csiszár and Bregman divergences [Amari 2009]. The α-divergences can be traced back to the works of Chernoff [1952] on evaluating classification errors, Rényi [1961] on generalizing the notion of entropy, and Havrda and Charvát [1967] on quantifying classification processes. It was rediscovered by Tsallis [1988] for non-extensive entropies in physics, and by Amari [1982] for information geometry and statistical inference in curved exponential families. For discrete positive measures, these divergences can be introduced as follows.

Example 1.4. An interesting parametric family of Csiszár divergences parametrized by a number $\alpha \in \mathbb{R}$ is provided by the α-divergences:

$$d_\alpha^{(a)}(y\|y') = \frac{1}{\alpha(1-\alpha)}(\alpha y + (1-\alpha)y' - y^\alpha y'^{1-\alpha}) \ . \tag{1.25}$$

Remark 1.40. For $\alpha \in \{0, 1\}$, the definition still holds by considering the limits using a Taylor series or l'Hôpital's rule, and respectively leads to the dual Kullback-Leibler divergence $d_0^{(a)}(y\|y') = y' \log(y'/y) - y' + y$, and to the Kullback-Leibler divergence $d_1^{(a)}(y\|y') = y \log(y/y') - y + y'$. Other particular cases are given by $\alpha \in \{-1, 1/2, 2\}$, leading respectively to the Neyman's χ^2, Hellinger, and Pearson's χ^2 distances.

Remark 1.41. The α-divergence is easily seen to be a Csiszár divergence for the differentiable convex function $\varphi_\alpha(y) = \frac{1}{\alpha(1-\alpha)}(\alpha + (1-\alpha)y - y^{1-\alpha})$, which is such that $\varphi_\alpha(1) = \varphi'_\alpha(1) = 0$. In the limit case $\alpha \in \{0, 1\}$, we have $\varphi_0(y) = -y + y \log y + 1$, and $\varphi_1(y) = y - \log y - 1$. The formulation of the α-divergence as a Bregman divergence is yet somewhat trickier to obtain, and the form of its generator function is technically less convenient.

1.2.3. Skew Jeffreys-Bregman divergences

We now discuss the general class of Bregman divergences and their skew Jeffreys-Bregman extension. The class of Bregman divergences was first studied by Bregman [1967] for solving convex optimization problems. These divergences encompass some well-known distance measures such as the squared Euclidean and Mahalanobis distances, or the Kullback-Leibler and Itakura-Saito divergences. For discrete positive measures, and assuming separability, these divergences can be introduced as follows.

Definition 1.19. Let $\varphi \colon \mathbb{R}_+^* \to \mathbb{R}$ be a differentiable convex function. The *Bregman φ-divergence* is the scalar divergence $d_\varphi^{(B)}$ defined as follows:

$$d_\varphi^{(B)}(y\|y') = \varphi(y) - \varphi(y') - (y - y')\varphi'(y') \quad \text{for all } y, y' \in \mathbb{R}_+^* \ . \tag{1.26}$$

Remark 1.42. The non-negativity is a direct consequence of the tangent inequality applied to the differentiable convex function φ, that is, $\varphi(y) \geq \varphi(y') + (y - y')\varphi'(y')$. Moreover, the identity trivially holds since we have $\varphi(y) - \varphi(y) - (y - y)\varphi'(y) = 0$, on the diagonal.

Remark 1.43. The class of Bregman divergences is not stable under swapping the arguments, so it is possible to create new divergences by considering symmetrized or skew versions of these divergences.

The class of Bregman divergences contains the relevant parametric family of β-divergences among others. This family notably encompasses the squared Euclidean distance, as well as the Kullback-Leibler and Itakura-Saito divergences. It was studied by Basu et al. [1998], Eguchi and Kano [2001], to robustify maximum likelihood estimation. For discrete positive measures, this parametric family can be introduced as follows.

Example 1.5. An interesting parametric family of Bregman divergences parametrized by a number $\beta \in \mathbb{R}$ is provided by the β-divergences:

$$d_\beta^{(b)}(y\|y') = \frac{1}{\beta(\beta-1)}(y^\beta + (\beta-1)y'^\beta - \beta y y'^{\beta-1}) \ . \tag{1.27}$$

Remark 1.44. For $\beta \in \{0,1\}$, the definition again holds by considering the limits using a Taylor series or l'Hôpital's rule, leading respectively to the Itakura-Saito divergence $d_0^{(b)}(y\|y') = y/y' - \log(y/y') - 1$, and to the Kullback-Leibler divergence $d_1^{(b)}(y\|y') = y\log(y/y') - y + y'$. A relevant particular case is given by $\alpha = 2$, which corresponds to the squared Euclidean distance.

Remark 1.45. The β-divergence is easily seen to be a Bregman divergence for the differentiable convex function $\varphi_\beta(y) = \frac{1}{\beta(\beta-1)}(y^\beta - \beta y + \beta - 1)$. In the limit case for $\beta \in \{0,1\}$, we have $\varphi_0(y) = y - \log y - 1$, and $\varphi_1(y) = -y + y\log y + 1$.

The natural way to skew a given Bregman divergence is by considering a convex combination of the divergence and of its swapped version. A symmetric divergence is then defined by taking the midpoint of this combination. This construction has recently been explored by Nielsen and Nock [2009b] in relation to computing information-theoretic centroids. Special instances of the construction lead to the well-known Jeffreys divergence as a symmetric version of the Kullback-Leibler divergence, as well as the cosh distance which arises from symmetrizing the Itakura-Saito divergence.

Definition 1.20. Let $\varphi\colon \mathbb{R}_+^* \to \mathbb{R}$ be a differentiable convex function, and $\lambda \in [0,1]$ be a skewing parameter. The *skew Jeffreys-Bregman* (φ,λ)*-divergence* is the scalar divergence $d_{\varphi,\lambda}^{(JB)}$ defined as follows:

$$d_{\varphi,\lambda}^{(JB)}(y\|y') = \lambda d_\varphi(y\|y') + (1-\lambda)d_\varphi(y'\|y) \quad \text{for all } y, y' \in \mathbb{R}_+^* \ . \tag{1.28}$$

In particular, for $\lambda = 1/2$, the corresponding scalar divergence $d_\varphi^{(JB)}$ is called the *Jeffreys-Bregman* φ*-divergence*.

Remark 1.46. In particular, the symmetric Jeffreys-Bregman φ-divergence simplifies as $d_\varphi^{(JB)} = (y-y')(\varphi'(y) - \varphi'(y'))/2$.

1.2.4. Skew Jensen-Bregman divergences

A second relevant way of skewing Bregman divergences exists, and is closely related to the Burbea-Rao divergences and their skew versions. The latter divergences were introduced and studied by Burbea and Rao [1982a,b] as Jensen differences in relation to entropy functions. A famous particular case of this is given by the Jensen-Shannon divergence as another symmetric version of the Kullback-Leibler divergence. The skewing of Burbea-Rao divergences and their links with Bregman divergences have recently been explored by Nielsen and Boltz [2011] for computing information-theoretic centroids, where the constructed divergences have also revealed connected, for exponential families, with the well-known Bhattacharyya distance introduced earlier by Bhattacharyya [1943].

Definition 1.21. Let $\varphi\colon \mathbb{R}_+^* \to \mathbb{R}$ be a differentiable convex function, and $\lambda \in (0,1)$ be a skewing parameter. The *skew Jensen-Bregman* (φ,λ)*-divergence* is the scalar divergence $d_{\varphi,\lambda}^{(JB')}$ defined as follows:

$$d_{\varphi,\lambda}^{(JB')}(y\|y') = \lambda d_\varphi(y\|\lambda y+(1-\lambda)y')+(1-\lambda)d_\varphi(y'\|\lambda y+(1-\lambda)y') \quad \text{for all } y,y' \in \mathbb{R}_+^* \ . \tag{1.29}$$

In particular, for $\lambda = 1/2$, the corresponding scalar divergence $d_\varphi^{(JB')}$ is called the *Jensen-Bregman* φ*-divergence*.

Remark 1.47. The limit cases $\lambda \in \{0,1\}$ are not included in the definition since they lead to a trivial null divergence.

Definition 1.22. Let $\varphi\colon \mathbb{R}_+^* \to \mathbb{R}$ be a differentiable convex function, and $\lambda \in (0,1)$ be a skewing parameter. The *skew Burbea-Rao* (φ,λ)*-divergence* is the scalar divergence $d_{\varphi,\lambda}^{(BR)}$ defined as follows:

$$d_{\varphi,\lambda}^{(BR)}(y\|y') = \lambda\varphi(y) + (1-\lambda)\varphi(y') - \varphi(\lambda y + (1-\lambda)y') \quad \text{for all } y,y' \in \mathbb{R}_+^* \ . \tag{1.30}$$

In particular, for $\lambda = 1/2$, the corresponding scalar divergence $d_\varphi^{(BR)}$ is called the *Burbea-Rao* φ*-divergence*

Remark 1.48. The limit cases $\lambda \in \{0,1\}$ are again excluded from the definition to avoid trivial null divergences.

Remark 1.49. The skew Jensen-Bregman (φ,λ)-divergence and the skew Burbea-Rao (φ,λ)-divergence coincide, $d_{\varphi,\lambda}^{(JB')} = d_{\varphi,\lambda}^{(BR)}$. In particular, the Jensen-Bregman φ-divergence and the Burbea-Rao φ-divergence coincide, $d_\varphi^{(JB')} = d_\varphi^{(BR)}$. Setting $y_\lambda = \lambda y + (1-\lambda)y'$, the equivalence can be seen as follows:

$$\begin{aligned} d_{\varphi,\lambda}^{(JB')}(y\|y') &= \lambda d_\varphi(y\|y_\lambda) + (1-\lambda)d_\varphi(y'\|y_\lambda) && (1.31)\\ &= \lambda(\varphi(y) - \varphi(y_\lambda) - (y-y_\lambda)\varphi'(y_\lambda)) \\ &\qquad + (1-\lambda)(\varphi(y') - \varphi(y_\lambda) - (y'-y_\lambda)\varphi'(y_\lambda)) && (1.32)\\ &= d_{\varphi,\lambda}^{(BR)}(y\|y') - (\lambda y + (1-\lambda)y' - y_\lambda)\varphi'(y_\lambda) && (1.33)\\ &= d_{\varphi,\lambda}^{(BR)}(y\|y') \ . && (1.34) \end{aligned}$$

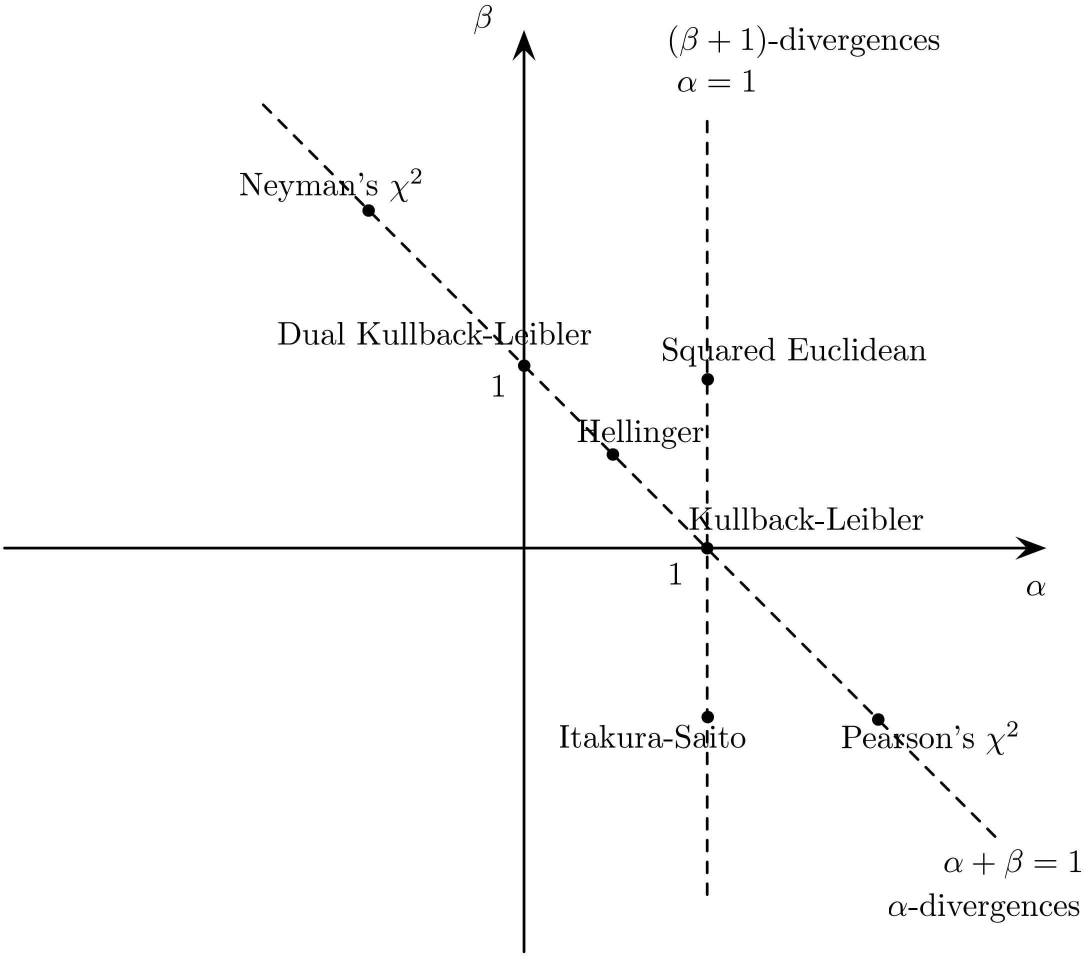

Figure 1.2.: Parametric family of (α, β)-divergences. The class of (α, β)-divergences encompasses many common information divergences, including all α-divergences and β-divergences.

Remark 1.50. Since the class of skew Burbea-Rao divergences is clearly stable under swapping the arguments, which amounts to replacing λ with $1-\lambda$, the class of skew Jensen-Bregman divergences is also, which is not obvious at first sight from their definition.

1.2.5. Skew (α, β, λ)-divergences

Recently, in the context of non-negative matrix factorization, Cichocki et al. [2011] proposed an elegant parametrization of a class of scalar divergences that encompasses both α-divergences and β-divergences among others, as shown in Figure 1.2. Furthermore, this family is potentially robust against noise and outliers, because it combines the respective scaling properties of the α-divergences and β-divergences, hence providing both zooming and weighting factors that can be tuned to improve estimation. This parametric family of (α, β)-divergences can be introduced as follows.

Definition 1.23. Let $\alpha, \beta \in \mathbb{R}$ be scalar parameters. The (α, β)-*divergence* is the scalar divergence $d^{(ab)}_{\alpha,\beta}$ defined as follows:

$$d^{(ab)}_{\alpha,\beta}(y\|y') = \frac{1}{\alpha\beta(\alpha+\beta)}(\alpha y^{\alpha+\beta} + \beta y'^{\alpha+\beta} - (\alpha+\beta) y^{\alpha} y'^{\beta}) \quad \text{for all } y, y' \in \mathbb{R}^*_+ \ . \tag{1.35}$$

Remark 1.51. The non-negativity of (α, β)-divergences can be proved with Young's inequality, for three different combinations of the signs of $\alpha\beta, \alpha(\alpha+\beta), \beta(\alpha+\beta)$. The function vanishing on the diagonal holds trivially.

Remark 1.52. As soon as either of α or β is null, the definition still holds in the respective limit cases $d^{(ab)}_{\alpha,0}(y\|y') = \frac{1}{\alpha^2}(y^{\alpha}\log(y^{\alpha}/y'^{\alpha}) - y^{\alpha} + y'^{\alpha})$ for $\alpha \neq 0$, and $d^{(ab)}_{0,\beta}(y\|y') = \frac{1}{\beta^2}(y'^{\beta}\log(y'^{\beta}/y^{\beta}) - y'^{\beta} + y^{\beta})$ for $\beta \neq 0$. When $\alpha + \beta = 0$, the definition is also valid with the limits $d^{(ab)}_{\alpha,-\alpha}(y\|y') = \frac{1}{\alpha^2}(\log(y'^{\alpha}/y^{\alpha}) + y^{\alpha}/y'^{\alpha} - 1)$ for $\alpha \neq 0$, and $d^{(ab)}_{0,0}(y\|y') = \frac{1}{2}(\log y - \log y')^2$.

Remark 1.53. As special cases, the (α, β)-divergences reduce to the α-divergences $d^{(ab)}_{\alpha,\beta} = d^{(a)}_{\alpha}$ for $\alpha + \beta = 1$, and to the β-divergences $d^{(ab)}_{\alpha,\beta} = d^{(b)}_{\beta+1}$ for $\alpha = 1$.

We finally introduce a direct but novel parametric family of information divergences as an extension of (α, β)-divergences by standard skewing.

Definition 1.24. Let $\alpha, \beta \in \mathbb{R}$ be scalar parameters, $\lambda \in [0, 1]$ be a skewing parameter. The skew (α, β, λ)-*divergence* is the scalar divergence $d^{(ab)}_{\alpha,\beta,\lambda}$ defined as follows:

$$d^{(ab)}_{\alpha,\beta,\lambda}(y\|y') = \lambda d^{(ab)}_{\alpha,\beta}(y\|y') + (1-\lambda) d^{(ab)}_{\alpha,\beta}(y'\|y) \quad \text{for all } y, y' \in \mathbb{R}^*_+ \ . \tag{1.36}$$

These divergences notably encompass all α-divergences and β-divergences, as well as their skew Jeffreys-Bregman versions. They will also reveal useful in the sequel to unify the results for non-negative matrix factorization based on these divergences, as done by Cichocki et al. [2011] for the non-skew versions.

2. Sequential Change Detection with Exponential Families

In this chapter, we elaborate methods for sequential change detection within the computational framework of information geometry. We consider a standard generic framework for sequential change detection applied to exponential families. This framework therefore encompasses many common statistical models as discussed in Chapter 1. More precisely, we follow a non-Bayesian approach where change detection is considered as a statistical decision problem with multiple hypotheses, and is solved using generalized likelihood ratio test statistics. A major drawback of previous sequential approaches in this context is to consider only known parameters before change, or to approximate the exact statistics when these parameters are actually unknown. This is addressed by employing exact generalized likelihood ratios with arbitrary estimators, and by expanding them for exponential families. By showing tight links between the computation of these statistics and maximum likelihood estimates, we derive a generic scheme for change detection with exponential families, under common scenarios with known or unknown parameters, and arbitrary estimators. We also interpret this scheme within the dually flat information geometry of exponential families, hence providing both statistical and geometrical intuitions to the problem, and bridging the gap between statistical and distance-based approaches to change detection. The scheme is finally revisited through convex duality, leading to an attractive procedure with closed-form sequential updates for the exact generalized likelihood ratio statistics, when both the parameters before and after change are unknown and are estimated by maximum likelihood. This is notably applied in Chapter 4 to devise a general and unifying framework for real-time audio segmentation.

2.1. Context

In this section, we first provide some background information on the problem of change detection, with focus on sequential approaches. We then discuss the motivations of our approach to this problem. We finally sum up our main contributions in this context.

2.1.1. Background

Let us consider a time series $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of observations that are sampled according to an unknown discrete-time stochastic process. In general terms, the problem of *change detection* is to decide whether the distribution of the process presents some structural modifications of interest along time, as depicted in Figure 2.1. This

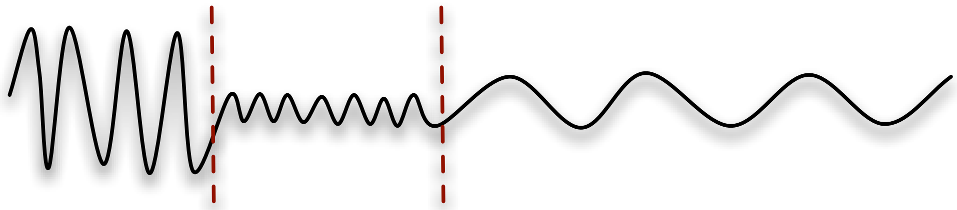

Figure 2.1.: Schematic view of change detection. The problem of change detection consists in finding variations of interest within the temporal structure of a process.

decision is often coupled with the estimation of the times when changes in the distribution occur. These time instants are called *change points* and delimit contiguous temporal regions called *segments*. In addition to estimating the change points, we sometimes also need to estimate the underlying distributions within the different segments.

Historically, change detection arose as a sequential problem in the area of quality control, with the *control charts* of Shewhart [1925, 1931]. The formulations of change detection have primarily focused on statistical frameworks, with consideration of a single change point and known distributions before and after change, by using *likelihood ratio* (LR) statistics. The first main approaches were the Bayesian methods introduced by Girshick and Rubin [1952], and the non-Bayesian procedures such as the *cumulative sum* (CUSUM) and the *finite moving average* charts of Page [1954], as well as the *geometric moving average* charts of Roberts [1959].

Later on, Shiryaev [1963, 1978] and Roberts [1966] proved some optimality properties of the sequential Bayesian detection rule with a geometric prior over the change time, hence known as *Shiryaev-Roberts* (SR) rule. This was also shown optimal in an asymptotic context by Pollak [1985]. In the meantime, Lorden [1971] discussed results on the asymptotic optimality of the non-Bayesian CUSUM rule, and introduced the *generalized likelihood ratio* (GLR) statistics to replace the LR statistics in CUSUM when the parameter after change is unknown and the distributions belong to a one-parameter exponential family. Optimality results were later proved in a non-asymptotic context by Moustakides [1986] and Ritov [1990]. As an alternative to the GLR statistics, Pollak and Siegmund [1975] introduced a weighted CUSUM rule using *mixture likelihood ratio* (MLR) statistics, also known as *weighted likelihood ratio* (WLR) statistics. These statistics were further used by Pollak [1987] to extend the Bayesian SR detection rule.

Many of these sequential approaches focused on detecting an additive change point in the mean of some independent univariate data under normality assumptions. Since then, several hundreds of papers have proposed specific extensions to relax these assumptions. We refer to the seminal book of Basseville and Nikiforov [1993], and paper of Lai [1995], for a comprehensive review and unifying framework. The recent books of Poor and Hadjiliadis [2009], and Chen and Gupta [2012], provide more up-to-date accounts respectively on sequential and retrospective approaches. The recent paper of Polunchenko and Tartakovsky [2012] presents the state-of-the-art results on optimal procedures for sequential change detection with known parameters before and after change. Some complementary viewpoints are treated in the books of Brodsky and Darkhovsky [1993], Csörgő and Horváth [1997], and Gustafsson [2000], with respective focus on non-parametric methods for change detection, asymptotic behaviors of change detection procedures, and change detection in adaptive filters. The field is still in active research today, and a forthcoming book on the topic has been written by leading researchers [Basseville et al. 2013]. We can sum up the main distinctions between the different approaches as follows.

The principal distinction, because of both theoretical and philosophical issues, is between non-Bayesian and Bayesian approaches. This distinction historically lies in the consideration of the unknown change points either as deterministic or as random quantities. On the one hand, in non-Bayesian approaches, a change time is assumed to be an unknown parameter, and the detection is roughly speaking based on the likelihood of a change compared to no change. On the other hand, in Bayesian approaches, a change time is assumed to be a latent variable with a given prior probability distribution, and the detection is rather based on the posterior probability of a change. In later approaches, this distinction also lies on whether the unknown parameters, if any, are considered as deterministic or as random quantities. In non-Bayesian approaches, the detection of a change relies on the point estimation of the unknown parameters, whereas it relies on their marginalization using suitable priors in Bayesian approaches.

Other important distinctions can be made depending on the problem assumptions. Relevant examples include sequential versus retrospective settings, single versus multiple change points, additive versus non-additive changes, known versus unknown distributions before or after change, univariate versus multivariate data, continuous versus discrete data, independent versus non-independent observations, parametric versus non-parametric distributions, scalar versus vector parameters.

Concerning optimality, the two principal criteria are the *average detection delay* and the *false alarm rate*. The average detection delay is related to the latency of the system and quantifies the time lag between a change point and the *alarm time* at which it is detected, while the false alarm rate is related to the robustness of the system and quantifies the number of alarms triggered wrongly before a change really occurs. Another less used criterion is the *misdetection rate* which is related to the sensibility of the system and quantifies the number of occurring changes that are missed. Often, optimality is considered by fixing the false alarm rate and minimizing the average detection delay, leading to methods for *quickest detection*. It is yet intuitive that, depending on the application at hand, a compromise has to be found between improving the average detection delay and decreasing the detection errors of false alarms and misdetections.

The applicative fields of change detection are numerous and various in nature. In addition to quality control in industrial production [Wetherill and Brown 1991], applications include fault detection in technological processes [Willsky 1976], or automatic surveillance for intrusion and abnormal behavior in security monitoring [Tartakovsky et al. 2006], and more generally many problems in signal processing [Basseville and Benveniste 1983b, Basseville 1988]. In this context change detection has been applied to data from various domains such as geophysics [Basseville and Benveniste 1983a], econometrics [Broemeling and Tsurumi 1987], audio [André-Obrecht 1988], medicine [Sonesson and Bock 2003], image [Radke et al. 2005].

Modern approaches to change detection have also intersected several techniques in machine learning, with online and offline algorithms for solving respectively the sequential and retrospective change detection problems. Some common employed techniques in this context are kernel methods, support vector machines, and convex optimization. Most related algorithms for change detection then consider some notion of distance in order to define either a cost function for measuring and optimizing the quality of the segmentation [Harchaoui and Lévy-Leduc 2008, 2010, Vert and Bleakley 2010], or a dissimilarity measure in a high-dimensional feature space for determining and thresholding the amount of novelty between successive windows of data [Desobry et al. 2005, Harchaoui et al. 2009a]. The latter methods can actually be linked with CUSUM schemes by using exponential families and reproducing kernel Hilbert spaces [Canu and Smola 2006].

2.1.2. Motivations

The ubiquity of change detection techniques in various applicative fields highlights the interests in providing generic methods that can handle data of heterogeneous types. In many approaches, however, either the distributions before and after change are assumed to be completely known in advance, or particular statistical models are employed for the unknown parameters, most of the time normal distributions, and the procedures are derived specifically for these models. Alternatives do exist, with non-parametric approaches, as well as parametric approaches based on generic families of distributions, notably with assumptions of independent observations in one-parameter exponential families. We concentrate on the latter approach, in the light of computational information geometry with general exponential families.

We thus seek to formulate a unifying and generic framework for statistical change detection in a times series of independent observations drawn from an exponential family. We try to encompass different scenarios where scalar or vector parameters before and after change can be known or unknown, with additive or non-additive changes, using univariate or multivariate, and continuous or discrete data, indifferently. We also aim at bridging the gap between classic statistical and contemporary machine learning approaches to change detection, by showing tight links between the statistical models involved and some associated distance functions.

Another motivation of our approach comes from the design of online methods for change detection. In the literature, change detection is still often addressed in a sequential rather than retrospective setting, that is, the time series is considered as a data stream that unfolds in time and is processed incrementally. This criterion is vital in contexts where causality is mandatory, such as real-time applications where one does not have access to the future. Yet a causal design may also be relevant in other contexts, not only to keep computations tractable when dealing with a large amount of data, but also to account for the inherent temporal flow of the time series. In other words, change detection may be viewed as finding a sufficient amount of novelty, a rupture of information content between a given time point and its relative past. We therefore focus here on online methods for change detection.

In general, sequential procedures are designed to detect a single change point in the incoming data stream. When multiple change points need to be detected, the following scheme is employed. We start with an empty window $\vec{\mathbf{x}} \leftarrow ()$ and process the available data incrementally. At each time increment $n = 1, 2, \dots,$ we concatenate the incoming observation $\mathbf{x}_n$ with the previous ones as $\vec{\mathbf{x}} \leftarrow \vec{\mathbf{x}} \,|\, \mathbf{x}_n$, and attempt to detect a change. If a change is detected, then we discard the observations before the estimated change point i, and restart the procedure with an initial window $\vec{\mathbf{x}} \leftarrow (\mathbf{x}_{i+1}, \dots, \mathbf{x}_n)$. The differences between approaches generally lie in computational heuristics such as using minimum and maximum window sizes, as well as window growing and sliding factors before attempting to detect a change. In these approaches, the sequential detection of multiple change points can therefore be reduced to the detection of a single change point in a given data sample $\bar{\mathbf{x}} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ of size $n > 1$.

We notice, however, that this problem reduction is disputable. First, it requires the precise estimation of the different change points, and does not take into account the uncertainty about these estimated change points. Second, it supposes that if no change point has been detected in the current window yet, then adding some extra sample observations may only introduce one change point. This is a reasonable assumption in general but it does not take into account the possibility that a change point has been missed, or that several change points occur in the added observations. This may occur when the sampling distributions before and after change are very similar and not enough sample observations are available to discriminate between them, or when a small drift in the sampling distribution occurs. In such situations, one would need to consider several change points or model the drift in some way.

Nevertheless, we focus on the widespread framework of *abrupt change detection* where considerations on smooth changes such as drifts are left aside. We also assume that the change points are detected fast enough so that the alarm times are triggered before other changes occur. It permits to employ the standard sequential schemes discussed above. A noticeable advantage of these schemes is that of discarding completely the past information when a change point is detected. It provides an important computational advantage over methods that would require storing some past information to deal with multiple change points and uncertainty in their estimation.

Finally, we concentrate on approaches similar to CUSUM detection rules with LR statistics, and their extensions with GLR statistics for unknown parameters. A major theoretical issue of previous sequential approaches in this context is to consider only known parameters before change. This is suitable for applications such as quality control where a normal regime is completely known in advance, but this is limited in many other real-world applications. The problem when considering unknown parameters before change, is that it breaks down the recursivity and computational efficiency of CUSUM schemes. Therefore, some approximations of the exact GLR test statistics are in general made to accommodate these situations, such as learning the distribution before change in a training set of samples, or estimating it directly at once for all hypotheses, either on the whole window, or in a dead region at its beginning where change detection is turned off. These approximate GLR statistics, however, result in practical shortcomings as soon as changes occur too rapidly, because of the estimation errors.

A few particular exact statistics have yet been studied. For example, Siegmund and Venkatraman [1995] considered the exact GLR statistics for unknown mean before and after change in univariate normals with a fixed and known variance, which relies on a specific invariance by translation. Recently, Mei [2006] proposed a framework for unknown parametric distributions with a compact parameter space before and after change in the case of independent observations, by using a point estimation before change, and a mixing prior after change, with no prior distribution on the change point. It seems however more natural to consider either a full non-Bayesian or a full Bayesian framework, as noted by Lai and Xing [2010] who developed a full Bayesian framework for sequential change detection with unknown parameters before and after change, based on the derivation of convenient expressions of the MLR statistics for exponential families, and proved some asymptotic optimality results of this procedure under the assumption of a geometric prior over change.

Nevertheless, this Bayesian framework is not suitable to all applications. Indeed, it first requires some expert knowledge or some training data to learn the distribution of the parameters in a supervised fashion before performing change detection. Such prior knowledge or data are unfortunately not always available. Moreover, the assumption of a geometric prior over change is not well-suited to all signals, which may exhibit more complicated distribution profiles for time intervals between change points. To overcome this, we seek to employ sequential change detection schemes without any a priori on the respective distributions of the change points and parameters.

We therefore position in this continuity but rather develop non-Bayesian methods for change detection with independent observations from an exponential family, when both the parameters before and after change may be unknown. Nevertheless, we do not discuss further theoretical optimality here and redirect instead the interested reader to the given references and citations therein. The problem, in our opinion, is that there is no widely accepted consensus on how to define exactly optimality, not only depending on Bayesian and non-Bayesian approaches, but also on asymptotic and non-asymptotic contexts, or even known and unknown parameters. Moreover, the theoretical optimality results may not always corroborate practical situations since the data are often not distributed exactly according to the models considered.

2.1.3. Contributions

To sum up our positions in the field of change detection, we follow a standard non-Bayesian approach for detecting changes in a sequence of mutually independent observations from a dominated statistical model. In this approach, change detection is seen as a decision problem with multiple statistical hypotheses, and is solved by employing GLR test statistics. In sequential schemes, approximations of the exact GLR test statistics are most of the time employed for tractability issues. Our main contributions in this context can be summarized as follows.

We apply the change detection framework to exponential families, and notably study the GLR test statistics within the dually flat information geometry. To the best of our knowledge, this is the first time change detection is studied in the light of information geometry, hence providing some geometrical intuitions to this statistical problem. Because of the known correspondence between exponential families and their associated Bregman divergences, these intuitions provides a unified view of change detection for many common statistical models and corresponding distance functions, and bridges the gap between certain statistical and distance-based approaches to change detection. We also generalize the estimation in the GLR test statistics to arbitrary estimators on top of commonly used maximum likelihood estimators. As a technical by-product, it allows the unification of different scenarios with known or unknown parameters before and after change, that are in general treated separately. In this context, the GLR statistics with arbitrary estimators are actually shown to be intrinsically linked with maximum likelihood estimation, where the chosen estimators behave as corrective terms compared to the maximum likelihood estimators. We also revisit the statistics through convex duality, which provides further evidence of the corrective role of the chosen estimators compared to maximum likelihood. Last but not least, it leads to an alternative expression for the exact GLR test statistics, which simplifies its computation when both parameters are unknown and are estimated by maximum likelihood. The derived expression is obtained in closed form in terms of the conjugate of the log-normalizer for the exponential family. Moreover, it can be updated sequentially, hence providing a computationally efficient scheme for exact inference about change detection in exponential families when both the parameters before and after change are unknown. The remainder of this chapter is organized as follows.

In Section 2.2, we review the standard non-Bayesian framework for change detection in a sequence of mutually independent observations from a dominated statistical model. In Subsection 2.2.1, we formalize the change detection problem as a multiple hypothesis problem. In Subsection 2.2.2, we define the statistical decision rule based on the common likelihood ratio test statistic, or on its generalized version for unknown parameters by introducing arbitrary estimators in the statistic. In Section 2.3, we apply the formulated framework to derive methods for change detection in exponential families. In Subsection 2.3.1, we notably develop the GLR test statistics with arbitrary estimators and actually show that it is intrinsically linked with the maximum likelihood estimators, which behave as corrective terms compared to the chosen estimators. In Subsections 2.3.2 and 2.3.3, we then obtain specific forms of the generic scheme under common scenarios by considering different combinations of estimators for the respective parameters. In particular, we expand the form of the known LR statistics when both the parameters before and after change are known, and of the standard GLR statistics when the parameter before change only is known. We also compare the standard approximate GLR and the exact GLR statistics when both parameters are unknown. In Subsection 2.3.4, we revisit the proposed generic updates through convex duality for exponential families, by reparametrizing the problem from the natural to the expectation parameter space. It provides further evidence for the corrective role of the chosen estimators compared to maximum likelihood in the GLR statistics. In Subsection 2.3.5, we finally derive a simplified expression for the exact GLR statistics when both parameters are unknown and are estimated by maximum likelihood.

2.2. Statistical framework

In this section, we formalize a standard statistical framework for sequential change detection. The detection of a change point in a given data sample is seen as a decision problem with multiple hypotheses. To solve this problem, we then introduce the common test statistics of likelihood ratio for known parameters and the corresponding non-Bayesian decision rule, as well as their proposed extension through the generalized likelihood ratio test statistics with arbitrary estimators.

2.2.1. Multiple hypothesis problem

To unify the problem formulation and discussion, we restrict to the widespread case where the observations are independently sampled according to distributions from a given dominated parametric statistical model.

Problem 2.1. Let $\mathcal{P} = \{P_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \Xi}$ be a dominated parametric statistical model on a measurable space $(\mathcal{X}, \mathcal{A})$, and let $X_1, \ldots, X_n$ be $n > 1$ mutually independent random variables that are distributed according to probability distributions from $\mathcal{P}$. The problem of *change detection* is to decide, on the basis of sample observations $\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^n$, whether the random variables $X_1, \ldots, X_n$ are identically distributed or not.

Remark 2.1. This problem and subsequent derivations can be formulated straightforward with adequate notational changes for non-parametric models. We state it in the parametric case for convenience, and for its direct application to the parametric models of exponential families later.

Remark 2.2. The formulation can also be extended to non-independent observations with more technical efforts. This is done by introducing the filtered probability space with the natural filtration associated to the stochastic process, that is, based on the increasing sequence of σ-algebras generated by the accumulated random variables. We then consider conditional distributions on the accumulated past observations and the observations in a segment are not necessarily identically distributed, but the parameter of interest $\boldsymbol{\xi}$ does not change.

Remark 2.3. The assumption that the model is dominated can also be dropped from the problem statement. It is however essential to the subsequent derivations of statistics based on probability densities.

As discussed previously, we suppose that there is at most one change point. Hence, the problem of change detection can be seen as a statistical decision between the null hypothesis of no change against the alternative hypothesis of one change.

Definition 2.1. Let Ξ_0, Ξ_0^i, Ξ_1^i be subsets of Ξ, for any $1 \leq i \leq n-1$. The *hypothesis of no change* and the *hypothesis of a change* are respectively the null and the alternative statistical hypotheses H_0 and H_1 defined as follows:

$$H_0 : \; X_1, \ldots, X_n \sim P_{\boldsymbol{\xi}_0}, \; \boldsymbol{\xi}_0 \in \Xi_0 \,, \tag{2.1}$$

$$H_1 : \; X_1, \ldots, X_i \sim P_{\boldsymbol{\xi}_0^i}, \; \boldsymbol{\xi}_0^i \in \Xi_0^i \,, \; X_{i+1}, \ldots, X_n \sim P_{\boldsymbol{\xi}_1^i}, \; \boldsymbol{\xi}_1^i \in \Xi_1^i \,, \; i \in [\![1, n-1]\!] \;. \tag{2.2}$$

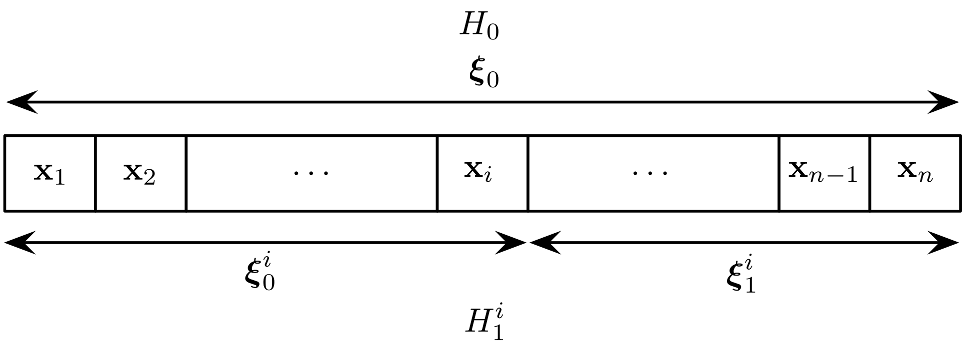

Figure 2.2.: Multiple hypotheses for a change point. The problem of change detection can be seen as comparing the plausibility of the respective hypotheses that changes occur at the different time instants, with the hypothesis that no change occurs at all.

Remark 2.4. We notice in this definition that the alternative hypothesis H_1 may encompass the null hypothesis H_0, depending on the subsets Ξ_0, Ξ_0^i, Ξ_1^i. Indeed, we do not require explicitly that $P_{\boldsymbol{\xi}_0^i} \neq P_{\boldsymbol{\xi}_1^i}$. Nevertheless, this is not an issue because the decision is not based exactly on whether the alternative is more plausible than the null hypothesis, but rather on to what extent it is more plausible, similarly to a model selection problem with nested models.

When a change occurs, we also need to estimate the change point. We therefore partition the hypothesis of a change into multiple hypotheses of a change at the respective time points. The different hypotheses of no change, and of a change at the respective time points, are illustrated in Figure 2.2.

Definition 2.2. The *hypothesis of a change at time* i, for some $1 \leq i \leq n-1$, is the alternative statistical hypothesis H_1^i defined as follows:

$$H_1^i : X_1, \ldots, X_i \sim P_{\boldsymbol{\xi}_0^i}, \ \boldsymbol{\xi}_0^i \in \Xi_0^i \ , \ X_{i+1}, \ldots, X_n \sim P_{\boldsymbol{\xi}_1^i}, \ \boldsymbol{\xi}_1^i \in \Xi_1^i \ . \tag{2.3}$$

Remark 2.5. For notational reasons, we employ the convention that the change points refer to the last points of the respective segments rather than the first points of the subsequent segments.

Remark 2.6. In certain scenarios, the parameters before and after change may be completely known in advance, being equal respectively to $\boldsymbol{\xi}_{\text{bef}}$ and $\boldsymbol{\xi}_{\text{aft}}$. In this situation, all hypotheses are simple since we have $\boldsymbol{\xi}_0, \boldsymbol{\xi}_0^i \in \{\boldsymbol{\xi}_{\text{bef}}\}$, and $\boldsymbol{\xi}_1^i \in \{\boldsymbol{\xi}_{\text{aft}}\}$. Nevertheless, the parameters before and after change are most of the time unknown, thus making some hypotheses to be composite. In the general case, all subsets Ξ_0, Ξ_0^i, Ξ_1^i, may be chosen to be different. Often, the subsets before change, respectively after change, are equal either to Ξ itself, or to a singleton if the corresponding parameter is completely known. These subsets act as a priori information about the problem, and allow the unification of different scenarios with known and unknown parameters which are in general treated separately in the literature.

2.2.2. Test statistics and decision rules

To assess the plausibility of the alternative hypotheses compared to the null hypothesis, some test statistics are needed. The aim of these statistics is to quantify to what extent a hypothesis is more plausible than another. The assumption that the models considered are dominated is crucial for this since it permits to employ the corresponding probability densities to construct the test. In this context, most test statistics rely in informal terms on the ratio $p(\bar{\mathbf{x}}|H_0)/p(\bar{\mathbf{x}}|H_1^i)$ between the plausibility of the data under the respective hypotheses. The different formulations depend on whether the parameters are known or unknown, and how the unknown parameters are dealt with. When the parameters before and after change are known in advance, we can directly use the respective likelihoods of the data under the different hypotheses.

Definition 2.3. Suppose that the sets Ξ_0, Ξ_0^i, Ξ_1^i are singletons, for any $1 \leq i \leq n-1$. The *likelihood ratio at time i* is the test statistic Λ^i defined as follows:

$$\Lambda^i(\bar{\mathbf{x}}) = -2 \log \frac{\prod_{j=1}^{n} p_{\boldsymbol{\xi}_0}(\mathbf{x}_j)}{\prod_{j=1}^{i} p_{\boldsymbol{\xi}_0^i}(\mathbf{x}_j) \prod_{j=i+1}^{n} p_{\boldsymbol{\xi}_1^i}(\mathbf{x}_j)} \quad \text{for all } \bar{\mathbf{x}} \in \mathcal{X}^n \ . \tag{2.4}$$

Remark 2.7. The LR statistic Λ^i vanishes whenever the likelihoods under H_0 and H_1^i are equal, meaning informally that they are equally plausible. Moreover, Λ^i is positive when the likelihood under H_0 is less than under H_1^i, meaning that H_0 is less plausible than H_1^i. Conversely, Λ^i is negative when the likelihood under H_0 is greater than under H_1^i, so that H_0 is more plausible than H_1^i.

Remark 2.8. As limit cases, when the likelihood under H_0, respectively H_1^i, is null, Λ^i equals $+\infty$, respectively $-\infty$. In the indeterminate case where both likelihoods under H_0 and H_1^i are null, it is convenient to use the convention that Λ^i equals 0 since H_0 and H_1^i are both equally non-plausible.

Remark 2.9. In the usual case where the parameters $\boldsymbol{\xi}_{\text{bef}}, \boldsymbol{\xi}_{\text{aft}}$, before and after change are completely known in advance, we have $\Xi_0 = \Xi_0^i = \{\boldsymbol{\xi}_{\text{bef}}\}$, and $\Xi_1^i = \{\boldsymbol{\xi}_{\text{aft}}\}$, for all $1 \leq i \leq n-1$, and the LR simplifies to the common cumulative sum statistic employed in the CUSUM procedure:

$$\frac{1}{2} \Lambda^i(\bar{\mathbf{x}}) = \sum_{j=i+1}^{n} \log \frac{p_{\boldsymbol{\xi}_{\text{aft}}}(\mathbf{x}_j)}{p_{\boldsymbol{\xi}_{\text{bef}}}(\mathbf{x}_j)} \ . \tag{2.5}$$

These statistics can in general be computed efficiently with a sequential update scheme, making the CUSUM algorithm an attractive online procedure.[1]

When a parameter is unknown, the likelihood under the corresponding composite hypothesis cannot be defined anymore and other test statistics are thus required. The usual non-Bayesian approach is to replace the unknown parameters in the hypotheses with their maximum likelihood estimates, and to write the generalization of the LR corresponding to these fixed parameters. We generalize this approach by considering arbitrary estimates. It permits to unify the different combinations of

[1] Actually, only the maximum of these statistics is needed, which can be computed recursively.

known and unknown parameters before and after change, as well as approximations of the test statistics as discussed later, and to employ any other estimator than the maximum likelihood estimator when needed.

Definition 2.4. Let $\widehat{\boldsymbol{\xi}}_0, \widehat{\boldsymbol{\xi}}_0^i, \widehat{\boldsymbol{\xi}}_1^i : \mathcal{X}^n \to \Xi$, be estimators of the parameters $\boldsymbol{\xi}_0, \boldsymbol{\xi}_0^i, \boldsymbol{\xi}_1^i$, for any $1 \leq i \leq n-1$. The *generalized likelihood ratio at time i* is the test statistic $\widehat{\Lambda}^i$ defined as follows:

$$\widehat{\Lambda}^i(\bar{\mathbf{x}}) = -2 \log \frac{\prod_{j=1}^n p_{\widehat{\boldsymbol{\xi}}_0(\bar{\mathbf{x}})}(\mathbf{x}_j)}{\prod_{j=1}^i p_{\widehat{\boldsymbol{\xi}}_0^i(\bar{\mathbf{x}})}(\mathbf{x}_j) \prod_{j=i+1}^n p_{\widehat{\boldsymbol{\xi}}_1^i(\bar{\mathbf{x}})}(\mathbf{x}_j)} \quad \text{for all } \bar{\mathbf{x}} \in \mathcal{X}^n \ . \tag{2.6}$$

Remark 2.10. To simplify notations, we consider estimators for $\bar{\mathbf{x}} \in \mathcal{X}^n$, and conflate the estimators of the individual sample models for $\mathbf{x}_j \in \mathcal{X}$, with the estimators of the i.i.d. sampling models for $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^n$, $(\mathbf{x}_1, \ldots, \mathbf{x}_i) \in \mathcal{X}^i$, $(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_n) \in \mathcal{X}^{n-i}$. We also allow arbitrary estimators in Ξ since it simplifies the following discussion without loss of generality. This is up to the user to choose estimators in the respective correct subsets Ξ_0, Ξ_0^i, Ξ_1^i when needed.

Remark 2.11. Rather than using estimators for the unknown parameters, the common Bayesian alternative to the GLR is to integrate out the unknown parameters using prior measures on the parameter space as in the MLR. Nonetheless, it requires some prior knowledge on the distribution of the parameters, which is not always available. Moreover, the computations may become intractable unless specific conjugate prior measures are employed, which may not always represent reliably the true distribution of the parameters.

Remark 2.12. In the general case, the GLR is a sum of two cumulative sums:

$$\frac{1}{2} \widehat{\Lambda}^i(\bar{\mathbf{x}}) = \sum_{j=1}^{i} \log \frac{p_{\widehat{\boldsymbol{\xi}}_0^i(\bar{\mathbf{x}})}(\mathbf{x}_j)}{p_{\widehat{\boldsymbol{\xi}}_0(\bar{\mathbf{x}})}(\mathbf{x}_j)} + \sum_{j=i+1}^{n} \log \frac{p_{\widehat{\boldsymbol{\xi}}_1^i(\bar{\mathbf{x}})}(\mathbf{x}_j)}{p_{\widehat{\boldsymbol{\xi}}_0(\bar{\mathbf{x}})}(\mathbf{x}_j)} \ . \tag{2.7}$$

As a special case, the GLR coincides with the LR when the parameters before and after change are known and thus taken as the estimates. When the parameter before change only is known, the GLR still simplifies to cumulative sums:

$$\frac{1}{2} \widehat{\Lambda}^i(\bar{\mathbf{x}}) = \sum_{j=i+1}^{n} \log \frac{p_{\widehat{\boldsymbol{\xi}}_1^i(\bar{\mathbf{x}})}(\mathbf{x}_j)}{p_{\boldsymbol{\xi}_{\mathrm{bef}}}(\mathbf{x}_j)} \ . \tag{2.8}$$

This expression is yet computationally more demanding than the LR. Indeed, the estimates after change need to be computed and may differ for all hypotheses, and for all successive windows. Therefore, there is no general known scheme for computing these statistics efficiently in a sequential fashion.[2]

Remark 2.13. When the parameter before change is unknown, the GLR cannot be written with only one cumulative sum anymore, even if the parameter after change is known. Its expression thus becomes even more expensive to compute. This is why it is in general approximated, by assuming the parameter before change known, while actually estimating it at once for all hypotheses, either on the whole data sample, or in a dead region at the beginning of the window where no change is assumed to occur.

[2]Moreover, the maximum of the statistics is in general not recursive anymore.

Remark 2.14. When the maximum likelihood estimates of the parameters in the hypotheses exist and are taken as the estimates, the GLR as defined here for arbitrary estimates specializes to the usual definition of an exact GLR statistic:

$$\frac{1}{2}\widehat{\Lambda}^i(\bar{\mathbf{x}}) = \log \frac{\sup_{\boldsymbol{\xi}_0^i \in \Xi_0^i} \prod_{j=1}^{i} p_{\boldsymbol{\xi}_0^i}(\mathbf{x}_j) \ \sup_{\boldsymbol{\xi}_1^i \in \Xi_1^i} \prod_{j=i+1}^{n} p_{\boldsymbol{\xi}_1^i}(\mathbf{x}_j)}{\sup_{\boldsymbol{\xi}_0 \in \Xi_0} \prod_{j=1}^{n} p_{\boldsymbol{\xi}_0}(\mathbf{x}_j)} \ . \tag{2.9}$$

In particular, when the parameter before change is completely known in advance, or is assumed to be known while actually being roughly determined, it leads to the standard GLR statistics employed in practice:

$$\frac{1}{2}\widehat{\Lambda}^i(\bar{\mathbf{x}}) = \sup_{\boldsymbol{\xi}_1^i \in \Xi_1^i} \sum_{j=i+1}^{n} \log \frac{p_{\boldsymbol{\xi}_1^i}(\mathbf{x}_j)}{p_{\boldsymbol{\xi}_{\mathrm{bef}}}(\mathbf{x}_j)} \ . \tag{2.10}$$

This test statistic is however an approximation of the exact GLR statistics as soon as the parameter before change is unknown.

Based on the chosen test statistics, we eventually need a decision rule to trigger a change point or not. Since the GLR as defined here encompasses the LR and classic or approximate GLR, we thus focus on decision rules based on it. The GLR statistics quantify how much the respective alternative hypotheses of a change at the different time points are plausible compared to the null hypothesis of no change. The classic non-Bayesian decision rule then amounts to thresholding the maximum of the GLR along time to detect a change.

Definition 2.5. Let $\lambda > 0$ be a threshold. The *non-Bayesian decision rule for a change* is the statistical decision rule defined as follows:

$$\max_{1 \leq i \leq n-1} \widehat{\Lambda}^i(\bar{\mathbf{x}}) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \quad \text{for all } \bar{\mathbf{x}} \in \mathcal{X}^n \ . \tag{2.11}$$

Remark 2.15. The change point is then estimated by maximum likelihood estimation, as the first time point where the maximum of the test statistics $\widehat{\Lambda}^i(\bar{\mathbf{x}})$ is reached.

Remark 2.16. As an alternative to the non-Bayesian point estimation with the maximum of the GLR, the usual Bayesian decision rule relies on integrating the MLR with a prior discrete measure on the different time points, or in simpler terms, on considering a weighted sum of the MLR. This requires some prior knowledge on the distribution of the time intervals between change points, which is not always available. Moreover, the computations may again become intractable except for certain specific priors. In particular, the literature has focused on a geometric prior over change, which is not always suited to model reliably arbitrary signals

To conclude this section, we insist again on the fact that when the parameter before change is unknown, almost all previous works employ approximations of the exact GLR in order to keep tractable updates in a sequential scheme. To this end, the parameter before change is assumed to be known, and is actually estimated at once and set equal in all hypotheses. For example, the estimation can be performed either on the whole window, or in a dead region at the beginning of the window where change detection is turned off. In the GLR, the estimators before change are then set equal to this fixed estimated value for all hypotheses. Such approximations may work when the time intervals between successive changes are important so that the approximation is valid, but their results break down because of estimation errors as soon as the changes occur more often. We argue after that we can still employ computationally efficient decision schemes based on the GLR statistics, for the large class of exponential families, and without using such approximations.

2.3. Methods for exponential families

In this section, we elaborate on the proposed methods for change detection when the parametric statistical model is an exponential family. We first develop a generic scheme for exact GLR statistics with arbitrary estimators, and interpret it in relation to maximum likelihood estimation within the dually flat geometry of exponential families. We then particularize this generic scheme to common scenarios with known and unknown parameters before and after change. We finally revisit the generic scheme through convex duality and provide a specific scheme with closed-form sequential updates for the exact generalized likelihood ratio statistics, when both the parameters before and after change are unknown, and are estimated by maximum likelihood.

2.3.1. Generic scheme

We recall that the non-Bayesian decision rule amounts to thresholding the maximum of the GLR along time. It appears that for exponential families, the GLR with arbitrary estimators is closely related to the maximum likelihood estimators. In the sequel, we restrict without loss of generality to minimal standard exponential families. For technical regularity that guarantees the existence and tractability of maximum likelihood estimates to a certain extent, we further assume that the minimal standard exponential family is also full and steep.

Theorem 2.1. *Suppose that $\mathcal{P} = \{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \mathcal{N}}$ is a full minimal steep standard exponential family, and let $\widehat{\boldsymbol{\theta}}^i_{0\,\mathrm{ml}}, \widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}$, be the maximum likelihood estimators of the parameters $\boldsymbol{\theta}^i_0, \boldsymbol{\theta}^i_1$, for any $1 \leq i \leq n-1$. The generalized likelihood ratio $\widehat{\Lambda}^i$ at time i verifies the following relation:*

$$\begin{aligned}\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,&\left\{D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})}\right) - D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}^i_0(\bar{\mathbf{x}})}\right)\right\}\\ +(n-i)\,&\left\{D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})}\right) - D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}^i_1(\bar{\mathbf{x}})}\right)\right\} \quad \text{for all } \bar{\mathbf{x}} \in \mathcal{K}^i_0 \cap \mathcal{K}^i_1 \ ,\end{aligned} \tag{2.12}$$

where the sets $\mathcal{K}^i_0, \mathcal{K}^i_1$, are defined as $\mathcal{K}^i_0 = \{\bar{\mathbf{x}} \in (\mathbb{R}^m)^n \colon \frac{1}{i}\sum_{j=1}^{i} \mathbf{x}_j \in \operatorname{int}\mathcal{K}\}$, and $\mathcal{K}^i_1 = \{\bar{\mathbf{x}} \in (\mathbb{R}^m)^n \colon \frac{1}{n-i}\sum_{j=i+1}^{n} \mathbf{x}_j \in \operatorname{int}\mathcal{K}\}$.

Proof. Let $\bar{\mathbf{x}} \in \mathcal{K}^n$ be some sample observations. Replacing the densities with their exponential representations, the GLR at time i can be developed as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = \sum_{j=1}^{i} \log \frac{\exp\left\{\widehat{\boldsymbol{\theta}}_0^i(\bar{\mathbf{x}})^\top \mathbf{x}_j - \psi(\widehat{\boldsymbol{\theta}}_0^i(\bar{\mathbf{x}}))\right\}}{\exp\left\{\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})^\top \mathbf{x}_j - \psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))\right\}} + \sum_{j=i+1}^{n} \log \frac{\exp\left\{\widehat{\boldsymbol{\theta}}_1^i(\bar{\mathbf{x}})^\top \mathbf{x}_j - \psi(\widehat{\boldsymbol{\theta}}_1^i(\bar{\mathbf{x}}))\right\}}{\exp\left\{\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})^\top \mathbf{x}_j - \psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))\right\}} \ . \tag{2.13}$$

Simplifying the logarithms and exponentials, and regrouping terms, we obtain:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = \sum_{j=1}^{i} \left\{(\widehat{\boldsymbol{\theta}}_0^i(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))^\top \mathbf{x}_j - \psi(\widehat{\boldsymbol{\theta}}_0^i(\bar{\mathbf{x}})) + \psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))\right\} + \sum_{j=i+1}^{n} \left\{(\widehat{\boldsymbol{\theta}}_1^i(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))^\top \mathbf{x}_j - \psi(\widehat{\boldsymbol{\theta}}_1^i(\bar{\mathbf{x}})) + \psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))\right\} \ . \tag{2.14}$$

Assuming now that the sample observations belong to $\mathcal{K}_0^i \cap \mathcal{K}_1^i$, the maximum likelihood estimates of the parameters $\boldsymbol{\theta}_0^i, \boldsymbol{\theta}_1^i$, exist and are unique by steepness of the family. Moreover, they belong to $\operatorname{int}\mathcal{N}$, and are given in expectation parameters by the average of the respective sufficient observations. The GLR can therefore be written as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\left\{\psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})) - \psi(\widehat{\boldsymbol{\theta}}_0^i(\bar{\mathbf{x}})) + (\widehat{\boldsymbol{\theta}}_0^i(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))^\top \nabla\psi(\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}^i(\bar{\mathbf{x}}))\right\} + (n-i)\left\{\psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})) - \psi(\widehat{\boldsymbol{\theta}}_1^i(\bar{\mathbf{x}})) + (\widehat{\boldsymbol{\theta}}_1^i(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}))^\top \nabla\psi(\widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}}))\right\} \ . \tag{2.15}$$

We can also add and subtract the maximum likelihood estimates $\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}^i(\bar{\mathbf{x}}), \widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}})$, and their log-normalizers $\psi(\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}^i(\bar{\mathbf{x}})), \psi(\widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}}))$, to make Bregman divergences B_ψ appear as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,\left\{B_\psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}^i(\bar{\mathbf{x}})) - B_\psi(\widehat{\boldsymbol{\theta}}_0^i(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}^i(\bar{\mathbf{x}}))\right\} + (n-i)\,\left\{B_\psi(\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}})) - B_\psi(\widehat{\boldsymbol{\theta}}_1^i(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}}))\right\} \ . \tag{2.16}$$

This proves the theorem by rewriting the Bregman divergences on the natural parameters as Kullback-Leibler divergences on the swapped corresponding distributions. □

Remark 2.17. Actually, we can restrict to minimal standard families without loss of generality because the GLR statistics are invariant, up to set of null measure, under the transformations involved in the reduction of an arbitrary exponential family to a minimal standard one, that is, sufficiency, reparametrization, and choice of a dominating measure. This makes the related statistical decision intrinsic to the underlying statistical manifold, which is confirmed by the expression of the

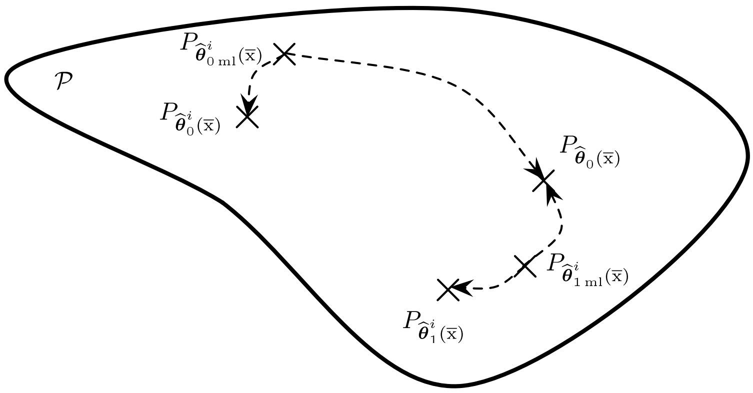

Figure 2.3.: Geometrical interpretation of change detection. The problem of change detection with exponential families and exact generalized likelihood ratio test statistics based on arbitrary estimators, can be seen as computing certain information divergences between the estimated distributions and the maximum likelihood distributions in the different hypotheses.

test statistics in terms of Kullback-Leibler divergences on the probability densities, since the Kullback-Leibler divergence is invariant under the same transformations. Moreover, the equivalence between different minimal standard reductions up to an affine transformation of the natural parameters, can be seen in the expression of the test statistics in terms of Bregman divergences, since Bregman divergences are invariant under affine transformations of their generating function.

Remark 2.18. As illustrated in Figure 2.3, the generic GLR scheme can be interpreted as (i) computing the divergences between the maximum likelihood estimates before change, respectively after change, and the chosen estimate with no change, (ii) correcting the chosen estimates before change, respectively after change, compared to the maximum likelihood estimates before change, respectively after change, (iii) weighting by the number of sample observations before change, respectively after change.

Using different combinations of estimators for the respective parameters in the hypotheses, we can derive specific forms of the GLR in many scenarios with known and unknown parameters, with arbitrary estimates or maximum likelihood estimates, and even with approximate statistics. Before discussing these different scenarios, we state a direct corollary which encompasses most of them, except from the exact GLR when both parameters are unknown, in the case where all estimators before change are set equal.

Corollary 2.2. *Suppose that the estimators $\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}^i_0$ are equal, for any $1 \leq i \leq n-1$. The generalized likelihood ratio $\widehat{\Lambda}^i$ at time i verifies the following relation:*

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = (n-i)\,\left\{D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\| P_{\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})}\right) - D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\| P_{\widehat{\boldsymbol{\theta}}^i_1(\bar{\mathbf{x}})}\right)\right\} \quad \textit{for all } \bar{\mathbf{x}} \in \mathcal{K}^i_1\,. \tag{2.17}$$

Proof. This follows from the theorem in the case where $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}^i_0$. More precisely, it is not required in this case that $\bar{\mathbf{x}} \in \mathcal{K}^i_0$ in the proof of the theorem. Indeed, the terms with the sufficient observations before change vanish so that we need not introduce the maximum likelihood estimate before change. □

Remark 2.19. This specific GLR scheme can therefore be interpreted as (i) computing the divergence between the maximum likelihood estimate after change, and the chosen estimator with no change, or equivalently before change since they are equal, (ii) correcting the chosen estimator after change compared to the maximum likelihood estimate after change, (iii) weighting by the number of sample observations after change.

Remark 2.20. We have supposed implicitly that the chosen estimator $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}_0^i$ is well-defined everywhere on $(\mathbb{R}^m)^n$, or at least on $\mathcal{K}_1^i$. If it is not the case, then its exact domain of definition needs to be considered as an intersection with $\mathcal{K}_1^i$. For example, if we use the maximum likelihood estimator in a dead region of $n_0 < n$ samples at the beginning of the window, then the correct domain is $\mathcal{K}_0^{n_0} \cap \mathcal{K}_1^i$, for any $n_0 \leq i \leq n-1$. The reasoning is similar for the maximum likelihood estimator on the whole window.

Remark 2.21. A similar relation can be derived when the estimators after change and with no change are equal. Nevertheless, this case is in general not relevant in practical situations so we do not discuss it further.

2.3.2. Case of a known parameter before change

We consider here the common scenario where the parameter before change is assumed to be known. We begin with the simpler case where the parameter after change is also known. This actually corresponds to the simple LR statistics, expressed here for exponential families in terms of information divergences.

Example 2.1. For the problem of change detection with exponential families where both the parameters $\boldsymbol{\theta}_{\mathrm{bef}}, \boldsymbol{\theta}_{\mathrm{aft}}$ before and after change are known, the respective estimators in the hypotheses are taken constant as $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}_0^i = \boldsymbol{\theta}_{\mathrm{bef}}$, and $\widehat{\boldsymbol{\theta}}_1^i = \boldsymbol{\theta}_{\mathrm{aft}}$, for all $1 \leq i \leq n-1$. The test statistics can therefore be expressed as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = (n-i)\,\left\{D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\boldsymbol{\theta}_{\mathrm{bef}}}\right) - D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\boldsymbol{\theta}_{\mathrm{aft}}}\right)\right\} \ . \tag{2.18}$$

Now considering that the parameter after change is unknown, we obtain a similar expression as for the LR statistics. If we also assume that the unknown parameter is estimated by maximum likelihood in the respective hypotheses, the expression can be simplified further. This actually corresponds to the standard GLR statistics for a known parameter before change, expressed here for exponential families in terms of a simple information divergence.

Example 2.2. For the problem of change detection with exponential families where the parameter $\boldsymbol{\theta}_{\mathrm{bef}}$ before change is known, and the parameter after change is unknown, the respective estimators before change in the hypotheses are taken constant as $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}_0^i = \boldsymbol{\theta}_{\mathrm{bef}}$, for all $1 \leq i \leq n-1$. The test statistics can therefore be expressed as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = (n-i)\,\left\{D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\boldsymbol{\theta}_{\mathrm{bef}}}\right) - D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}^i_1(\bar{\mathbf{x}})}\right)\right\} \ . \tag{2.19}$$

Supposing further that the maximum likelihood estimators are chosen for the estimators after change as $\widehat{\boldsymbol{\theta}}_1^i = \widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i$, for all $1 \leq i \leq n-1$, the test statistics can be simplified as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = (n-i)\,D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}})}\middle\|P_{\boldsymbol{\theta}_{\mathrm{bef}}}\right) \quad . \tag{2.20}$$

2.3.3. Case of unknown parameters before and after change

We now turn to the common scenario where both the parameters before and after change are unknown. In the literature, this situation is most of the time addressed by setting the estimators before change equal in the respective hypotheses, as discussed previously. It corresponds to the general case of the above corollary. Further considering the maximum likelihood estimators, the statistics actually specialize to the approximate GLR statistics commonly employed.

Example 2.3. For the problem of change detection with exponential families where both the parameters before and after change are unknown, the respective estimators before change in the hypotheses can be set equal as $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}_0^i$, for all $1 \leq i \leq n-1$. Supposing further that the maximum likelihood estimators are chosen for the estimators after change as $\widehat{\boldsymbol{\theta}}_1^i = \widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i$, for all $1 \leq i \leq n-1$, the test statistics can be simplified as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = (n-i)\,D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}_0(\bar{\mathbf{x}})}\right) \quad . \tag{2.21}$$

The single estimator before or with no change $\widehat{\boldsymbol{\theta}}_0$ can then be chosen as the maximum likelihood estimator over the whole window or over a dead region at the beginning of the window, as discussed previously.

We propose here an alternative to approximate GLR statistics, by considering exact GLR statistics where the estimators before change are estimated separately in the respective hypotheses. In the general case, it corresponds to the generic updates of the above theorem. Further considering the maximum likelihood estimators, the statistics specialize to the exact GLR statistics defined in the literature, but replaced in practice with the approximate GLR statistics for computational reasons.

Example 2.4. For the problem of change detection with exponential families where both the parameters before and after change are unknown, the respective estimators in the hypotheses can be chosen to be the maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}$, $\widehat{\boldsymbol{\theta}}_0^i = \widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}^i$, $\widehat{\boldsymbol{\theta}}_1^i = \widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i$, for all $1 \leq i \leq n-1$. The test statistics can then be expressed as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}^i(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}\right) + (n-i)\,D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}_{1\,\mathrm{ml}}^i(\bar{\mathbf{x}})}\middle\|P_{\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}\right) \quad . \tag{2.22}$$

This directly follows from the theorem, after remarking that the maximum likelihood estimate with no change is well-defined. Indeed, if the observations belong to $\mathcal{K}_0^i \cap \mathcal{K}_1^i$, then they also belong to $\mathcal{K}_0 = \{\bar{\mathbf{x}} \in (\mathbb{R}^m)^n : \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j \in \operatorname{int}\mathcal{K}\}$ by convexity of $\operatorname{int}\mathcal{K}$.

Remark 2.22. The exact GLR scheme with maximum likelihood estimators can therefore be interpreted as (i) computing the divergence between the maximum likelihood estimate before change, respectively after change, and the maximum likelihood estimate with no change, (ii) weighting by the number of sample observations before change, respectively after change.

2.3.4. Generic scheme revisited through convex duality

The above expressions of the GLR variations in terms of information divergences give both statistical and geometrical intuitions to change detection in exponential families. Moreover, because of the correspondence between exponential families and their associated Bregman divergences, this shows tight links between statistical and distance-based approaches to change detection. Further taking advantage of the dually flat information geometry of exponential families, we now rely on convex duality to reparametrize the problem from the natural to the expectation parameter space. This provides additional evidence for the corrective role of the chosen estimators compared to maximum likelihood in the generic GLR scheme.

Proposition 2.3. *Suppose that $\mathcal{P} = \{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \mathcal{N}}$ is a full minimal steep standard exponential family, and let $\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}, \widehat{\boldsymbol{\theta}}^i_{0\,\mathrm{ml}}, \widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}$, be the maximum likelihood estimators of the parameters $\boldsymbol{\theta}_0, \boldsymbol{\theta}^i_0, \boldsymbol{\theta}^i_1$, for any $1 \leq i \leq n-1$. The generalized likelihood ratio $\widehat{\Lambda}^i$ at time i verifies the following relation:*

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,\phi(\widehat{\boldsymbol{\eta}}^i_0(\bar{\mathbf{x}})) + (n-i)\,\phi(\widehat{\boldsymbol{\eta}}^i_1(\bar{\mathbf{x}})) - n\,\phi(\widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})) + \Delta^i_{\mathrm{ml}}(\bar{\mathbf{x}}) \quad \textit{for all } \bar{\mathbf{x}} \in \mathcal{K}^i_0 \cap \mathcal{K}^i_1 \;, \tag{2.23}$$

where the corrective term Δ^i_{ml} at time i compared to maximum likelihood estimation is expressed as follows:

$$\begin{aligned}\Delta^i_{\mathrm{ml}}(\bar{\mathbf{x}}) = i\,\big(\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\eta}}^i_0(\bar{\mathbf{x}})\big)^{\top} \nabla\phi(\widehat{\boldsymbol{\eta}}^i_0(\bar{\mathbf{x}})) + (n-i)\,\big(\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\eta}}^i_1(\bar{\mathbf{x}})\big)^{\top} \nabla\phi(\widehat{\boldsymbol{\eta}}^i_1(\bar{\mathbf{x}}))\\ - n\,\big(\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})\big)^{\top} \nabla\phi(\widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})) \quad \textit{for all } \bar{\mathbf{x}} \in \mathcal{K}^i_0 \cap \mathcal{K}^i_1 \;.\end{aligned} \tag{2.24}$$

Proof. Rewriting the generic GLR at time i with Bregman divergences B_ϕ on the expectation parameters, we obtain the following expression:

$$\begin{aligned}\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,\big\{B_\phi(\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})) - B_\psi(\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\eta}}^i_0(\bar{\mathbf{x}}))\big\}\\ + (n-i)\,\big\{B_\phi(\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})) - B_\psi(\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) \| \widehat{\boldsymbol{\eta}}^i_1(\bar{\mathbf{x}}))\big\} \;.\end{aligned} \tag{2.25}$$

Developing the Bregman divergences with their standard expressions and regrouping terms, the GLR can then be expressed as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,\phi(\widehat{\boldsymbol{\eta}}^i_0(\bar{\mathbf{x}})) + (n-i)\,\phi(\widehat{\boldsymbol{\eta}}^i_1(\bar{\mathbf{x}})) - n\,\phi(\widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})) + \Delta^i_{\mathrm{ml}}(\bar{\mathbf{x}}) \;, \tag{2.26}$$

where $\Delta^i_{\mathrm{ml}}(\bar{\mathbf{x}})$ writes as follows:

$$\begin{aligned}\Delta^i_{\mathrm{ml}}(\bar{\mathbf{x}}) = i\,\big(\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\eta}}^i_0(\bar{\mathbf{x}})\big)^{\top} \nabla\phi(\widehat{\boldsymbol{\eta}}^i_0(\bar{\mathbf{x}})) + (n-i)\,\big(\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) - \widehat{\boldsymbol{\eta}}^i_1(\bar{\mathbf{x}})\big)^{\top} \nabla\phi(\widehat{\boldsymbol{\eta}}^i_1(\bar{\mathbf{x}}))\\ - \big(i\,\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) + (n-i)\,\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) - n\,\widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})\big)^{\top} \nabla\phi(\widehat{\boldsymbol{\eta}}_0(\bar{\mathbf{x}})) \;.\end{aligned} \tag{2.27}$$

The last term can be re-expressed to provide the final expression of the corrective term $\Delta^i_{\mathrm{ml}}(\bar{\mathbf{x}})$. Indeed, it suffices to observe that the maximum likelihood estimate with no change is the barycenter of the maximum likelihood estimates before and after a change at time i, that is:

$$n\,\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) = i\,\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) + (n-i)\,\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) \enspace . \tag{2.28}$$

This is easily seen from the equation of the maximum likelihood estimates as averages of the corresponding sufficient observations. Here we have used implicitly that since the sample observations belong to $\mathcal{K}^i_0 \cap \mathcal{K}^i_1$, they also belong to $\mathcal{K}_0$ by convexity of $\operatorname{int}\mathcal{K}$, so that the maximum likelihood estimate with no change exist and is unique by steepness of the family. This proves the proposition. $\square$

2.3.5. Case of unknown parameters and maximum likelihood

We finally use the above results to revisit the common scenario where both the parameters before and after change are unknown, in particular for exact GLR statistics based on maximum likelihood estimators. In this case, the corrective term vanishes and simplifies the expression of the statistics.

Example 2.5. For the problem of change detection with exponential families where both the parameters before and after change are unknown, and where the estimators in the respective hypotheses are chosen to be the maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}$, $\widehat{\boldsymbol{\theta}}^i_0 = \widehat{\boldsymbol{\theta}}^i_{0\,\mathrm{ml}}$, $\widehat{\boldsymbol{\theta}}^i_1 = \widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}$, for all $1 \leq i \leq n-1$, the test statistics can be expressed as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,\phi(\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}})) + (n-i)\,\phi(\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})) - n\,\phi(\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})) \enspace . \tag{2.29}$$

This alternative expression for the exact GLR statistics greatly simplifies its computation. It is obtained in closed form in terms of the conjugate ϕ of the log-normalizer ψ for the exponential family. Moreover, because maximum likelihood estimates between successive windows are related by simple time shifts or barycentric updates in expectation parameters, this provides a computationally efficient scheme for calculating the statistics in a sequential fashion. For example, if no change has been detected in the previous window, the statistics can then be simply updated as $\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow \widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}})$, $\widehat{\boldsymbol{\eta}}^{n-1}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow \widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})$, $\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow ((n-i-1)\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})+\mathbf{x}_n)/(n-i)$, $\widehat{\boldsymbol{\eta}}^{n-1}_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow \mathbf{x}_n$, $\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow ((n-1)\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) + \mathbf{x}_n)/n$, for all $1 \leq i < n-1$. Similar updates can be obtained when a change point has been detected. Moreover, certain values at which the conjugate ϕ is evaluated actually reappear because of time shifts, and can therefore be stored to facilitate tractability. Nonetheless, the computational complexity in a given window is still linear in the number of observations since the GLR statistics need to be computed for every potential change point.[3]

[3] In a sequential change detection scheme where the window is growing until a change is detected, it may therefore be necessary to employ heuristics to bound the number of tested change points, for example, by using a maximum window size or by pruning the lowest statistics.

2.4. Discussion

In this chapter, we proposed generic methods for sequential change detection with exponential families. The developed framework therefore unifies change detection for many common statistical models. We followed a standard non-Bayesian approach where change detection is formulated as a statistical decision problem with multiple hypotheses and solved by employing generalized likelihood ratio test statistics. We also introduced arbitrary estimators in the definition of the generalized likelihood ratios. Applying this to exponential families, we developed a generic scheme for change detection under common scenarios with known or unknown parameters and arbitrary estimators, in close relation to maximum likelihood estimation. We also interpreted this scheme within the dually flat geometry of exponential families, hence providing both statistical and geometrical intuitions, and bridging the gap between statistical and distance-based approaches to change detection. We finally revisited this scheme through convex duality, and derived an attractive scheme with closed-form sequential updates for the exact generalized likelihood ratio statistics, when both the parameters before and after change are unknown and are estimated by maximum likelihood. Several directions of improvement were however left out for future work.

To begin with, some direct extensions of the framework are possible. An example is the generalization of the obtained results to non-full families. We can actually show that the results derived here still hold for certain curved exponential families, with slight modifications when revisiting the schemes through convex duality, so as to account for the maximum likelihoods before and after change not being simply linked anymore with the maximum likelihood for no change through a barycentric relation. Interestingly, this extension relies on the generalized Pythagorean theorem, and on projections according to information divergences onto autoparallel submanifolds in the dually flat geometry of exponential families. We yet did not develop this extension here for the sake of conciseness.

Another example is to consider non-steep family as well. This requires however a few additional assumptions. For these families, the maximum likelihood estimates exist and are unique under the same conditions as for steep families, that is, if the average of the sufficient observations lies in the interior of the convex support of the dominating measure. Nevertheless, the maximum likelihood estimate in expectation parameters does not necessarily equals the average of the sufficient observations. This is because the open expectation parameter space is a proper subset of the convex support of the dominating measure. Therefore, the presented results still hold under the condition that this average is indeed in the expectation parameter space. On the contrary to steep families, it does not unfortunately occur with probability one as the sample size tends to infinity, so that the schemes may actually never apply in practical situations.

A third example is the analysis of specific schemes when using other estimators than the maximum likelihood. For instance, we can derive a similar sequential update scheme as for maximum likelihood estimates, when using convex duality for maximum a posteriori estimates based on standard exponential family conjugate priors. The scheme is slightly more demanding since the corrective term does not simplify anymore. The maximum a posteriori estimates are however also related by simple time shifts and barycentric updates which facilitate tractability. Other estimators could also be investigated such as using quasi likelihoods to account for potential model misspecification. This idea is worth exploring.

A last example is the consideration of aggregates, or closures, of exponential families [Barndorff-Nielsen 1978, Brown 1986, Csiszár and Matúš 2005]. This would permit to include the domain boundaries as well as the limit distributions in the schemes, while guaranteeing the existence of maximum likelihood estimates and of their simple expression in all situations. Further considerations are however needed on this line to rigorously extend the obtained results.

In complement to the statistical assumptions of mutual independence considered in the proposed framework, we also want to address the statistical dependence between the random variables. As discussed in the text, the statistical framework exposed based on multiple hypothesis testing can be extended to arbitrary conditional models. Nevertheless, the issue in this context rather becomes the tractability of the test statistics as soon as the parameters are unknown. Specific change detection schemes have yet been proposed for particular models, notably for autoregressive models as in [André-Obrecht 1988]. More generally, we would like to address change detection in linear or even non-linear systems. Online schemes based on particle filtering have been proposed for instance in [Fearnhead and Liu 2007] to detect changes in non-linear systems, but such schemes suffer from computational loads when properly considering unknown parameters and exact inference. An alternative based on CUSUM-like test statistics has recently been proposed in [Vaswani 2007]. A different approach would be to model the observations through marginals and copulas, and detect changes either linked with the marginal distribution or with the dependency structure of the data.

On another perspective, we could also formulate sequential change detection in a Bayesian framework to complement the non-Bayesian framework developed here. This implies proposing relevant distributions to model both the run length between change points and the unknown parameters, seen as random variables. In this context, several frameworks have already been proposed, for example in [Adams and MacKay 2007, Turner et al. 2009, Lai et al. 2009, Lai and Xing 2010, Fearnhead and Liu 2011], certain dealing notably with exponential families. The inference schemes, however, are in general more demanding than for non-Bayesian approaches, even when using convenient conjugate priors on parameters. Moreover, conjugate priors do not necessarily model adequately the true distributions so that alternatives may be required. Using mixtures of conjugate priors, or equivalently hyperpriors in a hierarchical model, may provide interesting solutions to address this.

Finally, we would like to address the use of alternative statistics than likelihoods. This could be achieved by reversing the problem and starting from divergences. Here we considered test statistics and derived expressions in terms of information divergences within the dually flat geometry of exponential families. Another approach is to directly design statistics through divergences in order to obtain more robust estimators and tests while maintaining sufficient efficiency [Eguchi 1983, Basu et al. 1998, Eguchi and Kano 2001, Mihoko and Eguchi 2002, Pardo 2006, Broniatowski and Keziou 2009, Eguchi 2009, Basu et al. 2011]. This was left out for future work.

3. Non-Negative Matrix Factorization with Convex-Concave Divergences

In this chapter, we elaborate methods for non-negative matrix factorization within the computational framework of information geometry. We notably formulate a generic and unifying framework for non-negative matrix factorization with convex-concave divergences. This framework encompasses many common information divergences presented in Chapter 1, such as Csiszár divergences, certain Bregman divergences, and in particular all α-divergences and β-divergences. A general optimization scheme is developed based on variational bounding with surrogate auxiliary functions for almost arbitrary convex-concave divergences. Monotonically decreasing updates are then obtained by minimizing the auxiliary function. The proposed framework also permits to consider symmetrized and skew divergences for the cost function. In particular, the generic updates are specialized to provide updates for Csiszár divergences, certain skew Jeffreys-Bregman divergences, skew Jensen-Bregman divergences. It leads to several known multiplicative updates, as well as novel multiplicative updates, for α-divergences, β-divergences, and their symmetrized or skew versions. These results are also generalized by considering the family of skew (α, β, λ)-divergences. This is applied in Chapter 5 to design a real-time system for polyphonic music transcription.

3.1. Context

In this section, we first provide some background information on the problem of non-negative matrix factorization. We then discuss the motivations of our approach to this problem. We finally sum up our main contributions in this context.

3.1.1. Background

Let us consider a dataset $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ of size n consisting of non-negative multivariate observations of dimension m, and stack these observations into an $m \times n$ non-negative matrix $\mathbf{Y}$ whose rows and columns represent respectively the different variables and observations. As sketched in Figure 3.1, the problem of *non-negative matrix factorization* (NMF) is to find an approximate factorization of $\mathbf{Y}$ into an $m \times r$ non-negative matrix $\mathbf{A}$ and an $r \times n$ non-negative matrix $\mathbf{X}$, such that $\mathbf{Y} \approx \mathbf{AX}$, where the integer $r < \min(m, n)$ is the *rank of factorization*. In this linear model, each observation $\mathbf{y}_j$ can then be expressed as $\mathbf{y}_j \approx \mathbf{A}\mathbf{x}_j$. The matrix $\mathbf{A}$ thus forms

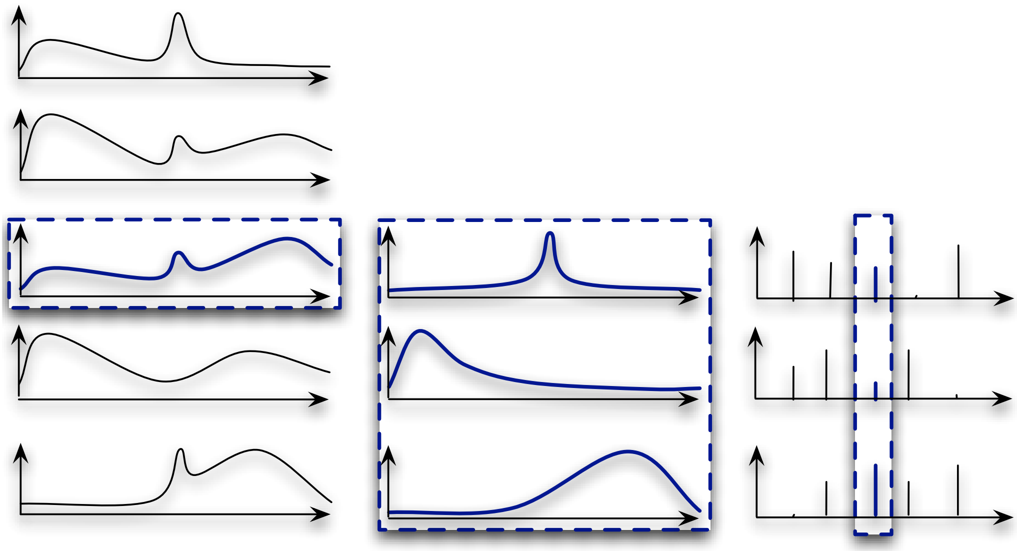

Figure 3.1.: Schematic view of non-negative matrix factorization. The problem of non-negative matrix factorization consists in reconstructing non-negative observations, as the addition of a small number of non-negative atoms with non-negative weights.

a *basis* or *dictionary*, and the columns of $\mathbf{X}$ are a *decomposition* or *encoding* of the respective columns of $\mathbf{Y}$ into this basis. Moreover, the rank of factorization is generally chosen such that $mr + rn \ll mn$, so that the factorization can be thought of as a compression or reduction of the observed data.

The NMF problem is therefore an unsupervised technique for multivariate data analysis and dimensionality reduction, such as principal component analysis (PCA) and independent component analysis (ICA). The distinction with the two latter techniques, however, is that no other constraints than non-negativity of the observations and factors are required in NMF. These constraints are dropped in PCA and ICA in favor of constraints for basis vectors that are respectively statistically uncorrelated or independent. As a result, cancellation is not allowed in the decomposition of NMF, whereas it is allowed in the decomposition of PCA and ICA through the subtraction of basis vectors. The main philosophy of NMF is thus to explain the observations in a constructive manner by addition of a small number of non-negative parts shared by several observations. Such assumptions are particularly interesting when negative values cannot be interpreted, for example, the pixel intensity for images, the word occurrence for texts, or the frequency power spectrum for sounds.

Because of the low-rank factorization, the NMF model is most of the time only approximate. As a result, the NMF problem is often formulated as an optimization problem, where the aim is to find a factorization which optimizes a given functional, called *cost function* or *objective function*, that quantifies the quality of the factorization. In general, the rank of factorization is kept fixed in the optimization, either chosen by the user or sometimes estimated directly from the data. The optimization is then performed over all possible pairs of non-negative factors $\mathbf{A}$ and $\mathbf{X}$. Moreover, most if not all works on NMF focus on *separable* cost functions, that is, that are the sum of a given scalar cost function considered element-wise.

Historically, the formulation of NMF is attributed to Paatero and Tapper [1994], Paatero [1997], who solved the problem for a weighted Euclidean cost function by using an alternating non-negative least squares algorithm, and applied it to the analysis of environmental data in chemometrics. It became however more popular after the work of Lee and Seung [1999, 2001], who investigated some simple and useful algorithms based on *multiplicative updates* for the Euclidean and Kullback-Leibler cost functions, with applications to the analysis of facial images and of text documents. Since then, there has been a growing interest for NMF in the communities of machine learning and signal processing, and a flourishing literature has developed about other algorithms and extensions to the standard NMF problem. We refer the reader to the survey article of Berry et al. [2007] for a general introduction, and to the book of Cichocki et al. [2009] for a comprehensive treatment as well as a discussion of the main applicative domains of NMF algorithms, including bioinformatics, spectroscopy, email surveillance, and analysis of text, image or audio data.

The differences between the variations of NMF can be summarized in three principal directions. First, the standard model can be modified, for example, by using non-negative tensors instead of non-negative matrices. Second, the standard constraints can be changed, for example, by imposing the sparsity of the factors in addition to their non-negativity. Third, the standard cost functions can be enhanced, for example, by using more general divergences, or by adding penalization terms to regularize the solutions. Several other algorithms than alternating least squares and multiplicative updates have also been developed to solve NMF and its extensions, notably based on gradient descent methods, adapted to the non-negativity constraints by using exponentiation, line search, backtracking, or projections.

More recently, NMF has also been considered from statistical perspectives. In this context, several authors have studied the links between optimization problems for NMF and statistical inference under generative distributional assumptions of the dataset [Schmidt and Laurberg 2008, Schmidt et al. 2009, Virtanen et al. 2008, Virtanen and Cemgil 2009, Févotte et al. 2009, Févotte and Cemgil 2009, Févotte 2011, Cemgil 2009, Zhong and Girolami 2009, Bertin et al. 2010, Hoffman et al. 2010, Dikmen and Févotte 2011, Lefèvre et al. 2011a]. In summary, equivalence is known between the NMF problem with either the Euclidean, Kullback-Leibler, or Itakura-Saito cost function, and the maximum likelihood estimation under either a normal, a Poisson, or a gamma observation model, respectively. There also exist equivalent composite models with superimposed components, which exploit either the closure under summation of the normal and Poisson observation models for the Euclidean and Kullback-Leibler cost functions, or a specific property of the circular complex normal distribution for the Itakura-Saito cost function.

Based on such statistical formulations, it is then possible to use many tools from statistical inference to solve the related NMF problems, such as the expectation-maximization algorithm and its generalizations. Moreover, it is possible to add statistical prior information to the problem, which can be seen as adding penalty terms to the cost function for regularization. For example, a prior exponential distribution on the activations is known to favor a sparse solution through an ℓ_1-norm penalty. A full Bayesian framework can be developed in this manner. One then rather resorts to posterior estimation methods for solving the NMF problem, by using Monte-Carlo methods such as Markov chain Monte Carlo schemes and in particular Gibbs sampling, or by using variational Bayes methods.

In parallel, more general matrix factorization models have also been considered by using the correspondence between exponential families and Bregman divergences [Collins et al. 2002, Singh and Gordon 2008, Mohamed et al. 2009, 2012]. The general setting is that of latent variable models, where the sample observations $\mathbf{y}_1, \dots, \mathbf{y}_n$ are supposed to be the sufficient observations of an exponential family with respective expectation parameters $\mathbf{A}\mathbf{x}_1, \dots, \mathbf{A}\mathbf{x}_n$.[1] Under mild assumptions, the negative log-likelihood $-\log p(\mathbf{Y}|\mathbf{A}, \mathbf{X})$ actually equals the sum of Bregman divergences $\sum_{j=1}^{n} B_\phi(\mathbf{y}_j \| \mathbf{A}\mathbf{x}_j)$ up to constant terms, where ϕ is the conjugate of the log-normalizer ψ for the exponential family. As a result, maximum likelihood estimation in this model amounts to a right-sided approximate matrix factorization problem with a Bregman divergence. It therefore elucidates the relations between the cost function and the distributional assumptions, and unifies many matrix factorizations models such as PCA, ICA, as well as NMF and the related techniques of *probabilistic latent semantic analysis* (PLSA) [Hofmann 1999, Gaussier and Goutte 2005], and *probabilistic latent component analysis* (PLCA) [Smaragdis and Raj 2007, Shashanka et al. 2008]. Again, a full Bayesian framework with adapted methods for statistical inference is also possible in this setting.

3.1.2. Motivations

Because of the wide applicative range of NMF techniques, there is a strong motivation in providing generic methods that handle data of heterogeneous types under different distributional assumptions. In many NMF algorithms, however, a particular cost function or statistical model is assumed. Therefore, a generic statistical framework based on the correspondence between exponential families and Bregman divergences, as discussed above, seems an interesting approach to the NMF problem. Indeed it provides tight links between statistical inference in the models involved and optimization of the related cost functions for a great variety of problems.

Nevertheless, because of the generality of this framework, the inference schemes employed may undergo theoretical and computational issues, related to convergence, efficiency and tractability. Moreover, it is not always evident to deal soundly with the non-negative constraints in generic inference schemes when the priors or model do not guarantee inherently this constraint. As a result, specific optimization schemes are still often derived to address particular NMF problems [Tan and Févotte 2009, Psorakis et al. 2011].

Other issues are that we do not always know the exact distribution of the data to analyze, and that the NMF problem may suffer from model misspecification, hence undermining the robustness of inference. A similar problem arises from the potential presence of outliers, which requires robust inference methods to be dealt with. In statistical inference, an alternative to tackle robustness issues is to minimize other divergences than that for the maximum likelihood estimator, for example, the left-sided related divergence corresponding to the maximum likelihood predictor [Barndorff-Nielsen 1978, Brown 1986], or by employing more general families of

[1] It can also be generalized to other parametrizations through the use of link functions.

divergences designed to improve robustness while maintaining sufficient efficiency [Eguchi 1983, Basu et al. 1998, Eguchi and Kano 2001, Mihoko and Eguchi 2002, Pardo 2006, Broniatowski and Keziou 2009, Eguchi 2009, Basu et al. 2011].

In the context of NMF, such alternatives have been considered by several authors. Cichocki et al. [2006] studied the right-sided NMF problem with certain Csiszár divergences and proposed heuristic multiplicative updates. Cichocki et al. [2008] focused on the right-sided NMF problem with α-divergences, and derived multiplicative updates that decrease monotonically the cost function, by using an auxiliary function based on the convexity of the cost function and Jensen's inequality. In the meantime, Dhillon and Sra [2006] studied the right- and left-sided NMF problems with Bregman divergences, and provided heuristic updates for the right-sided problem, as well as monotonically decreasing updates for the left-sided problem based again on an auxiliary function using convexity of the cost function and Jensen's inequality. Kompass [2007] considered the right-sided NMF problem with a subfamily of β-divergences for which the cost function is convex, and proposed multiplicative updates that decrease monotonically the cost function with the very same approach.

More recently, Nakano et al. [2010a] extended this to the right-sided NMF problem with any β-divergence, still using an auxiliary function relying on Jensen's inequality for the convex part of the cost function, and on the tangent inequality for the concave part. Févotte and Idier [2011] independently obtained the same results, and introduced other monotonically decreasing updates based on the equalization of the auxiliary function, rather than on its minimization as in the previous approaches. Cichocki et al. [2011] generalized some results to the right-sided NMF problem with (α, β)-divergences and provided monotonically decreasing multiplicative updates, again by minimizing an auxiliary function built with the same approach.

We position in the continuity of these works and seek to formulate a unifying framework for NMF optimization based on general families of divergences, with guaranteed monotonic decrease and thus convergence of the cost function.[2] Moreover, we aim at studying indifferently left- and right-sided NMF problems, as well as considering ways of symmetrization and skewing of the cost functions, which has been left aside from the literature up to now.

In the sequel, we consider algorithms based on alternating updates. Because the NMF problem is in general not jointly convex in both factors and uneasy to solve, most algorithms iteratively update the factors $\mathbf{A}$ and $\mathbf{X}$ separately. After initializing $\mathbf{A}$ and $\mathbf{X}$, a generic NMF scheme can be stated as follows. We first fix $\mathbf{X}$ and update $\mathbf{A}$ so as to improve the quality of the factorization, then fix $\mathbf{A}$ and update $\mathbf{X}$ so as to improve the quality of the factorization. These two steps are repeated in turn until a termination criterion is met. We can of course consider the updates in the reverse order, and the two initializations are not always required depending on the algorithm and the order of the updates. In this context, it thus seems crucial from theoretical and practical viewpoints to us, even if heuristic updates are still often used, that the respective updates guarantee at least the monotonic decrease of the cost function so that the factorization is improved at each iteration.

[2]The convergence of the cost function does not prove however its convergence to a global or local minimum, nor the convergence of the sequence of updates itself, which are theoretically more difficult to obtain and are not discussed here.

We also consider NMF with separable divergences. This basic assumption allows viewing the factors $\mathbf{A}$ and $\mathbf{X}$ similarly. It seems intuitive at first sight, since if $\mathbf{Y} \approx \mathbf{AX}$, then $\mathbf{Y}^\top \approx \mathbf{X}^\top \mathbf{A}^\top$, so that the roles of the factors are interchangeable. In this context, the NMF problem can be conveniently reduced to finding an update of the respective columns $\mathbf{x}_j$ of $\mathbf{X}$, where $\mathbf{A}$ is kept fixed. We thus concentrate without loss of generality on solving the *supervised non-negative matrix factorization* problem, also called *non-negative decomposition* problem, $\mathbf{y} \approx \mathbf{Ax}$, where the non-negative vector $\mathbf{y}$ and matrix $\mathbf{A}$ are known, and the non-negative vector $\mathbf{x}$ is chosen so as to optimize the quality of factorization with respect to a separable cost function. Nevertheless, we notice that the separability assumption, even if intuitive and used in almost all works, is disputable. Indeed, in the statistical framework exposed above, it is closely related to assumptions on the independence or exchangeability of the observations which are not always met in practice. Yet we leave these considerations apart from the present work, and focus on separable divergences.

3.1.3. Contributions

We follow a common optimization framework for non-negative matrix factorization with separable divergences where the factors are updated in turn iteratively, so that we can reduce the original problem to a non-negative decomposition problem. As mentioned above, the main issue of most approaches in this context is to rely on specific cost functions, some of which may lead to convergence issues and in particular to a non-monotonic decrease of the cost function. Nonetheless, several independent approaches have tackled this issue by employing families of cost functions, notably certain Csiszár and Bregman divergences, as well as the parametric α-divergences, β-divergences and (α, β)-divergences. Our main contributions in this continuity can be summarized as follows.

We formulate a generic and unifying framework for non-negative matrix factorization with convex-concave divergences. These general divergences encompass many common information divergences, and therefore permits to handle the majority of cost functions proposed in the literature so far, as well as novel ones, with a single generic methodology. It notably includes all Csiszár divergences in both the left and right arguments, all Bregman divergences in the left argument, and certain Bregman divergences in the right argument. In particular, all α-divergences and β-divergences are actually convex-concave in both arguments. We propose a general optimization scheme for non-negative decomposition with convex-concave divergences under mild regularity assumptions. It relies on the optimization technique of variational bounding, or majorization, where majorizing auxiliary functions are constructed for convex-concave cost functions by using Jensen's inequality for the convex part and on the tangent inequality for the concave part, hence extending several approaches based on auxiliary functions discussed above. Generic updates are obtained by minimizing the auxiliary functions around the previous solution at each iteration, which guarantees to make the cost function decrease monotonically until its convergence. The generic updates are discussed in contrast to the well-known concave-convex procedure, where we show how the use of Jensen's inequality encodes the non-negativity, and how its coupling with separability permits to reduce the multidimensional NMF optimization problem into a simpler problem with a system of one-dimensional independent equations that can be solved more efficiently. We also provide insights into the simplification of this system into closed-form updates, and notably clarify a reasonable assumption in relation to Pexider's functional equations to obtain attractive multiplicative updates. This assumption shows that information divergences based on power functions and their limit cases are the reasonable candidates for obtaining multiplicative updates. This includes the parametric families of α-divergences and β-divergences, as well as the (α, β)-divergences. The proposed framework also permits to consider non-negative matrix factorization with symmetrized and skew divergences. To the best of our knowledge, it is the first time that this is considered in the context of non-negative matrix factorization. The remainder of this chapter is organized as follows.

In Section 3.2, we formalize the optimization framework for non-negative matrix factorization with separable divergences. In Subsection 3.2.1, we formulate the non-negative decomposition problem as a cost function minimization problem. In Subsection 3.2.2, we introduce the framework of variational bounding, where majorizing auxiliary functions act as surrogates for the iterative minimization of the cost function. In Section 3.3, we apply the formulated framework to derive methods for non-negative matrix factorization with convex-concave divergences. In Subsection 3.3.1, we notably develop generic updates that guarantee the monotonic decrease of the cost function under mild assumptions. In Subsections 3.3.2, 3.3.3, 3.3.4, and 3.3.5, we then obtain specific forms of the generic updates for common information divergences. In particular, the proposed generic updates are specialized to provide updates for Csiszár divergences, certain skew Jeffreys-Bregman divergences, skew Jensen-Bregman divergences. It leads notably to several known multiplicative updates, as well as novel ones, for α-divergences, β-divergences, and their skew versions. We also generalize these results by considering the family of skew (α, β, λ)-divergences, for which multiplicative updates are derived in certain parameter regions.

3.2. Optimization framework

In this section, we formalize a standard optimization framework for NMF. The non-negative decomposition problem is considered as the minimization of a cost function built with a given separable divergence. Since this problem cannot be solved straightforward in general, we then introduce variational bounding as a generic optimization method for optimizing the cost function iteratively.

3.2.1. Cost function minimization problem

We formulate the reduced problem of non-negative decomposition, where we keep the dictionary matrix fixed, and seek encoding coefficients of the observations into this dictionary.

Problem 3.1. Let $\mathbf{y} \in \mathbb{R}_+^m$ be an observation vector of size $m \geq 1$, and let $\mathbf{A} \in \mathbb{R}_+^{m \times r}$ be a dictionary matrix made of $r \geq 1$ basis vectors. The problem of *non-negative* *decomposition* is to find an encoding vector $\mathbf{x} \in \mathbb{R}_+^r$ of $\mathbf{y}$ into $\mathbf{A}$ such that the approximation $\mathbf{y} \approx \mathbf{A}\mathbf{x}$ is of sufficient quality with respect to a given criterion.

Remark 3.1. A more general decomposition problem with arbitrary or no constraints can be formulated straightforward. The non-negativity constraints, or even positivity constraints, will however reveal crucial in the derivation of a generic algorithm for decomposition with convex-concave divergences, because of the need for positive weights in Jensen's inequality and of the multiplicative group structure of $\mathbb{R}_+^*$ for stability under multiplication and inversion.

As discussed above, we measure the quality of the approximate factorization via a cost function built with a separable divergence D on a set $\mathcal{Y} = Y^m$, generated by a given scalar divergence d on Y. In the sequel, we suppose without loss of generality that Y is contained in $\mathbb{R}_+$, otherwise it suffices to consider a restricted subset $Y \cap \mathbb{R}_+$. For the non-negative decomposition problem to make sense, we further suppose that Y is non-empty and that the observation vector $\mathbf{y}$ belongs to $\mathcal{Y}$. We now define the set of encoding vectors $\mathbf{x}$ that are feasible for the problem of non-negative decomposition.

Definition 3.1. The *feasible set* is the set $\mathcal{X}$ defined as follows:

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}_+^r \colon \mathbf{A}\mathbf{x} \in \mathcal{Y}\} \ . \tag{3.1}$$

Remark 3.2. The feasible set gathers all encoding vectors for which $\mathbf{A}\mathbf{x}$ lies in $\mathcal{Y}$ so that we can measure the quality of the approximate factorization via the separable divergence. For other encoding vectors, the problem does not make sense anymore since we cannot measure this quality according to the chosen criterion.

Definition 3.2. The *cost function* is the function C defined as follows:

$$C(\mathbf{x}) = D(\mathbf{y}\|\mathbf{A}\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathcal{X} \ . \tag{3.2}$$

Remark 3.3. In this definition, we employ the convention that the observations y_i are in the first argument of the scalar divergence. This should not confuse the reader: there is no loss of generality with this convention, and we can consider left- and right-sided problems, or obviously symmetric problems if the scalar divergence is symmetric. For a right-sided problem, it suffices to replace the scalar divergence $d(y\|y')$ with the scalar divergence $d^*(y\|y') = d(y'\|y)$ with swapped arguments.

Remark 3.4. We remark that even if the divergence is separable, the cost function cannot be seen as a separable function on the entries of the encoding vector $\mathbf{x}$ in general. It is however separable on the entries of the observation vector $\mathbf{y}$ since it can be developed as follows:

$$C(\mathbf{x}) = \sum_{i=1}^{m} d\left(y_i \,\middle\|\, \sum_{k=1}^{r} a_{ik} x_k\right) \ . \tag{3.3}$$

Remark 3.5. Using the cost function $C(\mathbf{x})$, the NMF problem can be seen as an optimization problem, more precisely a constrained minimization problem:

$$\text{minimize } C(\mathbf{x}) \quad \text{subject to } \mathbf{x} \in \mathcal{X} \ . \tag{3.4}$$

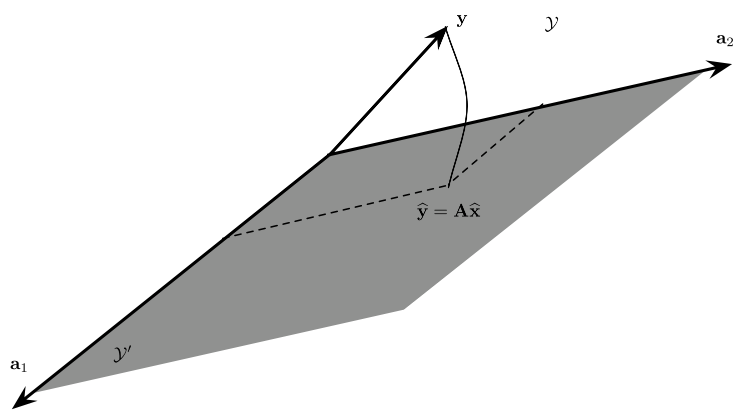

Figure 3.2.: Geometrical interpretation of non-negative decomposition. The optimization problem of non-negative decomposition can be seen as a projection of the observation vector onto the intersection of the conical hull of the basis vectors with the domain of the divergence.

In general, this problem cannot be solved straightforward and iterative methods for optimization are thus employed. Nevertheless, we notice that in an alternating update scheme for solving the NMF problem with respect to both factors, we do not need necessarily to solve exactly the above non-negative decomposition subproblem, but rather to find an encoding vector $\mathbf{x} \in \mathcal{X}$ that makes the cost function decrease compared to the current encoding vector $\bar{\mathbf{x}} \in \mathcal{X}$.

Remark 3.6. The non-negative decomposition problem also has a nice geometrical interpretation. Indeed, under some regularity assumptions, the problem can be seen as a right-sided projection $\widehat{\mathbf{y}} = \arg\min_{\mathbf{y}' \in \mathcal{Y}'} D(\mathbf{y} \| \mathbf{y}')$ of the point $\mathbf{y} \in \mathcal{Y}$ onto the subset $\mathcal{Y}' \subseteq \mathcal{Y}$ with respect to the divergence D, where $\mathcal{Y}'$ is the intersection of the conical hull of the set of basis vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ with the domain $\mathcal{Y}$ of the divergence, as illustrated in Figure 3.2. The optimal encoding vector $\widehat{\mathbf{x}}$ then represents the non-negative coordinates of the projection $\widehat{\mathbf{y}}$ with respect to the basis vectors, provided that the dictionary matrix $\mathbf{A}$ is of full rank. Solving for this projection is in general not easy, even when $\mathcal{Y}'$ is convex, in part because it is considered with respect to a divergence and not the Euclidean distance, but also because of the additional domain constraints on the problem. We notice however that for the left-sided decomposition problem with Bregman divergences, this projection can be solved under mild assumptions by using alternate Bregman projections on hyperplanes [Dhillon and Tropp 2008]. This procedure is yet slow to converge. Moreover, it cannot be extended in general to right-sided Bregman divergences or other divergences.

3.2.2. Variational bounding and auxiliary functions

As discussed above, the cost function in the non-negative decomposition problem is not straightforward to optimize in general. To address this, we rely on the framework of *variational bounding*, also called *majorization*, which is an iterative technique

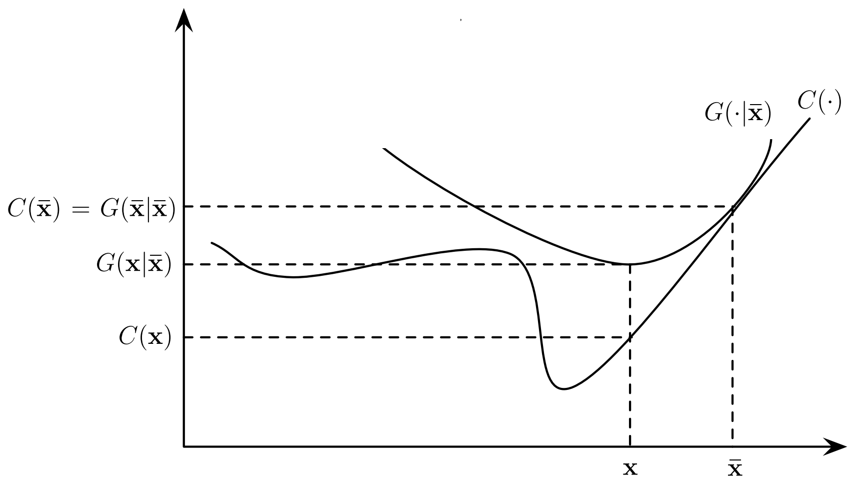

Figure 3.3.: Auxiliary function for the cost function. The auxiliary function defines a majorizing function above the current solution, which can be used as a surrogate for optimizing the cost function.

for minimization problems where we replace the cost function at each step with a surrogate majorizing function that we optimize instead [Rustagi 1976, Hunter and Lange 2004]. This technique is in general employed for solving optimization problems with possibly non-convex cost functions, and often reveals efficient provided that the majorizing auxiliary functions at each step are easier to optimize than the original cost function. The difficulty lies in choosing appropriate majorizing functions, that can be optimized efficiently, and that provide tight bounds in order to fit well the original cost function and make it decrease fast enough. Such majorizing functions can be defined in general terms as follows.

Definition 3.3. An *auxiliary function* is a function $G\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $G(\bar{\mathbf{x}}|\bar{\mathbf{x}}) = C(\bar{\mathbf{x}})$ and $G(\mathbf{x}|\bar{\mathbf{x}}) \geq C(\mathbf{x})$ for all $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}$.

Remark 3.7. We actually need not always define an auxiliary function everywhere on $\mathcal{X} \times \mathcal{X}$. If the updates have to stay in a given subset of $\mathcal{X}' \subseteq \mathcal{X}$, we can just define the auxiliary function on $\mathcal{X}' \times \mathcal{X}'$ and optimize it on $\mathcal{X}'$, provided that the algorithm is correctly initialized with a point in $\mathcal{X}'$. This is sometimes the case for non-negative decomposition with multiplicative updates, where we want the updates to stay in $\mathbb{R}_+^*$ instead of $\mathbb{R}_+$ in order to avoid divisions by zero and trivial fixed points that are not optimal. Of course, it is nonetheless possible that the algorithm converges to a boundary point while staying in the interior so we do not exclude these potential solutions of the problem.

The interest of using an auxiliary function lies in the fact that if we can optimize it, then we can also make the original cost function decrease, that is, we can find a new point $\mathbf{x} \in \mathcal{X}$ that improves the cost function compared to the current point $\bar{\mathbf{x}} \in \mathcal{X}$. This is illustrated in Figure 3.3 and formalized in the following lemma.

Lemma 3.1. *Let* $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}$. *If* $G(\mathbf{x}|\bar{\mathbf{x}}) \leq G(\bar{\mathbf{x}}|\bar{\mathbf{x}})$, *then* $C(\mathbf{x}) \leq C(\bar{\mathbf{x}})$.

Proof. Let $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}$. By definition, we have $C(\mathbf{x}) \leq G(\mathbf{x}|\bar{\mathbf{x}})$ and $C(\bar{\mathbf{x}}) = G(\bar{\mathbf{x}}|\bar{\mathbf{x}})$. Now if $G(\mathbf{x}|\bar{\mathbf{x}}) \leq G(\bar{\mathbf{x}}|\bar{\mathbf{x}})$, then we have $C(\mathbf{x}) \leq G(\mathbf{x}|\bar{\mathbf{x}}) \leq G(\bar{\mathbf{x}}|\bar{\mathbf{x}}) = C(\bar{\mathbf{x}})$, which proves the lemma. □

Remark 3.8. We also have strict decrease of the cost function as soon as we choose a vector $\mathbf{x}$ that makes the auxiliary function strictly decrease as $G(\mathbf{x}|\bar{\mathbf{x}}) < G(\bar{\mathbf{x}}|\bar{\mathbf{x}})$.

Remark 3.9. This justifies the use of an auxiliary function to minimize or at least make the original cost function decrease. Indeed, if the current solution is given by $\bar{\mathbf{x}} \in \mathcal{X}$, then choosing a point $\mathbf{x} \in \mathcal{X}$ such that $G(\mathbf{x}|\bar{\mathbf{x}}) \leq G(\bar{\mathbf{x}}|\bar{\mathbf{x}})$ provides a better solution. This may be iterated until a termination criterion is met. In general, when it is possible, the point $\mathbf{x}$ is chosen as a minimizer of the auxiliary function at $\bar{\mathbf{x}}$, so that we need to solve the following optimization problem:

$$\text{minimize } G(\mathbf{x}, \bar{\mathbf{x}}) \quad \text{subject to } \mathbf{x} \in \mathcal{X} \ . \tag{3.5}$$

The minimization can be done on an arbitrary subset $\mathcal{X}' \subseteq \mathcal{X}$ as soon as $\bar{\mathbf{x}} \in \mathcal{X}'$. It is also possible to equalize the auxiliary function instead, or to choose any point in between the minimization and the equalization.

We show in the sequel that we can build auxiliary functions for a wide range of common information divergences presented in Chapter 1. We can therefore optimize the respective cost functions by variational bounding. We will focus on *majorization-minimization* schemes where the auxiliary function is iteratively minimized to update the solution as discussed above.

3.3. Methods for convex-concave divergences

In this section, we elaborate on the proposed methods for solving the problem of non-negative decomposition with convex-concave divergences. We first develop generic updates by constructing and minimizing a general auxiliary function. We then particularize these generic updates to common families of information divergences, including Csiszár divergences, skew Jeffreys-Bregman divergences, skew Jensen-Bregman divergences, and skew (α, β, λ)-divergences. It leads to known and novel closed-form multiplicative updates for all α-divergences, β-divergences, almost all (α, β)-divergences, and certain of their symmetric or skew versions.

3.3.1. Generic updates

Up from now, we restrict to the case where $Y = \mathbb{R}_+^*$ since it concerns the scalar divergences considered later. This ensures that Y is stable under multiplication and inversion, which reveals useful in the subsequent derivations. Moreover, it guarantees that the feasible set $\mathcal{X}$ contains at least all positive vectors as soon as it is non-empty.

Lemma 3.2. *If $Y = \mathbb{R}_+^*$, then $\mathcal{X}$ is non-empty iff $(\mathbb{R}_+^*)^r \subseteq \mathcal{X}$.*

Proof. On the one hand, if $(\mathbb{R}_+^*)^r \subseteq \mathcal{X}$, then $\mathcal{X}$ is clearly non-empty. On the other hand, if $\mathcal{X}$ is non-empty, then we remark that the dictionary $\mathbf{A}$ cannot have a null row. Indeed, if it were the case, then the corresponding row of $\mathbf{Ax}$ would be null for any encoding vector $\mathbf{x} \in \mathbb{R}_+^r$, so that $\mathcal{X}$ would be empty, leading to a contradiction. As a result, there is no null row in $\mathbf{A}$, and for any encoding vector $\mathbf{x} \in (\mathbb{R}_+^*)^r$, we also have $\mathbf{A}\mathbf{x} \in (\mathbb{R}_+^*)^m$, so that $\mathbf{x} \in \mathcal{X}$ and $(\mathbb{R}_+^*)^r \subseteq \mathcal{X}$. □

Remark 3.10. We also have equivalence with the dictionary matrix $\mathbf{A}$ having no null row. If it were the case, then the non-negative decomposition would be degenerate and we could remove this row of $\mathbf{A}$ and the corresponding entry of the observation vector $\mathbf{y}$, to end up with a non-degenerate problem with this respect.

We therefore suppose that the feasible set $\mathcal{X}$ is non-empty, meaning that the non-negative decomposition problem is feasible. In other terms there exists at least one encoding vector $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} \in \mathcal{Y}$, and thus it is relevant to search for an encoding vector that leads to a sufficient or at least improved quality of factorization compared to others. It also implies that the feasible set $\mathcal{X}$ contains the entire positive orthant. Assuming now that the scalar divergence d is convex-concave, meaning that it can be expressed as the sum of a convex part and of a concave part with respect to either one of the two arguments, a convenient auxiliary function can be built by using Jensen's inequality for the convex part and the tangent inequality for the concave part. During optimization based on variational bounding with such auxiliary functions, we stay in the positive orthant for technical validity. This is discussed hereafter, focusing on convex-concave divergences in the second argument without loss of generality.

Proposition 3.3. *Suppose that the scalar divergence d is a convex-concave function in the second argument, and that the decomposition $d = \breve{d} + \widehat{d}$, where $\breve{d}$ and $\widehat{d}$ are respectively convex and concave in the second argument, can be chosen such that $\widehat{d}$ is differentiable in the second argument. Then, we can define an auxiliary function G for the cost function C as follows:*

$$G(\mathbf{x}|\bar{\mathbf{x}}) = \sum_{i=1}^{m} \left\{ \widehat{d}\left(y_i \,\middle\|\, \sum_{l=1}^{r} a_{il}\bar{x}_l \right) + \sum_{k=1}^{r} \frac{a_{ik}\bar{x}_k}{\sum_{l=1}^{r} a_{il}\bar{x}_l}\, \breve{d}\left(y_i \,\middle\|\, \sum_{l=1}^{r} a_{il}\bar{x}_l\, \frac{x_k}{\bar{x}_k} \right) \right.$$
$$\left. + \sum_{k=1}^{r} a_{ik}(x_k - \bar{x}_k)\, \frac{\partial \widehat{d}}{\partial y'}\left(y_i \,\middle\|\, \sum_{l=1}^{r} a_{il}\bar{x}_l \right) \right\} \quad \textit{for all } \mathbf{x}, \bar{\mathbf{x}} \in (\mathbb{R}_+^*)^r \ . \quad (3.6)$$

Proof. Let $\mathbf{x}, \bar{\mathbf{x}} \in (\mathbb{R}_+^*)^r \subseteq \mathcal{X}$. We separate the cost function $C(\mathbf{x}) = \sum_{i=1}^{m} C_i(\mathbf{x})$ as the element-wise scalar divergences $C_i(\mathbf{x}) = d(y_i \| \sum_{k=1}^{r} a_{ik}x_k)$ on the entries of the observation vector. We further decompose the element-wise cost functions $C_i(\mathbf{x}) = \breve{C}_i(\mathbf{x}) + \widehat{C}_i(\mathbf{x})$, into convex and concave parts $\breve{C}_i(\mathbf{x}) = \breve{d}(y_i \| \sum_{k=1}^{r} a_{ik}x_k)$ and $\widehat{C}_i(\mathbf{x}) = \widehat{d}(y_i \| \sum_{k=1}^{r} a_{ik}x_k)$. Our aim is to bound separately $\breve{C}_i, \widehat{C}_i$, with auxiliary functions $\breve{G}_i, \widehat{G}_i$, and then sum up everything to provide a global auxiliary function G for the cost function C. Beginning with the convex parts, we define the auxiliary functions as follows:

$$\breve{G}_i(\mathbf{x}|\bar{\mathbf{x}}) = \sum_{k=1}^{r} \frac{a_{ik}\bar{x}_k}{\sum_{l=1}^{r} a_{il}\bar{x}_l}\, \breve{d}\left(y_i \,\middle\|\, \sum_{l=1}^{r} a_{il}\bar{x}_l\, \frac{x_k}{\bar{x}_k} \right) \ . \quad (3.7)$$

The function $\breve{G}_i$ is well-defined since $\sum_{l=1}^r a_{il}\bar{x}_l \neq 0$, $\bar{x}_k \neq 0$, and $\sum_{l=1}^r a_{il}\bar{x}_l \frac{x_k}{\bar{x}_k} \in \mathbb{R}_+^*$, for all $1 \leq k \leq r$. We also have $\breve{G}_i(\bar{\mathbf{x}}|\bar{\mathbf{x}}) = \breve{C}_i(\bar{\mathbf{x}})$ which can be seen from:

$$\breve{G}_i(\bar{\mathbf{x}}|\bar{\mathbf{x}}) = \sum_{k=1}^r \frac{a_{ik}\bar{x}_k}{\sum_{l=1}^r a_{il}\bar{x}_l}\, \breve{d}\left(y_i \,\middle\|\, \sum_{l=1}^r a_{il}\bar{x}_l\right) \tag{3.8}$$

$$= \breve{d}\left(y_i \,\middle\|\, \sum_{l=1}^r a_{il}\bar{x}_l\right) \tag{3.9}$$

$$= \breve{C}_i(\bar{\mathbf{x}}) \ . \tag{3.10}$$

Moreover, we have $\breve{G}_i(\mathbf{x}|\bar{\mathbf{x}}) \geq \breve{C}_i(\mathbf{x})$ as a result of Jensen's inequality for the function $\breve{d}$, which is convex in the second argument, and using the weights $a_{ik}\bar{x}_k \in \mathbb{R}_+^*$ normalized by their positive sum:

$$\breve{G}_i(\mathbf{x}|\bar{\mathbf{x}}) \geq \breve{d}\left(y_i \,\middle\|\, \sum_{k=1}^r \frac{a_{ik}\bar{x}_k}{\sum_{l=1}^r a_{il}\bar{x}_l} \sum_{l=1}^r a_{il}\bar{x}_l \frac{x_k}{\bar{x}_k}\right) \tag{3.11}$$

$$= \breve{d}\left(y_i \,\middle\|\, \sum_{k=1}^r a_{ik}x_k\right) \tag{3.12}$$

$$= \breve{C}_i(\mathbf{x}) \ . \tag{3.13}$$

This proves that $\breve{G}_i$ is indeed an auxiliary function for $\breve{C}_i$. We now turn to the concave parts, and define the auxiliary functions as follows:

$$\widehat{G}_i(\mathbf{x}|\bar{\mathbf{x}}) = \widehat{d}\left(y_i \,\middle\|\, \sum_{l=1}^r a_{il}\bar{x}_l\right) + \sum_{k=1}^r a_{ik}(x_k - \bar{x}_k)\, \frac{\partial \widehat{d}}{\partial y'}\left(y_i \,\middle\|\, \sum_{l=1}^r a_{il}\bar{x}_l\right) \ . \tag{3.14}$$

The function $\widehat{G}_i$ is well-defined since the function $\widehat{d}$ is differentiable in the second argument. We also have $\widehat{G}_i(\bar{\mathbf{x}}|\bar{\mathbf{x}}) = \widehat{C}_i(\bar{\mathbf{x}})$ which can be seen from:

$$\widehat{G}_i(\bar{\mathbf{x}}|\bar{\mathbf{x}}) = \widehat{d}\left(y_i \,\middle\|\, \sum_{l=1}^r a_{il}\bar{x}_l\right) \tag{3.15}$$

$$= \widehat{C}_i(\bar{\mathbf{x}}) \ . \tag{3.16}$$

Moreover, we have $\widehat{G}_i(\mathbf{x}|\bar{\mathbf{x}}) \geq \widehat{C}_i(\mathbf{x})$, which arises from the tangent inequality applied to the differentiable concave functions $\widehat{C}_i$ as follows:

$$\widehat{G}_i(\mathbf{x}|\bar{\mathbf{x}}) = \widehat{C}_i(\bar{\mathbf{x}}) + (\mathbf{x} - \bar{\mathbf{x}})^\top \nabla \widehat{C}_i(\bar{\mathbf{x}}) \tag{3.17}$$

$$\geq \widehat{C}_i(\mathbf{x}) \ . \tag{3.18}$$

This proves that $\widehat{G}_i$ is indeed an auxiliary function for $\widehat{C}_i$. Putting everything together, we conclude that $G(\mathbf{x}|\bar{\mathbf{x}}) = \sum_{i=1}^m \breve{G}_i(\mathbf{x}|\bar{\mathbf{x}}) + \widehat{G}_i(\mathbf{x}|\bar{\mathbf{x}})$ is an auxiliary function for the cost function $C(\mathbf{x}) = \sum_{i=1}^m \breve{C}_i(\mathbf{x}) + \widehat{C}_i(\mathbf{x})$, which proves the proposition. □

Remark 3.11. The framework of convex-concave functions applies to many common information divergences, except for some Bregman divergences in the second argument and their skew Jeffreys-Bregman versions. The assumption that the concave part is differentiable is not too restrictive since most cost functions for non-negative decomposition are smooth. Moreover, it is well-known that an arbitrary concave function is actually differentiable at all but at most countably many points. Finally, it is also possible to extend the results by considering subgradients where the function is not differentiable.

Remark 3.12. The decomposition $d = \check{d} + \hat{d}$ is clearly arbitrary up to adding and subtracting the same differentiable convex function respectively to $\check{d}$ and $\hat{d}$. For the divergences considered here, it appears that there is a somewhat canonical decomposition, but using other decompositions may lead to different results since the auxiliary function depends on this decomposition.

To optimize the cost function based on majorization-minimization, we need to minimize the above auxiliary functions at each step. It guarantees the monotonic decrease of the cost function as discussed below.

Theorem 3.4. *Suppose that the scalar divergence d is a convex-concave function in the second argument, and that the decomposition $d = \check{d} + \hat{d}$, where $\check{d}$ and $\hat{d}$ are respectively convex and concave in the second argument, can be chosen such that $\check{d}$ and $\hat{d}$ are differentiable in the second argument. Then, for all $\bar{\mathbf{x}} \in (\mathbb{R}_+^*)^r$, we have $C(\mathbf{x}) \leq C(\bar{\mathbf{x}})$ for any point $\mathbf{x} \in (\mathbb{R}_+^*)^r$ that verifies the following system of equations:*

$$\sum_{i=1}^{m} a_{ik} \frac{\partial \check{d}}{\partial y'} \left(y_i \middle\| \sum_{l=1}^{r} a_{il} \bar{x}_l \, \frac{x_k}{\bar{x}_k} \right) = - \sum_{i=1}^{m} a_{ik} \frac{\partial \hat{d}}{\partial y'} \left(y_i \middle\| \sum_{l=1}^{r} a_{il} \bar{x}_l \right) \quad \textit{for all } k \in [\![1, r]\!] \ . \tag{3.19}$$

Proof. Let $\bar{\mathbf{x}} \in (\mathbb{R}_+^*)^r$. The auxiliary function G at $\bar{\mathbf{x}}$ can be separated on the entries of the encoding vector $\mathbf{x} \in (\mathbb{R}_+^*)^r$ as $G(\mathbf{x}|\bar{\mathbf{x}}) = \hat{d}(y_i \| \sum_{l=1}^{r} a_{il}\bar{x}_l) + \sum_{k=1}^{r} g_k(x_k|\bar{\mathbf{x}})$, where $g_k(x_k \| \bar{\mathbf{x}})$ is defined as follows:

$$g_k(x_k|\bar{\mathbf{x}}) = \sum_{i=1}^{m} \frac{a_{ik}\bar{x}_k}{\sum_{l=1}^{r} a_{il}\bar{x}_l} \, \check{d}\left(y_i \middle\| \sum_{l=1}^{r} a_{il} \bar{x}_l \, \frac{x_k}{\bar{x}_k} \right) + a_{ik}(x_k - \bar{x}_k) \, \frac{\partial \hat{d}}{\partial y'} \left(y_i \middle\| \sum_{l=1}^{r} a_{il} \bar{x}_l \right) \ . \tag{3.20}$$

The auxiliary function G at $\bar{\mathbf{x}}$ being convex and differentiable, it attains its minimum at a point $\mathbf{x} \in (\mathbb{R}_+^*)^r$ iff we have $\partial_{x_k} G(\mathbf{x}|\bar{\mathbf{x}}) = 0$, for any $1 \leq k \leq r$. Since the respective derivatives can be developed as $\partial_{x_k} G(\mathbf{x}|\bar{\mathbf{x}}) = \partial_{x_k} g_k(x_k|\bar{\mathbf{x}})$, the minimum is attained at the points $\mathbf{x}$ that are solutions of the following system of equations:

$$\sum_{i=1}^{m} a_{ik} \frac{\partial \check{d}}{\partial y'} \left(y_i \middle\| \sum_{l=1}^{r} a_{il} \bar{x}_l \, \frac{x_k}{\bar{x}_k} \right) + a_{ik} \frac{\partial \hat{d}}{\partial y'} \left(y_i \middle\| \sum_{l=1}^{r} a_{il} \bar{x}_l \right) = 0 \ . \tag{3.21}$$

Moreover, these global minima are such that $G(\mathbf{x}|\bar{\mathbf{x}}) \leq G(\bar{\mathbf{x}}|\bar{\mathbf{x}})$, and thus verify $C(\mathbf{x}) \leq C(\bar{\mathbf{x}})$ since G is an auxiliary function for C. □

Remark 3.13. The assumptions imply that d needs to be differentiable, but for the same reasons as discussed above, this is not too restrictive in our context.

Remark 3.14. We notice that if a column of the dictionary $\mathbf{A}$ is null, then the corresponding equation is always verified, simply meaning that adding this null column to any extent in the decomposition does not change anything. As a result, the problem is degenerate and we can remove this column from the dictionary and the corresponding entry of the encoding vector, to end up with a non-degenerate problem with this respect.

Remark 3.15. It is interesting to remark that because of the differentiation, these equations do not depend on arbitrary affine terms in the decomposition into convex and concave parts. As a result, constant terms can clearly be omitted in the decomposition, and linear terms can be put either in the convex part or in the concave part without changing the updates. The updates may however depend on other arbitrary convex terms that are not affine. For the considered information divergences, there is actually a somewhat canonical decomposition that leads to simple solutions as discussed later.

Remark 3.16. Here we considered optimization by majorization-minimization, and thus set the gradient of the auxiliary function to zero to obtain the system of equations for the updates. This scheme works as soon as the convex auxiliary function attains its minimum inside the positive orthant. When it is not the case, then an infimum is actually attained at a limit point. Since the auxiliary function can be separated on the respective dimensions, the search for such a limit point can also be separated across the dimensions. Considering a given element-wise convex auxiliary function, we notice that it cannot be affine with a negative slope, otherwise the majorized cost function would become negative at some point. As a result, either its minimum is attained in $\mathbb{R}_+^*$, or its infimum is attained at zero. Therefore, setting the updated coordinate anywhere on the open segment in between the current coordinate solution and zero would make the cost function decrease. The zero value should however be avoided for technicality reasons, even if the sequence of updates may converge to this point. We yet do not need such updates here since it does not concern the information divergences considered.

Remark 3.17. At this point, it is worth mentioning some relations between the proposed method and the concave-convex procedure (CCCP) [Yuille and Rangarajan 2003]. The CCCP lies in the same framework of variational bounding, but only considers the tangent inequality to majorize the concave part of the convex-concave cost function C. It then minimizes the convex auxiliary function by setting its gradient to zero. It leads to the update $\mathbf{x}$ of $\bar{\mathbf{x}}$ as $\nabla \check{C}(\mathbf{x}) = -\nabla \hat{C}(\bar{\mathbf{x}})$ when such a point exists. For several convex-concave problems, the convex part is actually strictly convex and its gradient one-to-one with a closed-form inverse, so that the updates can be computed efficiently. Applying this to the non-negative decomposition framework, however, is not straightforward in general, because of the linear model and the non-negative constraints. Indeed, the system of equations of the CCCP would write as follows:

$$\sum_{i=1}^{m} a_{ik} \frac{\partial \check{d}}{\partial y'} \left(y_i \middle\| \sum_{l=1}^{r} a_{il} x_l \right) = - \sum_{i=1}^{m} a_{ik} \frac{\partial \hat{d}}{\partial y'} \left(y_i \middle\| \sum_{l=1}^{r} a_{il} \bar{x}_l \right) . \tag{3.22}$$

Because of the weighted sum of derivatives, which arises from the linear model, the solutions of this system of equations is not available analytically in general, even if the inverse mapping of $\partial_{y'}\breve{d}$ is in closed form. As a result, we would need to solve for a system of r equations which all depend on the full vector $\mathbf{x}$ of dimension r. Moreover, there is a priori no consideration of the non-negative constraints when using this procedure. This makes the CCCP unsuited, even for the standard Euclidean cost function. In the proposed approach, we further use the Jensen's inequality for the convex part of the cost function. On the one hand, it results in general in a looser bound than for the CCCP. On the other hand, it greatly simplifies the minimization of the auxiliary function. Indeed, it allows separating the auxiliary function and its optimization on the entries of $\mathbf{x}$, thus leading to r independent equations in dimension one, which can be solved much more efficiently in general. Moreover, the non-negativity assumptions are taken into account in the weights of Jensen's inequality. Even if it does not guarantee a priori the non-negativity of the updates in general, we see later that it is nonetheless the case for many divergences through the derivation of closed-form multiplicative updates expressed with these weights.

Remark 3.18. A reasonable condition for the system of equations to admit an analytical solution, is that the derivative $f(y') = \partial_{y'}\breve{d}(y\|y')$ of the convex part allows separating the ratio $x_k/\bar{x}_k$ from the sum $\sum_{l=1}^{r} a_{il}\bar{x}_l$, either by taking a multiplicative form $f(y'_1 y'_2) = g(y'_1)h(y'_2)$, or an additive form $f(y'_1 y'_2) = g(y'_1) + h(y'_2)$, up to additive constants. Under mild assumptions, it is well-known that these two Pexider's functional equations admit respective solutions in the form of power functions $f(y') = k y'^p$, and of logarithmic functions $f(y') = k \log y'$, where $k, p \in \mathbb{R}$. Integrating these functions leads again to power functions, logarithmic functions, or even functions of the form $k\, y' \log y'$, up to affine terms. It therefore seems intuitive that scalar divergences defined using one of these three forms would provide analytical solutions for the updates of the non-negative decomposition. Combinations of such functions may also lead to closed-form updates depending on how the different terms simplify. This highlights the interests in using scalar divergences such as the α-divergences, β-divergences, and (α, β)-divergences, not only from a statistical and information-theoretic standpoint, but also from a computational perspective.

This theorem provides a generic method to update the encoding vector while ensuring monotonic decrease of a convex-concave cost function. For this, we need to solve r independent equations of dimension one. Provided that a solution does exist, solving the system can be done iteratively by using simple line search methods, or more elaborate and efficient schemes such as Newton's methods. In the sequel, we show that for common information divergences, these equations simplify further, and sometimes specifically lead to convenient closed-form multiplicative updates. As discussed above, we will assume without loss of generality that the dictionary matrix $\mathbf{A}$ has no null row nor column to derive these updates.

3.3.2. Case of Csiszár divergences

When considering the class of Csiszár divergences, the proposed generic method can be simplified in terms of the generator function φ. Since this class is stable under swapping the arguments and standard skewing, we specify without loss of generality the form taken by the updates for a right-sided non-negative decomposition problem.

Corollary 3.5. *Consider the right-sided non-negative decomposition problem with the Csiszár φ-divergence $d_\varphi^{(C)}$. For all $\bar{\mathbf{x}} \in (\mathbb{R}_+^*)^r$, we have $C(\mathbf{x}) \leq C(\bar{\mathbf{x}})$ for any point $\mathbf{x} \in (\mathbb{R}_+^*)^r$ that verifies the following system of equations:*

$$\sum_{i=1}^{m} a_{ik} \varphi' \left(\frac{1}{y_i} \sum_{l=1}^{r} a_{il} \bar{x}_l \frac{x_k}{\bar{x}_k} \right) = 0 \quad \text{for all } k \in [\![1, r]\!] \ . \tag{3.23}$$

Proof. The Csiszár φ-divergence $d_\varphi^{(C)}$ is clearly differentiable and convex in the second argument. We can thus decompose it in respective differentiable convex and concave parts $\breve{d}_\varphi^{(C)}(y\|y') = y\,\varphi(y'/y)$, and $\widehat{d}_\varphi^{(C)}(y\|y') = 0$. Applying the generic method for convex-concave divergences, the corollary follows by remarking that $\partial_{y'} \breve{d}_\varphi^{(C)}(y\|y') = \varphi'(y'/y)$, and that $\partial_{y'} \widehat{d}_\varphi^{(C)}(y\|y') = 0$. □

Example 3.1. For the right-sided non-negative decomposition problem with the α-divergences $d_\alpha^{(a)}$, where $\alpha \neq 0, 1$, the system of equations leads to the following closed-form multiplicative updates as the unique solution:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^{m} a_{ik} (y_i / \sum_{l=1}^{r} a_{il} \bar{x}_l)^\alpha}{\sum_{i=1}^{m} a_{ik}} \right)^{1/\alpha} \ . \tag{3.24}$$

For $\alpha = 1$, corresponding to the Kullback-Leibler divergence, solving the equations shows that the above multiplicative updates actually still hold. These updates coincide with that proposed by Cichocki et al. [2008], where the case $\alpha = 0$ is yet omitted. For $\alpha = 0$, corresponding to the dual Kullback-Leibler divergence, the updates take a different form as follows:

$$x_k = \bar{x}_k \times \exp \left(\frac{\sum_{i=1}^{m} a_{ik} \log (y_i / \sum_{l=1}^{r} a_{il} \bar{x}_l)}{\sum_{i=1}^{m} a_{ik}} \right) \ . \tag{3.25}$$

These updates were obtained for example by Dhillon and Sra [2006] as a left-sided problem with the Kullback-Leibler divergence seen as a Bregman divergence. Considering left-sided problems with α-divergences is straightforward since it actually suffices by symmetry to replace α with $1 - \alpha$, and apply the corresponding above updates.

Example 3.2. For the right-sided non-negative decomposition problem with the skew Jeffreys (α, λ)-divergences $d_{\alpha,\lambda}^{(a)}$, where $\alpha \neq 0, 1$, the system of equations can be developed as follows:

$$\begin{aligned} \lambda(1-\alpha) \sum_{i=1}^{m} a_{ik} \left(\frac{1}{y_i} \sum_{l=1}^{r} a_{il} \bar{x}_l \right)^{-\alpha} \left(\frac{x_k}{\bar{x}_k} \right)^{-\alpha} & \\ + (1-\lambda)\alpha \sum_{i=1}^{m} a_{ik} \left(\frac{1}{y_i} \sum_{l=1}^{r} a_{il} \bar{x}_l \right)^{\alpha-1} \left(\frac{x_k}{\bar{x}_k} \right)^{\alpha-1} & = (\alpha + \lambda - 2\alpha\lambda) \sum_{i=1}^{m} a_{ik} \ . \end{aligned} \tag{3.26}$$

Unfortunately, it does not admit a closed-form solution in the general case, and iterative methods are required as discussed above. It is neither the case when $\alpha \in \{0, 1\}$, corresponding to the skew Jeffreys divergence as a skew version of the Kullback-Leibler divergence, where the unknown variables appear both as logarithmic and inverse terms. For certain values of α, the equations can be written as polynomial equations, which can be solved more efficiently with specific methods such as root-finding algorithms, or with analytical solutions for lower degrees, when positive solutions do exist. For example, for $\alpha \in \{-1, 2\}$, corresponding respectively to skew versions of the Neyman's and Pearson's chi-square distances, we end up with a polynomial equation of degree three, while for $\alpha \in \{-1/2, 3/2\}$, we have a polynomial equation of degree four. Obviously, for $\alpha = 1/2$, corresponding to the symmetric Hellinger distance, the skewness has no effect and we have the same solutions for any value of $\lambda \in [0, 1]$, given as a special case of the above non-skew multiplicative updates. For $\lambda \in \{0, 1\}$, corresponding to a left- or right-sided non-skew problem, the updates also correspond to the above non-skew multiplicative updates. For the left-sided non-negative decomposition problem with the skew Jeffreys (α, λ)-divergences $d^{(a)}_{\alpha,\lambda}$, it actually suffices by symmetry to replace either α with $1-\alpha$, or λ with $1-\lambda$, to end up with a right-sided problem.

Example 3.3. A novel analytical scheme can be derived for a specific one-parameter family of skew Jeffreys (α, λ)-divergences. Indeed, if we consider the right-sided problem with the skew Jeffreys (α, λ)-divergences $d^{(a)}_{\alpha,\lambda}$, where $\alpha \in \mathbb{R} \setminus [0, 1]$, and $\lambda = \alpha/(2\alpha - 1) \in (0, 1) \setminus \{1/2\}$, the system of equations leads to the following closed-form multiplicative updates as the unique solution:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} (y_i / \sum_{l=1}^r a_{il} \bar{x}_l)^{\alpha}}{\sum_{i=1}^m a_{ik} (y_i / \sum_{l=1}^r a_{il} \bar{x}_l)^{1-\alpha}} \right)^{1/(2\alpha-1)} . \tag{3.27}$$

Notice that we are able to obtain closed-form updates, thanks to the constraint $\lambda = \alpha/(2\alpha - 1)$ which allows the constant to vanish in the equations. We also have an inherent symmetry within the one-parameter family, where replacing α with $1-\alpha$, is equivalent to replacing λ with $1-\lambda$. As a result, we need only consider the values of $\alpha > 1$ because the other values are redundant. In the limit case $\alpha \in \{0, 1\}$, we have $\lambda = 0$, $\lambda = 1$, which corresponds to the equivalent non-skew problems of left-sided non-negative decomposition with the dual Kullback-Leibler divergence, and of right-sided non-negative decomposition with the Kullback-Leibler divergence, respectively. As a result, the updates are still valid and actually coincide with the non-skew multiplicative updates derived above. We can also consider straightforwardly the left-sided problems with a one-parameter family of skew (α, λ)-divergences, where the constraint now writes $\lambda = (\alpha - 1)/(2\alpha - 1)$, and we have equivalence with the right-sided problems by replacing α with $1 - \alpha$, and λ with $1 - \lambda$.

The class of Csiszár divergences is also stable under the second type of skewing introduced for Jensen-Bregman divergences. We were not however able to derive interesting results to simplify the proposed generic method for skew Jensen (α, λ)-divergences, except for negative integer values of α where we obtain polynomial equations, and in particular for $\alpha = -1$, corresponding to the Neyman's chi-square distance, where the polynomial equation is of degree four. For the sake of conciseness, we thus do not develop this type of skewing further.

3.3.3. Case of skew Jeffreys-Bregman divergences

On the contrary to Csiszár divergences, the class of Bregman divergences is not stable under swapping the arguments, so that the left-sided problems are not equivalent to right-sided problems in general. Moreover, even if the Bregman divergences are always convex in the first argument, they are not in general convex-concave in the second argument. As a result, we cannot specialize the proposed generic method to the non-negative decomposition problems with arbitrary Jeffreys-Bregman divergences, except for the specific non-skew left-sided problems with Bregman divergences. We therefore state only generic results for the latter problems, where the proposed generic method simplifies in terms of the generator function φ.

Corollary 3.6. *Consider the left-sided non-negative decomposition problem with the Bregman φ-divergence $d_\varphi^{(B)}$. For all $\bar{\mathbf{x}} \in (\mathbb{R}_+^*)^r$, we have $C(\mathbf{x}) \leq C(\bar{\mathbf{x}})$ for any point $\mathbf{x} \in (\mathbb{R}_+^*)^r$ that verifies the following system of equations:*

$$\sum_{i=1}^{m} a_{ik} \varphi' \left(\sum_{l=1}^{r} a_{il} \bar{x}_l \, \frac{x_k}{\bar{x}_k} \right) = \sum_{i=1}^{m} a_{ik} \varphi'(y_i) \quad \textit{for all } k \in [\![1, r]\!] \ . \tag{3.28}$$

Proof. The Bregman φ-divergence $d_\varphi^{(B)}$ is clearly differentiable and convex in the first argument. We can thus decompose the swapped divergence in respective differentiable convex and concave parts $\check{d}_\varphi^{(B)}(y\|y') = \varphi(y') - \varphi(y) - (y' - y)\varphi'(y)$, and $\hat{d}_\varphi^{(B)}(y\|y') = 0$. Applying the generic method for convex-concave divergences, the corollary follows by remarking that $\partial_{y'} \check{d}_\varphi^{(B)}(y\|y') = \varphi'(y') - \varphi'(y)$, and that $\partial_{y'} \hat{d}_\varphi^{(B)}(y\|y') = 0$. ☐

Remark 3.19. These updates coincide with that found by Dhillon and Sra [2006], where the proof is actually a special instance of our generic proof for convex-concave divergences applied to the Bregman divergences, which are convex in the first argument so that the concave part and the tangent inequality disappear, and just the convex part and Jensen's inequality are used.

Example 3.4. A novel analytical scheme can be derived for the left-sided problem with the β-divergences $d_\beta^{(b)}$, where $\beta \neq 0, 1$, for which the system of equations leads to the following closed-form multiplicative updates as the unique solution:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^{m} a_{ik} y_i^{\beta-1}}{\sum_{i=1}^{m} a_{ik} (\sum_{l=1}^{r} a_{il} \bar{x}_l)^{\beta-1}} \right)^{1/(\beta-1)} \ . \tag{3.29}$$

For $\beta = 0$, corresponding to the Itakura-Saito divergence, solving the equations shows that the above multiplicative updates actually still hold. For $\beta = 1$, corresponding to the Kullback-Leibler divergence, the updates take a different form, which coincides with that for the right-sided problem with the dual Kullback-Leibler divergence seen as a Csiszár divergence:

$$x_k = \bar{x}_k \times \exp \left(\frac{\sum_{i=1}^{m} a_{ik} \log (y_i / \sum_{l=1}^{r} a_{il} \bar{x}_l)}{\sum_{i=1}^{m} a_{ik}} \right) \ . \tag{3.30}$$

Even if there is no generic results for right-sided problems with Bregman divergences, specific results can sometimes be obtained on a case-by-case analysis when the divergence is convex-concave in the second argument. This is the case for the convex α-divergences discussed above as a special instance of Csiszár divergences, or for the β-divergences that we discuss now.

Example 3.5. For the right-sided non-negative decomposition problem with the convex-concave β-divergences $d_\beta^{(b)}$, where $\beta \neq 0, 1$, three cases need to be distinguished. First, for $\beta \geq 2$, $d_\beta^{(b)}$ can be decomposed into differentiable convex and concave parts $\breve{d}_\beta^{(b)}(y\|y') = y'^\beta/\beta$, $\widehat{d}_\beta^{(b)}(y\|y') = -yy'^{\beta-1}/(\beta-1)$, up to constant terms with respect to the second argument. Applying the generic method for convex-concave divergences, and remarking that $\partial_{y'}\breve{d}_\beta^{(b)}(y\|y') = y'^{\beta-1}$, $\partial_{y'}\widehat{d}_\beta^{(b)}(y\|y') = -yy'^{\beta-2}$, we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i \big(\sum_{l=1}^r a_{il}\bar{x}_l\big)^{\beta-2}}{\sum_{i=1}^m a_{ik} \big(\sum_{l=1}^r a_{il}\bar{x}_l\big)^{\beta-1}} \right)^{1/(\beta-1)} . \tag{3.31}$$

Second, for $1 \leq \beta \leq 2$, $d_\beta^{(b)}$ can be decomposed into differentiable convex and concave parts $\breve{d}_\beta^{(b)}(y\|y') = y'^\beta/\beta - yy'^{\beta-1}/(\beta-1)$, $\widehat{d}_\beta^{(b)}(y\|y') = 0$, up to constant terms. The derivatives equal $\partial_{y'}\breve{d}_\beta^{(b)}(y\|y') = y'^{\beta-1} - yy'^{\beta-2}$, $\partial_{y'}\widehat{d}_\beta^{(b)}(y\|y') = 0$, and we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i \big(\sum_{l=1}^r a_{il}\bar{x}_l\big)^{\beta-2}}{\sum_{i=1}^m a_{ik} \big(\sum_{l=1}^r a_{il}\bar{x}_l\big)^{\beta-1}} \right) . \tag{3.32}$$

Third, for $\beta \leq 1$, $d_\beta^{(b)}$ can be decomposed into differentiable convex and concave parts $\breve{d}_\beta^{(b)}(y\|y') = -yy'^{\beta-1}/(\beta-1)$, $\widehat{d}_\beta^{(b)}(y\|y') = y'^\beta/\beta$, up to constant terms. The derivatives equal $\partial_{y'}\breve{d}_\beta^{(b)}(y\|y') = -yy'^{\beta-2}$, $\partial_{y'}\widehat{d}_\beta^{(b)}(y\|y') = y'^{\beta-1}$, and we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i \big(\sum_{l=1}^r a_{il}\bar{x}_l\big)^{\beta-2}}{\sum_{i=1}^m a_{ik} \big(\sum_{l=1}^r a_{il}\bar{x}_l\big)^{\beta-1}} \right)^{1/(2-\beta)} . \tag{3.33}$$

For $\beta \in \{0, 1\}$, corresponding to the Itakura-Saito and Kullback-Leibler divergences, the divergences are respectively convex-concave and convex in the second argument, and solving the equations shows that the above updates actually still hold. These updates coincide with that found by Nakano et al. [2010a], Févotte and Idier [2011], where the proof is a specific instance of that for the proposed generic method with use of both Jensen's inequality and the tangent inequality.

Remark 3.20. It is interesting to remark that the three multiplicative updates differ only by an exponent. This exponent step size varies between 0 and 1, and ensures the monotonic decrease of the cost function in the three domains considered. Where the cost function is convex, this exponent step size is relaxed to 1, while it is less than 1 for the two other domains. It is actually possible to show that the exponent can also be relaxed to one for $0 \leq \beta \leq 1$, while keeping monotonic descent of the cost function. This is akin to over-relaxation and produces larger steps for a faster convergence [Févotte 2011].

Similarly, specific results can sometimes be obtained for skew Jeffreys-Bregman divergences when the divergence is convex-concave in the second argument. This is the case for the convex skew Jeffreys α-divergences discussed above as a special instance of Csiszár divergences, and for the convex-concave skew Jeffreys β-divergences. By symmetry, we only develop the results for the right-sided problems.

Example 3.6. Novel analytical schemes can be derived for the right-sided non-negative decomposition problem with the skew Jeffreys (β, λ)-divergences $d^{(b)}_{\beta,\lambda}$, where $\beta \neq 0, 1$, and three cases need again to be distinguished. First, for $\beta \geq 2$, which implies $1 - 2\lambda + \lambda\beta > 0$, $d^{(b)}_{\beta,\lambda}$ can be decomposed into differentiable convex and concave parts $\breve{d}^{(b)}_{\beta,\lambda}(y\|y') = \frac{1-2\lambda+\lambda\beta}{\beta(\beta-1)} y'^{\beta} - \frac{1-\lambda}{\beta-1} y' y^{\beta-1}$, $\widehat{d}^{(b)}_{\beta}(y\|y') = -\frac{\lambda}{\beta-1} y y'^{\beta-1}$, up to constant terms. Applying the generic method, and remarking that $\partial_{y'} \breve{d}^{(b)}_{\beta}(y\|y') = \frac{1-2\lambda+\lambda\beta}{\beta-1} y'^{\beta-1} - \frac{1-\lambda}{\beta-1} y^{\beta-1}$, $\partial_{y'} \widehat{d}^{(b)}_{\beta}(y\|y') = -\lambda y y'^{\beta-2}$, we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} \left(\frac{\lambda(\beta-1)}{1-2\lambda+\lambda\beta} y_i + \frac{1-\lambda}{1-2\lambda+\lambda\beta} \sum_{l=1}^r a_{il}\bar{x}_l \right) (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-2}}{\sum_{i=1}^m a_{ik} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1}} \right)^{1/(\beta-1)} . \tag{3.34}$$

Second, for $1 - 2\lambda + \lambda\beta \leq 0$, which implies $\beta \leq 1$, $\lambda \neq 0$, $d^{(b)}_{\beta,\lambda}$ can be decomposed into differentiable convex and concave parts $\breve{d}^{(b)}_{\beta,\lambda}(y\|y') = -\frac{\lambda}{\beta-1} y y'^{\beta-1}$, $\widehat{d}^{(b)}_{\beta}(y\|y') = \frac{1-2\lambda+\lambda\beta}{\beta(\beta-1)} y'^{\beta} - \frac{1-\lambda}{\beta-1} y' y^{\beta-1}$, up to constant terms. The derivatives equal $\partial_{y'} \breve{d}^{(b)}_{\beta}(y\|y') = -\lambda y y'^{\beta-2}$, $\partial_{y'} \widehat{d}^{(b)}_{\beta}(y\|y') = \frac{1-2\lambda+\lambda\beta}{\beta-1} y'^{\beta-1} - \frac{1-\lambda}{\beta-1} y^{\beta-1}$, and we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-2}}{\sum_{i=1}^m a_{ik} \left(\frac{1-\lambda}{\lambda(1-\beta)} y_i^{\beta-1} + \frac{2\lambda-1-\lambda\beta}{\lambda(1-\beta)} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1} \right)} \right)^{1/(2-\beta)} . \tag{3.35}$$

For $\beta = 0$, $\lambda \in [1/2, 1]$, the decomposition and equations still hold so that the updates are also actually valid. Third, for $\beta \leq 2$, $1 - 2\lambda + \lambda\beta \geq 0$, $d^{(b)}_{\beta,\lambda}$ can be decomposed into differentiable convex and concave parts $\breve{d}^{(b)}_{\beta,\lambda}(y\|y') = \frac{1-2\lambda+\lambda\beta}{\beta(\beta-1)} y'^{\beta} - \frac{\lambda}{\beta-1} y y'^{\beta-1} - \frac{1-\lambda}{\beta-1} y' y^{\beta-1}$, $\widehat{d}^{(b)}_{\beta}(y\|y') = 0$, up to constant terms. The derivatives equal $\partial_{y'} \breve{d}^{(b)}_{\beta}(y\|y') = \frac{1-2\lambda+\lambda\beta}{\beta-1} y'^{\beta-1} - \lambda y y'^{\beta-2} - \frac{1-\lambda}{\beta-1} y^{\beta-1}$, $\partial_{y'} \widehat{d}^{(b)}_{\beta}(y\|y') = 0$, and we obtain the following system of equations:

$$\frac{1-2\lambda+\lambda\beta}{\beta-1} \sum_{i=1}^m a_{ik} \left(\sum_{l=1}^r a_{il}\bar{x}_l \right)^{\beta-1} \left(\frac{x_k}{\bar{x}_k} \right)^{\beta-1} - \lambda \sum_{i=1}^m a_{ik} y_i \left(\sum_{l=1}^r a_{il}\bar{x}_l \right)^{\beta-2} \left(\frac{x_k}{\bar{x}_k} \right)^{\beta-2} = \frac{1-\lambda}{\beta-1} \sum_{i=1}^m a_{ik} y_i^{\beta-1} . \tag{3.36}$$

Unfortunately, this system does not admit a closed-form solution in general when $\beta \neq 2$ and $1 - 2\lambda + \lambda\beta \neq 0$, and iterative methods are required. The same happens for $\beta = 1$, $\lambda \in (0, 1)$, corresponding to the skew Jeffreys divergence, as discussed in the case of skew Jeffreys (α, λ)-divergences, and where the system writes differently in terms of a logarithm and an inverse of the unknown variables. On the contrary, for $\beta = 0$, $\lambda \in [0, 1/2]$, the decomposition still holds so that the above system of equations is also valid. For certain values of β, the equations can be written as polynomial equations, which can be solved more efficiently with specific methods such as root-finding algorithms, or with analytical solutions for lower degrees, when positive solutions do exist. For example, for $\beta = 3/2$ and $\lambda \in (0, 1)$, or $\beta = 0$ and $\lambda \in (0, 1/2)$, we end up with a polynomial equation of degree two with a positive discriminant and a unique analytical positive solution, while for $\beta = 1/2$ and $\lambda \in (0, 2/3)$, or $\beta = -1$ and $\lambda \in (0, 1/3)$, we have a polynomial equation of degree three, and for $\beta = 2/3$ and $\lambda \in (0, 3/4)$, or $\beta = -2$ and $\lambda \in (0, 1/4)$, we have a polynomial equation of degree four. Obviously, as soon as $\lambda \in \{0, 1\}$ in any of the three distinctive cases, which corresponds to a left- or right-sided non-skew problem, the equations or updates lead to the corresponding non-skew multiplicative updates derived before. Finally, for a left-sided problem with the skew Jeffreys (β, λ)-divergences, it actually suffices by symmetry to replace λ with $1 - \lambda$, and to solve the corresponding right-sided problem.

Remark 3.21. It is interesting to interpret the effect of skewing in the two multiplicative updates, compared to that of a right-sided problem with the corresponding non-skew β-divergence, as a weighted convex mixing of the observations y_i with their current estimates $\sum_{l=1}^{r} a_{il}\bar{x}_l$, and as a weighted convex mixing of the current exponentiated estimates $(\sum_{l=1}^{r} a_{il}\bar{x}_l)^{\beta-1}$ with the exponentiated observations $y_i^{\beta-1}$, respectively.

Remark 3.22. We also notice that besides the natural decomposition employed here, we could also have separated the term $\frac{1-2\lambda+\lambda\beta}{\beta(\beta-1)} y'^{\beta}$, into two terms $\frac{1-\lambda}{\beta-1} y'^{\beta}$, $\frac{\lambda}{\beta} y'^{\beta}$. It may seem interesting at first sight since there are still three cases but the domains do not depend on λ anymore. The other side of the coin, however, is that it adds one extra term in the equations, which actually only lead to a general analytical solution for $\beta \geq 2$, and this solution corresponds to the one found above. It confirms the intuitive reasoning that when analyzing a convex or concave function with a multiplicative factor, it is more natural to consider only one term and analyze the sign of the multiplicative factor finely, than to separate different terms to simplify the analysis.

3.3.4. Case of skew Jensen-Bregman divergences

We now focus on the class of skew Jensen-Bregman divergences, which is actually equivalent to the class of skew Burbea-Rao divergences. On the contrary to Bregman divergences, it appears that the skewing procedure makes these divergences always convex-concave in the first argument and in the second argument. As a result, the proposed generic method applies to all these divergences, hence providing novel schemes for non-negative decomposition based on them. Since the class is stable under swapping the arguments, we concentrate without loss of generality on the right-sided non-negative decomposition problem with skew Jensen-Bregman divergences, where the proposed generic method simplifies in terms of the generator function φ.

Corollary 3.7. *Consider the right-sided non-negative decomposition problem with the skew Jensen-Bregman (φ, λ)-divergence $d_{\varphi,\lambda}^{(JB')}$. For all $\bar{\mathbf{x}} \in (\mathbb{R}_+^*)^r$, we have $C(\mathbf{x}) \leq C(\bar{\mathbf{x}})$ for any point $\mathbf{x} \in (\mathbb{R}_+^*)^r$ that verifies the following system of equations:*

$$\sum_{i=1}^{m} a_{ik} \varphi' \left(\sum_{l=1}^{r} a_{il} \bar{x}_l \, \frac{x_k}{\bar{x}_k} \right) = \sum_{i=1}^{m} a_{ik} \varphi' \left(\lambda y_i + (1-\lambda) \sum_{l=1}^{r} a_{il} \bar{x}_l \right) \quad \text{for all } k \in [\![1, r]\!] \ . \tag{3.37}$$

Proof. The skew Jensen-Bregman divergence (φ, λ)-divergence $d_{\varphi,\lambda}^{(JB')}$ is clearly differentiable and convex-concave in the second argument. We can thus decompose it in differentiable convex and concave parts $\breve{d}_{\varphi}^{(JB')}(y \| y') = \lambda \varphi(y) + (1-\lambda)\varphi(y')$, and $\widehat{d}_{\varphi}^{(JB')}(y \| y') = -\varphi(\lambda y + (1-\lambda) y')$. Applying the generic method for convex-concave divergences, the corollary follows by remarking that $\partial_{y'} \breve{d}_{\varphi}^{(JB')}(y \| y') = (1-\lambda)\varphi'(y')$, and $\partial_{y'} \widehat{d}_{\varphi}^{(JB')}(y \| y') = -(1-\lambda)\varphi'(\lambda y + (1-\lambda) y')$. □

Remark 3.23. It is interesting to interpret the effect of skewing in the equations, compared to that of a left-sided problem with the corresponding non-skew Bregman divergence, as a weighted convex mixing of the observations y_i with their current estimates $\sum_{l=1}^{r} a_{il} \bar{x}_l$.

Remark 3.24. We notice that besides the natural decomposition, we could have separated the convex term $(1-\lambda)\varphi(y')$, into a convex term $\varphi(y')$, and a concave term $-\lambda\varphi(y')$. In this case, the system of equations becomes:

$$\begin{aligned} \sum_{i=1}^{m} a_{ik} \varphi' \left(\sum_{l=1}^{r} a_{il} \bar{x}_l \, \frac{x_k}{\bar{x}_k} \right) = \lambda \sum_{i=1}^{m} a_{ik} \varphi' \left(\sum_{l=1}^{r} a_{il} \bar{x}_l \right) \\ + (1-\lambda) \sum_{i=1}^{m} a_{ik} \varphi' \left(\lambda y_i + (1-\lambda) \sum_{l=1}^{r} a_{il} \bar{x}_l \right) \ . \end{aligned} \tag{3.38}$$

It modifies the system of equations for the natural decomposition by a weighted convex mixing of the constant $\sum_{i=1}^{m} a_{ik} \varphi'(\lambda y_i + (1-\lambda) \sum_{l=1}^{r} a_{il} \bar{x}_l)$, which depends on both the observations y_i and their current estimates $\sum_{l=1}^{r} a_{il} \bar{x}_l$, with the constant $\sum_{i=1}^{m} a_{ik} \varphi'(\sum_{l=1}^{r} a_{il} \bar{x}_l)$, which only depends on the current estimates. It thus seems intuitive that as λ augments, the solutions are bound towards the current estimates since the value $x_k = \bar{x}_k$ gets closer to a solution, so that the updates get slowed down. This is confirmed formally below for the skew Jensen (β, λ)-divergences.

Example 3.7. A novel analytical scheme can be derived for the right-sided problem with the skew Jensen β-divergence $d_{\beta,\lambda}^{(b')}$, where $\beta \neq 0, 1$, for which the system of equations leads to the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^{m} a_{ik} (\lambda y_i + (1-\lambda) \sum_{l=1}^{r} a_{il} \bar{x}_l)^{\beta - 1}}{\sum_{i=1}^{m} a_{ik} (\sum_{l=1}^{r} a_{il} \bar{x}_l)^{\beta - 1}} \right)^{1/(\beta - 1)} \ . \tag{3.39}$$

For $\beta = 0$, corresponding to a skew version of the Itakura-Saito divergence, solving the equations shows that the above multiplicative updates actually still hold. For $\beta = 1$, corresponding to a skew version of the Kullback-Leibler divergence, the updates take a different form as follows:

$$x_k = \bar{x}_k \times \exp\left(\frac{\sum_{i=1}^m a_{ik} \log\left(\lambda y_i / \sum_{l=1}^r a_{il}\bar{x}_l + 1 - \lambda\right)}{\sum_{i=1}^m a_{ik}}\right) . \tag{3.40}$$

In particular for $\lambda = 1/2$, we have closed-form multiplicative updates for all symmetric Jensen β-divergences, including the well-known Jensen-Shannon divergence for $\beta = 1$ as a symmetric version of the Kullback-Leibler divergence, as well as the cosh distance for $\beta = 0$ as a symmetric version of the Itakura-Saito divergence. For a left-sided problem with the skew Jensen β-divergence $d_{\beta,\lambda}^{(b')}$, it actually suffices to replace λ with $1 - \lambda$, and to solve the corresponding right-sided problem.

Remark 3.25. Considering the alternative decomposition discussed in the above remark, the system of equations for the skew Jensen (β, λ)-divergences leads to closed-form multiplicative updates too. For $\beta \neq 1$, these updates write as follows:

$$x_k = \bar{x}_k \times \left(\lambda + (1-\lambda)\frac{\sum_{i=1}^m a_{ik}\left(\lambda y_i + (1-\lambda)\sum_{l=1}^r a_{il}\bar{x}_l\right)^{\beta-1}}{\sum_{i=1}^m a_{ik}\left(\sum_{l=1}^r a_{il}\bar{x}_l\right)^{\beta-1}}\right)^{1/(\beta-1)} . \tag{3.41}$$

For $\beta = 1$, corresponding to the skew Jensen-Shannon divergence, these updates write as follows:

$$x_k = \bar{x}_k \times \exp\left((1-\lambda)\frac{\sum_{i=1}^m a_{ik} \log\left(\lambda y_i / \sum_{l=1}^r a_{il}\bar{x}_l + 1 - \lambda\right)}{\sum_{i=1}^m a_{ik}}\right) . \tag{3.42}$$

It corroborates the remark that this decomposition intuitively reduces the progression of the updates as λ augments compared to the progression for the natural decomposition. It also confirms the reasoning that when analyzing a convex or concave function with a multiplicative factor, it is more natural to consider only one term and analyze the sign of the multiplicative factor finely, as discussed previously.

Unfortunately, we were not able to derive similar analytical results for the skew Jensen (α, λ)-divergences, as discussed previously when seen as Csiszár divergences.

3.3.5. Case of skew (α, β, λ)-divergences

We now consider the parametric family of skew (α, β, λ)-divergences. These divergences are almost always convex-concave in the first and in the second argument so that the proposed generic method applies. Nevertheless, depending on the parameter values, the convex-concave decompositions do differ, and several cases need to be considered. For simplicity, we begin with the non-skew case, that is, corresponding to $\lambda \in \{0, 1\}$, of non-negative decomposition with the (α, β)-divergences. Because these divergences are stable under swapping the arguments, we restrict without loss of generality to the right-sided problem.

Example 3.8. For the right-sided non-negative decomposition problem with the convex-concave (α, β)-divergences $d^{(ab)}_{\alpha,\beta}$, where $\alpha\beta(\alpha+\beta) \neq 0$, three cases need to be distinguished. First, for $\beta/\alpha \geq 1/\alpha$, $d^{(ab)}_{\alpha,\beta}$ can be decomposed into differentiable convex and concave parts $\breve{d}^{(ab)}_{\alpha,\beta}(y\|y') = y'^{\alpha+\beta}/\alpha(\alpha+\beta)$, $\widehat{d}^{(ab)}_{\alpha,\beta}(y\|y') = -y^{\alpha}y'^{\beta}/\alpha\beta$, up to constant terms. Applying the generic method for convex-concave divergences, and remarking that $\partial_{y'}\breve{d}^{(ab)}_{\alpha,\beta}(y\|y') = y'^{\alpha+\beta-1}/\alpha$, $\partial_{y'}\widehat{d}^{(ab)}_{\alpha,\beta}(y\|y') = -y^{\alpha}y'^{\beta-1}/\alpha$, we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^{\alpha} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1}}{\sum_{i=1}^m a_{ik} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha+\beta-1}} \right)^{1/(\alpha+\beta-1)} . \tag{3.43}$$

Second, for $1/\alpha - 1 \leq \beta/\alpha \leq 1/\alpha$, $d^{(ab)}_{\alpha,\beta}$ can be decomposed into differentiable convex and concave parts $\breve{d}^{(ab)}_{\alpha,\beta}(y\|y') = y'^{\alpha+\beta}/\alpha(\alpha+\beta) - y^{\alpha}y'^{\beta}/\alpha\beta$, $\widehat{d}^{(ab)}_{\alpha,\beta}(y\|y') = 0$, up to constant terms. The derivatives equal $\partial_{y'}\breve{d}^{(ab)}_{\alpha,\beta}(y\|y') = y'^{\alpha+\beta-1}/\alpha - y^{\alpha}y'^{\beta-1}/\alpha$, $\partial_{y'}\widehat{d}^{(ab)}_{\alpha,\beta}(y\|y') = 0$, and we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^{\alpha} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1}}{\sum_{i=1}^m a_{ik} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha+\beta-1}} \right)^{1/\alpha} . \tag{3.44}$$

Third, for $\beta/\alpha \leq 1/\alpha - 1$, $d^{(ab)}_{\alpha,\beta}$ can be decomposed into differentiable convex and concave parts $\breve{d}^{(ab)}_{\alpha,\beta}(y\|y') = -y^{\alpha}y'^{\beta}/\alpha\beta$ and $\widehat{d}^{(ab)}_{\alpha,\beta}(y\|y') = y'^{\alpha+\beta}/\alpha(\alpha+\beta)$, up to constant terms. The derivatives equal $\partial_{y'}\breve{d}^{(ab)}_{\alpha,\beta}(y\|y') = -y^{\alpha}y'^{\beta-1}/\alpha$, $\partial_{y'}\widehat{d}^{(ab)}_{\alpha,\beta}(y\|y') = y'^{\alpha+\beta-1}/\alpha$, and we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^{\alpha} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1}}{\sum_{i=1}^m a_{ik} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha+\beta-1}} \right)^{1/(1-\beta)} . \tag{3.45}$$

For the limit cases $\beta = 0$, or $\alpha + \beta = 0$, solving the equations shows that the above updates actually still hold as soon as $\alpha \neq 0$. For the limit case $\alpha = 0$, the updates are however not valid anymore. In the special case $\alpha = 0$, $\beta = 1$, corresponding to the dual Kullback-Leibler divergence, the equations lead to the already known multiplicative updates as follows:

$$x_k = \bar{x}_k \times \exp\left(\frac{\sum_{i=1}^m a_{ik} \log\left(y_i / \sum_{l=1}^r a_{il}\bar{x}_l\right)}{\sum_{i=1}^m a_{ik}} \right) . \tag{3.46}$$

The four obtained multiplicative updates coincide with that of Cichocki et al. [2011], where the proof is a special instance of ours with use of both Jensen's inequality and the tangent inequality.[3] Nevertheless, for $\alpha = 0$, $\beta \neq 1$, the above reasoning does not hold anymore as discussed in the remark below. We notice finally that for $\alpha + \beta = 1$, the obtained updates coincide with that of the α-divergences $d^{(a)}_{\alpha}$, while

[3] A slight difference appears in their proof, where the tangent inequality is first applied to the concave term, and the resulting affine term is added to the convex term before applying Jensen's inequality, but the results are identical since this extra term has no effect in Jensen's inequality.

for $\alpha = 1$, they coincide with that of the β-divergences $d^{(b)}_{\beta+1}$. Moreover, a left-sided problem can be considered straightforward by swapping α and β, and solving the corresponding right-sided problem, provided that $\beta \neq 0$, or that $\beta = 0$ and $\alpha = 1$. The updates then also coincide with that of α-divergences and of β-divergences.

Remark 3.26. For $\alpha = 0$, $\beta \neq 1$, the divergences exhibit terms whose convexity or concavity depends not only on the values of the parameters, but also on the value of the first argument, or even on different zones of the second argument. Therefore, the convex-concave decomposition depends not only on the parameter values, but also on the observations y_i and on the region considered for searching a solution. This situation is not considered properly in [Cichocki et al. 2011] where there is a technical flaw. It is argued that we can take the limit in the multiplicative updates to obtain the updates for $\alpha = 0$, $\beta \neq 1$. Taking this limit, the authors obtain trivial identity updates corresponding to a null exponent step size, that is, the updates actually do not modify the current solution. This of course does not make the cost function increase, but a more rigorous reasoning would be required to provide non-trivial updates. Some perspectives on this line are discussed later.

Remark 3.27. It is interesting to remark that the first three multiplicative updates differ only by their exponent step size which ensures the monotonic decrease of the cost function in the three domains considered. The four multiplicative updates can actually be unified, up to exponents, by using deformed exponentials and logarithms as done in [Cichocki et al. 2011]. This permits to extend the remark to the fourth update. Moreover, this normalizes the exponents between 0 and 1, where the relaxed exponent 1 is attained in the convex cases $\alpha \neq 0$ and $1/\alpha - 1 \leq \beta/\alpha \leq 1/\alpha$, or $\alpha = 0$ and $\beta = 1$. Last but not least, it is shown in [Cichocki et al. 2011] that the other exponents can be over-relaxed to 1 under certain conditions, roughly speaking when the estimates are close enough to the observations so that convexity holds locally.

We now turn to the more complex situation of a general non-negative decomposition problem with arbitrary skew (α, β, λ)-divergences. Because these divergences are also stable under swapping the arguments, we focus on the right-sided problem.

Example 3.9. Novel analytical schemes can be derived for the right-sided non-negative decomposition problem with the skew (α, β, λ)-divergences $d^{(ab)}_{\alpha,\beta,\lambda}$, where $\alpha\beta(\alpha + \beta) \neq 0$, for which seven cases need to be distinguished. There are actually three terms in the decomposition of $d^{(ab)}_{\alpha,\beta,\lambda}$ into convex and concave parts, up to constant terms. These three terms are $\frac{\lambda\beta+(1-\lambda)\alpha}{\alpha\beta(\alpha+\beta)} y'^{\alpha+\beta}$, $-\frac{\lambda}{\alpha\beta} y^{\alpha} y'^{\beta}$, $-\frac{1-\lambda}{\alpha\beta} y^{\beta} y'^{\alpha}$, with respective derivatives $\frac{\lambda\beta+(1-\lambda)\alpha}{\alpha\beta} y'^{\alpha+\beta-1}$, $-\frac{\lambda}{\alpha} y^{\alpha} y'^{\beta-1}$, $-\frac{1-\lambda}{\beta} y^{\beta} y'^{\alpha-1}$. For different values of the parameters, these terms can all become convex or concave, almost independently but just never all concave at the same time, hence the seven possible combinations. In the general case, we thus have equations with potentially monomials of degree $\alpha + \beta - 1$ if the first term is convex, $\beta - 1$ if the second term is convex, $\alpha - 1$ if the third term is convex, and 0 if at least one term is concave. As a result, we have in general analytical updates when only one term is convex, otherwise we end up a priori with non-trivial equations to solve. Special cases where this equation can be solved are discussed later. We now concentrate on the three principal cases where each term is convex in turn while the two others are concave. When the first term only is convex, which is equivalent to $\alpha, \beta \geq 1$ or $\alpha, \beta < 0$, we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} \left(\frac{\lambda\beta}{\lambda\beta + (1-\lambda)\alpha} y_i^\alpha (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1} + \frac{(1-\lambda)\alpha}{\lambda\beta + (1-\lambda)\alpha} y_i^\beta (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha-1} \right)}{\sum_{i=1}^m a_{ik} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha+\beta-1}} \right)^{1/(\alpha+\beta-1)} . \tag{3.47}$$

When the second term only is convex, which is equivalent to $0 < \alpha \leq 1$, $\beta < 0$, $\alpha/(\alpha - \beta) \leq \lambda \leq 1$, we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^\alpha (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1}}{\sum_{i=1}^m a_{ik} \left(\frac{(\lambda-1)\alpha}{\lambda\beta} y_i^\beta (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha-1} + \frac{\lambda\beta+(1-\lambda)\alpha}{\lambda\beta} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha+\beta-1} \right)} \right)^{1/(1-\beta)} . \tag{3.48}$$

When the third term only is convex, which is equivalent to $0 < \beta \leq 1$, $\alpha < 0$, $0 \leq \lambda \leq \alpha/(\alpha - \beta)$, we obtain the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^\beta (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha-1}}{\sum_{i=1}^m a_{ik} \left(\frac{\lambda\beta}{(\lambda-1)\alpha} y_i^\alpha (\sum_{l=1}^r a_{il}\bar{x}_l)^{\beta-1} + \frac{\lambda\beta+(1-\lambda)\alpha}{(1-\lambda)\alpha} (\sum_{l=1}^r a_{il}\bar{x}_l)^{\alpha+\beta-1} \right)} \right)^{1/(1-\alpha)} . \tag{3.49}$$

In the four other cases, there is no such solution, even if we may have polynomial equations for certain values of the parameters as discussed in the previous examples. Now in the limit case $\alpha + \beta = 0$, $\alpha, \beta \neq 0$, there are three non-constant terms in the decomposition too, whose derivatives coincide with the above ones, so that the respective systems of equations for the different possible combinations of convex and concave terms are still valid. It appears that the first term is never convex alone, so that there are two cases where we have general analytical updates, that is, when the second term only is convex, and when the third term only is convex. In these two cases, the updates coincide respectively with the two last multiplicative updates above. For the other limit cases, where at least $\alpha = 0$, or $\beta = 0$, the decomposition again depends on the values of the two arguments in general. As a result, the above reasoning does not hold anymore, except for the case $\lambda = 1$, $\alpha = 0$, $\beta = 1$, or the symmetric case $\lambda = 0$, $\alpha = 1$, $\beta = 0$, which correspond to right- and left-sided non-skew problems with the dual Kullback-Leibler and the Kullback-Leibler divergences, respectively. More generally, when $\lambda \in \{0, 1\}$, we obviously have the non-skew left- and right-sided problems with the (α, β)-divergences, and the same multiplicative updates as discussed above when they exist. For $\alpha = 1$, or $\beta = 1$, we end up with the problems for the skew Jeffreys β-divergences, while for $\alpha + \beta = 1$, we end up with the problems for the skew Jeffreys α-divergences, with the corresponding closed-form multiplicative updates when they exist. When $\alpha = \beta$, the problem is actually symmetric and is equivalent to the non-skew problem for any $\lambda \in [0,1]$, for which the multiplicative updates derived still hold as soon as $\alpha, \beta \neq 0$. Finally, a left-sided problem with the skew (α, β, λ)-divergences can be considered straightforward either by replacing λ with $1-\lambda$, or by swapping α and β, to end up with an equivalent right-sided problem.

Remark 3.28. It is again interesting to interpret the effect of skewing in the multiplicative updates, compared to that of a right-sided problem with the corresponding non-skew divergence, as a weighted convex mixing of the exponentiated observations with their current exponentiated estimates.

Remark 3.29. We could also have considered an alternative decomposition by separating the term $\frac{\lambda\beta+(1-\lambda)\alpha}{\alpha\beta(\alpha+\beta)} y'^{\alpha+\beta}$, into two terms $\frac{\lambda}{\alpha(\alpha+\beta)} y'^{\alpha+\beta}$, $\frac{1-\lambda}{\beta(\alpha+\beta)} y'^{\alpha+\beta}$, whose convexity or concavity does not depend on λ anymore. Nevertheless, this adds one extra term in the equations, which actually only lead to a general analytical solution for $\alpha, \beta \geq 1$, or $\alpha, \beta < 0$, and this solution corresponds to the one found above. This confirms again the intuitive reasoning that when analyzing a convex or concave function with a multiplicative factor, it is more natural to consider only one term and analyze the sign of the multiplicative factor finely, than to separate different terms to simplify the analysis.

Example 3.10. A novel analytical scheme can also be derived for a specific two-parameter family of skew (α, β, λ)-divergences. Considering the right-sided problem with the skew (α, β, λ)-divergences $d^{(ab)}_{\alpha,\beta,\lambda}$, where $(\alpha, \beta) \in (\mathbb{R}^*_- \times \mathbb{R}^*_+) \cup (\mathbb{R}^*_+ \times \mathbb{R}^*_-)$, $\lambda = \alpha/(\alpha - \beta) \in (0,1)$, the monomial of degree $\alpha+\beta-1$ in the system of equations vanishes. This provides closed-form solutions for the different possible combinations of convex and concave terms. There are actually three such combinations to distinguish. When the two terms are convex, which is equivalent to $\alpha < 0$ and $\beta \geq 1$, or $\beta < 0$ and $\alpha \geq 1$, the system of equations leads to the following closed-form multiplicative updates:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^\beta (\sum_{l=1}^r a_{il} \bar{x}_l)^{\alpha-1}}{\sum_{i=1}^m a_{ik} y_i^\alpha (\sum_{l=1}^r a_{il} \bar{x}_l)^{\beta-1}} \right)^{1/(\beta-\alpha)} . \tag{3.50}$$

As a special case when $\alpha + \beta = 1$, $\alpha \in \mathbb{R} \setminus [0,1]$, we obtain the exact same one-parameter family and multiplicative updates as discussed previously for skew Jeffreys (α, λ)-divergences. Now, when one term is convex while the other is concave, we end up with particular cases of two cases discussed above. The first one is when $0 < \alpha \leq 1$, $\beta < 0$, where the multiplicative updates specialize as follows:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^\alpha (\sum_{l=1}^r a_{il} \bar{x}_l)^{\beta-1}}{\sum_{i=1}^m a_{ik} y_i^\beta (\sum_{l=1}^r a_{il} \bar{x}_l)^{\alpha-1}} \right)^{1/(1-\beta)} . \tag{3.51}$$

The second one is when $0 < \beta \leq 1$, $\alpha < 0$, where the multiplicative updates specialize as follows:

$$x_k = \bar{x}_k \times \left(\frac{\sum_{i=1}^m a_{ik} y_i^\beta (\sum_{l=1}^r a_{il} \bar{x}_l)^{\alpha-1}}{\sum_{i=1}^m a_{ik} y_i^\alpha (\sum_{l=1}^r a_{il} \bar{x}_l)^{\beta-1}} \right)^{1/(1-\alpha)} . \tag{3.52}$$

In the limit case where either $\alpha = 0$, or $\beta = 0$, we respectively have $\lambda = 0$, or $\lambda = 1$, and we end up with non-skew left- and right-sided problems discussed above. Therefore, the reasoning does not hold anymore, except for $\beta = 1$, or for $\alpha = 1$, where we obtain respectively the equivalent left- and right-sided problems with the dual Kullback-Leibler and the Kullback-Leibler divergences, for which the above multiplicative updates still hold and specialize to the ones already known. We also notice that there is an inherent symmetry within the two-parameter family, where swapping α and β transforms λ in $1 - \lambda$. As a result, we need only consider the values of $(\alpha, \beta) \in \mathbb{R}_-^* \times \mathbb{R}_+^*$ because the other values are redundant. We can finally consider straightforwardly the left-sided problems with a two-parameter family of skew (α, β, λ)-divergences, where the constraint now writes $\lambda = \beta/(\beta - \alpha)$, and we have equivalence with the right-sided problems by swapping α and β, and replacing λ with $1 - \lambda$.

3.4. Discussion

In this chapter, we proposed methods for non-negative matrix factorization with convex-concave divergences. The proposed framework encompasses many common information divergences, such as Csiszár divergences, certain Bregman divergences, and in particular all α-divergences and β-divergences. We developed a general optimization scheme based on variational bounding with auxiliary functions that works for almost arbitrary convex-concave divergences. We obtained monotonically decreasing updates under mild conditions by minimizing the auxiliary function. We also considered symmetrized and skew divergences for the cost function. In particular, we specialized the generic updates to provide updates for Csiszár divergences, certain skew Jeffreys-Bregman divergences, skew Jensen-Bregman divergences, thus leading to several known multiplicative updates, as well as novel multiplicative updates, for α-divergences, β-divergences, and their symmetrized or skew versions. We also generalized this by considering the family of skew (α, β, λ)-divergences. Several directions of improvement were however left out for future work.

To begin with, we would like to enhance the standard factorization model considered here. Direct extensions can easily be handled in the proposed framework, such as generalizations to convex NMF models and non-negative tensor models through vectorization, as discussed for β-divergences in [Févotte and Idier 2011]. Other generalizations could also be investigated, for example, convolutive NMF models as proposed in [Smaragdis 2004].

In addition to extending the models, we could also extend the cost functions. Although the framework presented for convex-concave divergences unifies and generalizes the majority of information divergences employed in the literature, we did not discuss the possibility to add penalty terms to regularize the solutions. The proposed framework extends straightforward to convex-concave penalties by including the respective terms in the decomposition of the cost function into convex and concave parts for constructing the auxiliary functions. It notably includes penalizations with ℓ_p-norms for sparsity regularization. More specific penalties could also be considered on a case-by-case analysis, such as the group sparsity employed in

[Lefèvre et al. 2011a]. Nevertheless, even with the simple ℓ_1-norm penalty, we may not be able to systematically derive closed-form multiplicative updates extending those discussed here. It has already been observed in [Févotte and Idier 2011] for the specific β-divergences. This is because the introduced term in the equations can make them non-trivial to solve, as soon as there are two monomials of different degrees and a non-null constant term, or more than two monomials. For the same reasons, we were not able to derive systematic multiplicative updates on the whole parameter range for standard skewing of the α-divergences and β-divergences, or more generally of the (α, β)-divergences.

To handle such situations, it would be interesting to investigate tailored optimization schemes to solve the equations. In the present work, we focused on simplifications of the equations into attractive multiplicative updates. As discussed previously, the generic updates can nonetheless be solved in the general case by employing optimization schemes such as line search or Newton's methods in dimension one. More specific schemes can beneficially be derived to tailor optimization according to the cost function and penalties considered when no multiplicative updates are available.

A perspective concerning the standard skewing of α-divergences, β-divergences, and (α, β)-divergences, is to find links between the solutions of the skew problem, and the respective solutions of the left- and right-sided problems. We have found in this context that it is possible to express the update of the current solution of the skew problem, as an equation involving the left- and right-sided updates of this current solution. Nevertheless, it does not iterate since the obtained solution does not correspond in general to that of the left- and right-sided updates. As a result, the equation for the skew updates should be solved at each iteration. It would be more interesting to solve in parallel the left- and right-sided problem to derive the skew solution in the end only. We were not yet able to derive such relations.

To go further, other generic updates could also be investigated. As discussed previously, we focused here on the minimization of the auxiliary function. Nevertheless, any point that makes the auxiliary function decrease actually also makes the cost function decrease. This can be used to propose updates based on majorization-equalization, or any compromise in between minimization and equalization. On the one hand, the equalization may permit to improve the speed of convergence of the cost function by behaving as an over-relaxation compared to minimization. On the other hand, the equalization is less likely than the minimization to have a solution inside the positive orthant, so that tempered updates may be necessary. This has notably been developed for β-divergences in [Févotte and Idier 2011], and is worth extending to arbitrary convex-concave divergences. Another approach to provide alternative generic updates is to construct other auxiliary functions. In particular, we may require further assumptions on the convex-concave divergences. For example, it is well-known that any function with bounded Hessian is actually convex-concave. Using such functions, we may derive tighter bounds for the auxiliary functions by considering second-order approximations instead of the simple first-order approximation in the tangent inequality. Other specific properties may be employed such as strong convexity or logarithmic convexity.

On another perspective, we could relax slightly the convex-concave assumptions made here. Indeed, the proposed methodology for constructing the auxiliary functions can be extended as is to consider convex-concave decompositions that depend not only on the value of the first argument, but also on different values of the second argument. As a result, we may generalize the discussion to include divergences that are not strictly speaking convex-concave, as for certain (α, β)-divergences and their skew versions. Similar ideas have been considered in [Cichocki et al. 2011] to provide relaxed multiplicative updates, corresponding to a gradient descent, with guaranteed monotonic decrease in certain regions of the solution space that depend on the observations. A systematic analysis of such updates is worth exploring.

Last but not least, a major theoretical step for non-negative matrix factorization would be to prove strong convergence properties of the algorithms. In the present work, we have obtained the guaranteed monotonic decrease of the cost function, hence ensuring its convergence. Nevertheless, we were not able to prove convergence of the cost function to a global or local minimum, nor to a stationary point. Moreover, even if the cost function converges, the updates themselves may not converge, though we are guaranteed by the monotonic decrease that the output solutions have at least a certain quality. A first step in obtaining further results in this direction is the study of the supervised non-negative decomposition solely. For example, the obtained updates for non-negative decomposition with the Euclidean cost function actually converge to the global minimum, as a special case of a more general framework for quadratic programming with non-negative constraints [Sha et al. 2007]. With this respect, we believe that the recent convergence proof for the concave-convex procedure provided in [Sriperumbudur and Lanckriet 2012] could be adapted to the proposed framework. Interestingly, this proof relies on a more general theory for studying convergence properties of iterative algorithms [Zangwill 1969], which was also used in [Sha et al. 2007]. Concerning the convergence properties of the general NMF problem with alternate updates, results have again been proved for the Euclidean cost function in [Lin 2007], though the case of general divergences is more complicated. Interesting insights in this direction have been recently investigated by studying the Lyapunov stability of NMF with α-divergences [Yang and Ye 2012], and β-divergences [Badeau et al. 2010]. Further considerations are needed on this line to prove full convergence of the schemes for general convex-concave divergences.

Part II.

Real-Time Applications in Audio Signal Processing

4. Real-Time Audio Segmentation

In this chapter, we address the problem of automatic segmentation which is fundamental in audio signal processing. A major drawback of existing approaches in this context is that they consider specific signals and homogeneity criteria, or assume normality of the data distribution. Other issues arise from the potential computational complexity and non-causality of the schemes. To address these issues, we devise a generic and unifying framework for real-time audio segmentation based on the methods for sequential change detection with exponential families developed in Chapter 2. The proposed system can handle various types of signals and of homogeneity criteria, by controlling the information rate of the audio stream to detect changes in real time. The framework also bridges the gap between statistical and distance-based approaches to segmentation through the dually flat geometry of exponential families. We notably clarify the relations between various standard approaches to audio segmentation, and show how they can be unified and generalized in the proposed framework. Various applications are showcased to illustrate the generality of the framework, and a quantitative evaluation is performed for musical onset detection to demonstrate how the proposed approach can leverage modeling in complex problems.

4.1. Context

In this section, we first provide some background information on the problem of audio segmentation, and in particular for onset detection in music signals and for speaker segmentation in speech signals. We then discuss the motivations of our approach to the problem of real-time audio segmentation. We finally sum up our main contributions in this context.

4.1.1. Background

As depicted in Figure 4.1, the task of *audio segmentation* consists in determining time instants which partition a sound signal into homogeneous and continuous temporal regions, such that adjacent regions exhibit inhomogeneities. These time instants are called *boundaries*, while the continuous temporal regions between the boundaries are simply called *segments*. Audio segmentation has been widely studied in the literature, mainly for music and speech signals, and is of great interest for a variety of applications in audio analysis, indexing and information retrieval, as put in perspective in the popular papers of Tzanetakis and Cook [1999], and Foote [2000]. In this context, the following design element is of primary importance.

For the segmentation to be relevant, the segments must possess a certain consistency of their own information content, but a difference of information content

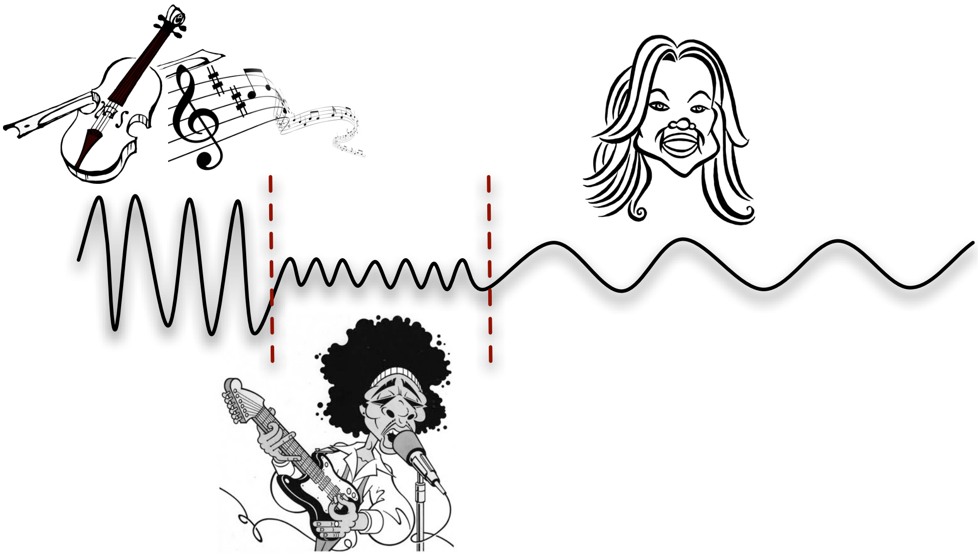

Figure 4.1.: Schematic view of the audio segmentation task. Starting from the audio signal, the goal is to find time boundaries such that the resulting segments are intrinsically homogeneous but differ from their neighbors.

with the previous and next segments. This therefore requires the definition of a criterion to quantify the homogeneity, or consistency, and various criteria may be employed depending on the types of signals considered. For instance, we may want to segment a conversation in terms of silence and speech, or in terms of different speakers. Similarly we may want to segment a music piece in terms of notes, or in terms of different instruments.

Early researches for the automatic segmentation of digital signals can be traced back to the pioneering work of Basseville and Benveniste [1983a,b] on the detection of changes according to different criteria, such as spectral characteristics, in various applicative domains. This framework was later applied by André-Obrecht [1988] to the segmentation of speech signals into homogeneous infra-phonemic regions. The problem of audio segmentation is still actively researched today, either for direct applications such as speaker segmentation in conversations and onset detection in music signals as discussed later, or as a front-end module in a broad class of tasks such as speaker diarization [Tranter and Reynolds 2006, Anguera Miro et al. 2012] and music structure analysis [Foote 1999, Paulus et al. 2010] among others.

In many works, audio segmentation relies on application-specific and high-level criteria of homogeneity in terms of semantic classes, and the supervised detection of changes is based on a system for automatic classification where the segments are created in function of the assigned classes. For example, the segmentation of a conversation into speakers would depend on a system for speaker recognition. Similarly, the segmentation of a music piece into notes would depend on a system for note recognition. Such an approach has yet the drawbacks to assume the existence and knowledge of classes, to rely on a potentially fallible classification, and to require some training data for learning the classes.

Some approaches without classification have been proposed to address these issues and perform an unsupervised, or blind, detection of changes. This is notably the case of onset detection in music signals [Bello et al. 2005, Dixon 2006]. In this context, a detection function is constructed from the signal by extracting certain sound features. The detection function is then used to localize the onsets by applying thresholding and peak-picking heuristics. For example, we can employ directly the energy envelope as a detection function, which is well-suited for percussions and instruments with a marked attack such as the piano. This basic idea was enhanced by Klapuri [1999] who considered the energy in non-overlapping frequency bands by applying a filter bank motivated by psychoacoustic considerations, with fusion of the onsets detected in the respective bands.

Complementary information is often required to deal both with percussive sounds, and with smooth sounds such as bowed instruments. This information is notably contained in the evolution of the spectral characteristics. For example, Duxbury et al. [2002] proposed to apply a filter bank, and to compute the energy in high-frequency bands coupled with spectral information in lower bands. Bello and Sandler [2003] considered the phase information of the spectrum to detect onsets based on its deviation. The phase information was further combined with the amplitude information by Duxbury et al. [2003], and with the energy by Bello et al. [2004].

In particular, variations of the so-called *spectral flux* method have been widely used. This method consists in computing the difference between the frequency spectra at successive time frames to define the detection function. The maxima of the detection function then correspond to locations where the spectrum greatly changes. The principal variants of this method differ both in the processing of the spectra, for example, choosing a certain frequency scale and an appropriate time-frequency transform, normalizing the spectra, emphasizing certain frequencies, and in the choice of the distance function used to measure the difference between successive spectra, for example, the simple Euclidean distance, the taxicab distance as a direct unsigned difference, the half-wave rectified difference, the Kullback-Leibler divergence as a distance between histograms.

Other unsupervised approaches have also been considered for the problem of speaker segmentation [Kemp et al. 2000, Kotti et al. 2008]. In this context, the audio frames are in general represented with timbre features that encode the spectral envelope, such as Mel-frequency cepstral coefficients (MFCC). Unsupervised methods for segmentation then usually consists in computing a metric to quantify the timbral changes along time represented by the evolution of the MFCC distribution, supposed to be characteristic of the speaker changes.

In particular, variations of the *cepstral flux* method have been widely used. This method relies on computing a distance between the MFCC observations at successive time frames, where the variants differ both in the processing of the MFCC observations, for example, the number of coefficients, their normalization, the number of bands, and in the distance function employed, for example, the simple Euclidean distance, or more elaborate distances between statistical models representing the MFCC distribution. When using statistical models, the MFCC observations are in general supposed to be Gaussian variables, and various metrics are employed, such as the Kullback-Leibler divergence on maximum likelihood estimates, considered first by Siegler et al. [1997], or statistics based on likelihoods, including the standard Akaike information criterion (AIC), or the Bayesian information criterion (BIC) proposed by Chen and Gopalakrishnan [1998], and the generalized likelihood ratio (GLR) proposed by Bonastre et al. [2000].

The BIC method became quite popular and is now the baseline method despite the relatively high computational cost of its naive implementation. In order to speed up the BIC method, several directions have been investigated. For instance, Tritschler and Gopinath [1999] introduced windowing heuristics by computing only the BIC at the center of a growing or sliding window, with growing and sliding factors of several time frames, hence making the scheme faster but less precise since not every time frame is tested as a potential boundary. Cettolo and Vescovi [2003] compared several tailored implementations by updating the BIC and related statistics incrementally.

An alternative to computational heuristics or optimizations is to consider simpler test statistics than the BIC, such as the cumulative sum (CUSUM) statistics based on likelihood ratio (LR) statistics, where approximations in the parameter estimation before change are realized to make the computation recursive and less expensive. It was notably used and compared to other methods for speaker segmentation by Omar et al. [2005]. Finally, several authors have considered two-pass schemes, where the first pass is done with a fast method to determine rough candidate boundaries, while the second pass is done with the BIC method on windows centered around the candidate boundaries for pruning and refinement, as in the well-known system proposed by Delacourt and Wellekens [2000].

4.1.2. Motivations

We are interested here in providing a unifying framework for audio segmentation, with arbitrary types of signals and of homogeneity criteria. Moreover, we do not assume any a priori on the existence of classes. Segmentation is thus distinct from classification for us, the classes being replaced with consistent informative entities. This distinction is coherent from the perceptual perspective of auditory scene analysis [Bregman 1990]. Indeed, human temporal segmentation is a more primitive process than sound recognition and identification. Since the characteristics of natural sound sources tend to vary smoothly along time, new sound events are in general indicated by abrupt changes. Therefore, the segmentation process does not require the interpretation of the incoming acoustic data, but rather follows the variation of its information content so as to detect structural ruptures. We think that taking such considerations into account would help the design of a generic framework to perform audio segmentation.

Most unsupervised approaches to audio segmentation are however tailored to a particular type of signal and of homogeneity criterion. Nonetheless, different segmentation tasks are sometimes addressed with similar approaches. In particular, this is the case of distance-based approaches such as spectral and cepstral flux methods in musical onset detection and speaker segmentation. Other interesting approaches employ information-theoretic distances, such as the parametric family of Rényi entropies [Liuni et al. 2011], or the Kullback-Leibler divergence [Cont et al. 2011], to perform a segmentation on the spectral distribution as a sound feature. More general audio segmentation frameworks employ kernel methods in relation to support vector machine (SVM) classifiers, so as to compute distances in a high-dimensional feature space [Davy and Godsill 2002, Desobry et al. 2005, Harchaoui et al. 2009b, Sadjadi and Hansen 2010]. These methods have notably been applied successfully in the context of speaker segmentation to define different metrics in variations of the cepstral flux method [Fergani et al. 2006, Lin et al. 2007, Kadri et al. 2008].

Actually, the idea of using general distances for quantifying homogeneity had already been highlighted by Tzanetakis and Cook [1999] who envisioned the development of a generic methodology for audio segmentation. This methodology can be developed in three abstract stages, where the user extracts sound features from the audio signal to provide a time series of observations, computes the distance between successive observations to build a detection function, and finds peaks in the detection function to localize segment boundaries. The authors instantiated the methodology to provide an offline system based on Mahalanobis distance and several heuristics for constructing the detection function and finding the peaks.

We believe that a statistical perspective based on information divergences is a relevant approach to generalize this, and to address a sound and unifying framework with restricted needs for arbitrary heuristics on the choice of particular distances or detection strategies. Statistical approaches have already been used in audio segmentation, notably with the BIC method for speaker segmentation as discussed above. A major drawback of these approaches, however, is that they consider only normal assumptions on the observations, which is obviously not well-suited to all sound features. Secondary issues come from the potential computational complexity, or on the design heuristics used for speeding up the schemes to the detriment of the segmentation quality. Last but not least, not all these approaches, especially the two-pass schemes discussed above, are causal and thus suited to real-time constraints.

Causality is indeed primordial in a real-time context where we do not have access to the future and where the segmentation must be performed online by comparing the present to the past. Moreover, because a sound signal is a stream that unfolds in time in a causal manner, we also think that a causal design is still relevant in an offline setting, and that it is pertinent from a perceptual viewpoint to account for the inherent temporal flow of the time series. We therefore view the process of segmentation as the detection of a sufficient amount of novelty, that is, a rupture of information content between a given time point and its relative past.

General statistical approaches to online segmentation are provided by the theory of *detection of abrupt changes* [Basseville and Nikiforov 1993]. In particular, CUSUM approaches naturally fit with most statistical models of probability distributions, with recursive and tractable computations that are compatible with the design of a real-time system. These approaches, however, undergo approximations for parameter estimation before change. More precisely, the parameters before change are assumed to be known in advance when forming the statistical hypotheses of a change at the respective time points of the window, and when computing the respective LR statistics. This is suitable for applications such as quality control where a normal regime is completely known beforehand. It is yet limited in many real-world applications where we do not seek to detect failures of a known standard regime.

When applying CUSUM in such scenarios, the parameters before change, which are supposed to be known, are actually estimated either on the whole window, or in a dead region at the beginning of the window where change detection is turned off. The LR statistics are thus replaced with approximate GLR statistics, where the parameter before change is the same in all hypotheses. This results in practical shortcomings, because of estimation errors, as soon as change points occur rapidly, which is the case for audio signals in general.

The problem when considering properly unknown parameters before change and when forming exact GLR statistics, is that it breaks down the recursivity and computational efficiency of the detection schemes. Therefore, the CUSUM approximations of the exact GLR statistics are in general still employed to accommodate sequential situations. A few specific exact GLR statistics have yet been studied, notably for unknown mean in univariate normals with a fixed and known variance [Siegmund and Venkatraman 1995], and some extensions to multivariate normals as employed in certain methods for speaker segmentation discussed above.

A more general Bayesian framework for independent observations in exponential families has been proposed recently to address the estimation of parameters before and after change [Lai and Xing 2010]. This Bayesian framework, however, relies on a geometric prior on the time between change points, which is not always well-suited for arbitrary audio signals. Moreover, it requires a prior knowledge on the distribution of the parameters in the respective segments, which is not always available. To overcome this, we rather seek to employ sequential change detection schemes with unknown parameters before and after change, but without any a priori on the respective distributions of the change points and parameters.

4.1.3. Contributions

As explained above, many different approaches to audio segmentation have been studied in the literature, notably based either on statistical assumptions or on distance heuristics. Most approaches, however, consider specific segmentation tasks, and rely on particular statistical models, often normal distributions, or on particular distance functions. In this context, our main contributions can be summed up as follows.

We propose a generic framework for real-time audio segmentation that can handle various types of signals and of homogeneity criteria. The proposed system detects changes by controlling the information rate of the incoming audio stream as it unfolds in time. As a by-product, the quantified units can then be characterized by using representative probabilistic models within the respective segments, hence permitting their processing for further applications. More specifically, the proposed system relies on the computation of a short-time sound representation from the incoming audio stream, and on the modeling of the observed sound features with parametric probability distributions. The segmentation then consists in monitoring the parameters of the distributions in real time so as to assess structural changes that indicate new segments. The choice of the audio features and of their associated statistical model is left to the user depending on the homogeneity criterion in the application at hand. Considering distributions from exponential families, a wide range of common statistical families can be employed to model combinations of audio features with various topologies, such as scalar or multidimensional, discrete or continuous data. The sequential change detection is performed with the GLR statistics for exponential families. In addition to their primary statistical interpretation, these statistics also found geometrical grounds through the dually flat geometry of exponential families. As a result, the proposed framework paves the way for unifying several statistical and distance-based approaches to audio segmentation. Such approaches include CUSUM schemes based on LR statistics, approximate and exact GLR statistics, AIC and BIC methods from model selection theory, as well as certain kernel methods relying on SVM classifiers and similar approaches. In particular, the baseline spectral and cepstral flux methods commonly employed in onset detection and speaker segmentation can be seen as special instances of this modular framework. The system is finally applied to a variety of audio signals and segmentation tasks. The remainder of this chapter is organized as follows.

In Section 4.2, we present the proposed approach in more detail. In Subsection 4.2.1, we describe the architecture of the proposed real-time system. In Subsection 4.2.2, we focus on the change detection core scheme. We then clarify the relations between various standard approaches to audio segmentation, and show how they can be unified and generalized in the proposed framework. In Section 4.3, we apply the proposed system to a variety of audio signals and segmentation tasks. In Subsections 4.3.1, 4.3.2, 4.3.3, and 4.3.4, we notably showcase applications based on the energy for segmentation into silence and activity regions, on timbral characteristics for segmentation into music and speech, or into different speakers, and on spectral characteristic for segmentation into polyphonic note slices, respectively. This illustrates the generality of the included applications on different problems, by adapting the proposed framework to the homogeneity criteria considered. In Subsection 4.3.5, we further perform a quantitative evaluation for the specific task of onset detection in music signals on a complex dataset, and demonstrate how the proposed approach can leverage modeling compared to baseline approaches on this task.

4.2. Proposed approach

In this section, we present the proposed approach to a generic framework for audio segmentation with arbitrary types of signals and of homogeneity criteria. We first outline the general architecture of the real-time system designed, which relies on an arbitrary short-term sound representation and on its modeling through a parametric statistical model. We then elaborate on the sequential change detection scheme employed to monitor the variations in the model parameters, and notably clarify the relations with various standard approaches to audio segmentation in order to demonstrate how they can be unified and generalized in the proposed framework.

4.2.1. System architecture

The general architecture of the system is depicted in Figure 4.2. We consider an audio stream that arrives incrementally to the system as successive time frames. These frames are represented with an arbitrary short-time sound representation to provide a time series $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of observations. These observations are modeled

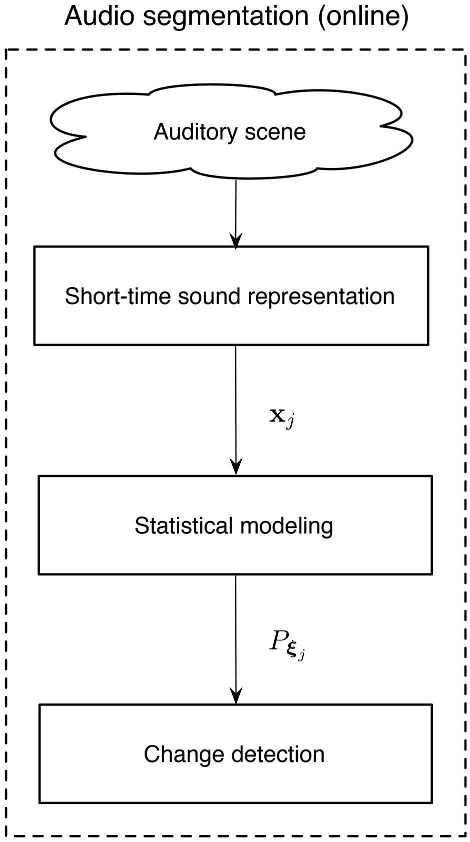

Figure 4.2.: Architecture of the proposed real-time system. The audio signal arrives online to the system, and is represented through arbitrary sound features that are modeled with parametric probability distributions, whose parameters are then monitored to detect changes.

with probability distributions $P_{\boldsymbol{\xi}_1}, P_{\boldsymbol{\xi}_2}, \ldots$ from a given parametric statistical family. The segmentation paradigm then consists in detecting sequentially the changes in the parameters $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots$ of the respective distributions. The sequential change detection procedure can be sketched as follows.

We start with an empty window $\vec{\mathbf{x}} \leftarrow (\,)$. For each time increment $n = 1, 2, \ldots,$ we accumulate the incoming observation $\mathbf{x}_n$ in the growing window $\vec{\mathbf{x}} \leftarrow \vec{\mathbf{x}} \,|\, \mathbf{x}_n$, and attempt to detect a change point in the parameters at any time i of the window. When a change point is detected, we discard the observations before the estimated change point i and start again the procedure with an initial window $\vec{\mathbf{x}} \leftarrow (\mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)$. The sequential change detection problem can therefore be reduced to finding one change point anywhere within a given data sample $\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. As a by-product, each segment can then be characterized by a statistical prototype corresponding to the constant parameters of the probability distribution in that segment.

Concerning the short-time sound representation, various and almost arbitrary sound features can be employed to produce the time series $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of observations. For example, we may compute a short-term Fourier transform, or any other time-frequency representation, so as to extract information on the spectral distribution of the audio stream and segment it according to spectral changes. Alternatively, we can employ timbral information for segmentation by computing MFCC observations, or loudness information through the energy envelope. The choice of these sound descriptors is left to the user depending on the types of signals and on the criteria for homogeneity considered. The proposed framework based on sequential change detection is able to handle any sound features in a unifying way, provided that some statistical assumptions are verified.

Because we seek a general framework for audio segmentation, we cannot rely on a simple statistical modeling such as using uniquely normal distributions. Indeed, normality assumptions are well-suited for certain types of sound features such as MFCC observations, but may fail to model reliably other features. For example, the simple energy descriptor provides a non-negative time series of observations which are quite unlikely to be distributed according to normal distributions whose support is the real line. More generally, sound features may be either scalar or multidimensional, exhibit various ranges, and take continuous or discrete values. Moreover, we may want also to associate several sound features with different topologies.

To handle such scenarios, we assume that the sound features can be reliably modeled with an exponential family of probability distributions. It leaves the choice for many common parametric families, such as Bernoulli, Dirichlet, Gaussian, Laplace, Pareto, Poisson, Rayleigh, Von Mises-Fisher, Weibull, Wishart, log-normal, exponential, beta, gamma, geometric, binomial, negative binomial, categorical, multinomial models, or for any association of such models. We now discuss a modular change detection paradigm that works for almost all exponential families.

4.2.2. Change detection

In the present work, we employ the GLR statistics for change detection with exponential families developed in Chapter 2. The decision rule based on the GLR statistics simply amounts to comparing the maximum of the GLR within the window with a threshold $\lambda > 0$. We also assume that the parametric statistical model under consideration is a full minimal steep standard exponential family $\mathcal{P} = \{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \mathcal{N}}$. We finally estimate the unknown parameters before and after change in the respective hypotheses by using maximum likelihood estimators. Under mild assumptions, the respective maximum likelihood estimates exist and are unique. The GLR statistics $\widehat{\Lambda}^i$ at the respective time points i of the window can then be expressed as follows:

$$\frac{1}{2}\widehat{\Lambda}^i(\bar{\mathbf{x}}) = \log \frac{\prod_{j=1}^{i} p_{\widehat{\boldsymbol{\theta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}(\mathbf{x}_j) \; \prod_{j=i+1}^{n} p_{\widehat{\boldsymbol{\theta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})}(\mathbf{x}_j)}{\prod_{j=1}^{n} p_{\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}(\mathbf{x}_j)} \; . \tag{4.1}$$

At this point, let us clarify some commonly confused relations between change detection methods based on the LR, approximate and exact GLR statistics, as well as on the AIC and BIC statistics. In informal terms, the LR and GLR statistics can be written as $-2\log(p(\bar{\mathbf{x}}|H_0)/p(\bar{\mathbf{x}}|H_1^i))$. The difference between the statistics lies in the consideration of the parameters in the hypotheses. In the LR statistics for CUSUM schemes, the parameters before and after change are completely known in advance. When they are both unknown, the exact GLR statistics estimate them in the different hypotheses by using the available data respectively before and after the hypothesized change point, as $\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j$, $\widehat{\boldsymbol{\eta}}^{i}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) = \frac{1}{i}\sum_{j=1}^{i}\mathbf{x}_j$, $\widehat{\boldsymbol{\eta}}^{i}_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) = \frac{1}{n-i}\sum_{j=i+1}^{n}\mathbf{x}_j$. In the approximate GLR statistics, however, only the parameter after change is actually estimated in the different hypotheses, while the parameter before change is estimated at once and set identical for all hypotheses. In general, it is estimated either on the whole window, as $\widehat{\boldsymbol{\eta}}^{i}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \approx \widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j$, or in a dead region at the beginning of the window where change detection is turned off, as $\widehat{\boldsymbol{\eta}}^{i}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \approx \widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \approx \frac{1}{n_0}\sum_{j=1}^{n_0}\mathbf{x}_j$, for some $n_0 < n$. It permits to keep the recursive form of the LR statistics and to use the computationally attractive CUSUM scheme, but it undermines the correct detection of changes, because of estimation errors, as soon as changes occur too rapidly.

Concerning the AIC and BIC methods for model selection, they consist in comparing the likelihood of the data under the respective hypotheses with a penalty increasing according to the number of free parameters, so as to favor sparse models over complex models and avoid overfitting, following the Occam's razor principle. The distinction between the two methods is actually that the penalty term augments with the sample size n in BIC, but not in AIC, hence favoring in general sparser models in BIC than in AIC. For detecting a change when both parameters are unknown, the differences of AIC or BIC between the hypothesis of no change and the respective hypotheses of a change are computed. These differences can be expressed respectively as $-2\log(p(\bar{\mathbf{x}}|H_0)/p(\bar{\mathbf{x}}|H_1^i)) + 2(k_0 - k_1^i)$, and $-2\log(p(\bar{\mathbf{x}}|H_0)/p(\bar{\mathbf{x}}|H_1^i)) + (k_0 - k_1^i)\log n$, where k_0 and k_1^i are the numbers of free scalar parameters estimated in the respective hypotheses. We actually have $k_0 = k_1^i/2 = d$, where d denotes the dimension of the parameter space. The model selection rule then simply amounts to detecting a change as soon as the maximum of the differences within the window is positive, that is, $-2\log(p(\bar{\mathbf{x}}|H_0)/p(\bar{\mathbf{x}}|H_1^i)) > 2d$, and $-2\log(p(\bar{\mathbf{x}}|H_0)/p(\bar{\mathbf{x}}|H_1^i)) > d\log n$, respectively. Therefore, the AIC and BIC methods are actually nothing else than particular exact GLR methods, with a constant threshold $2d$ for AIC, and with a threshold $d\log n$ that increases with the window length for BIC. A penalty parameter $\gamma > 0$ is also sometimes introduced to consider a penalized BIC with a threshold $\gamma\, d\log n$, which is again a specific GLR scheme.

Keeping this in mind, there is absolutely no reason that the AIC, BIC, or penalized BIC methods are computationally more demanding than the exact GLR method as sometimes argued in the literature. Nonetheless, the approximate GLR method is in general faster since it allows the implementation of a recursive CUSUM scheme, with the difference that the parameters after change need to be estimated in all hypotheses compared to the standard CUSUM with LR statistics for completely known parameters.

In addition to unifying LR, GLR, AIC, BIC methods and their variations, the proposed approach further provides a sound framework to bridge the gap between these statistical methods and geometrical methods based on the computation of distances. Indeed, the GLR statistics can also be expressed in terms of information divergences as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\, D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^{i}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})} \middle\| P_{\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}\right) + (n-i)\, D_{\mathrm{KL}}\left(P_{\widehat{\boldsymbol{\theta}}^{i}_{1\,\mathrm{ml}}(\bar{\mathbf{x}})} \middle\| P_{\widehat{\boldsymbol{\theta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})}\right) . \tag{4.2}$$

Rewriting the Kullback-Leibler divergences in terms of associated Bregman divergences on natural or expectation parameters $D_{\mathrm{KL}}\left(P_{\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}'}\right) = B_{\psi}(\boldsymbol{\theta}' \| \boldsymbol{\theta}) = B_{\phi}(\boldsymbol{\eta} \| \boldsymbol{\eta}')$, this paves the way for the design of arbitrary distance-based approaches to segmentation, provided that the considered distance is the canonical Bregman divergence of an exponential family. Furthermore, this lays a theoretical background to understand and guide the choice of a given distance in relation to the corresponding statistical assumptions on the distribution of the observations. This includes schemes based on the widely used Euclidean and Mahalanobis distances, and on the Kullback-Leibler and Itakura-Saito divergences, to name but a few.

In this context, it is worth comparing the distance-based segmentation derived from rigorous statistical considerations, to the heuristic distance-based segmentations usually employed in the literature. Most of the time, the segmentation based on a given distance relies on the direct computation of the distance between the observations at two successive time frames, or sometimes more generally between the left and right parts of a window after averaging the observations on the respective sides. In contrast, the proposed approach shows that the distance corresponding to the exact GLR statistics is rather computed separately for the two parts with respect to the global average as a reference. Interestingly, when using approximate GLR statistics where all parameters before change are assumed equal, the first distance vanishes and we end up with a scheme similar to heuristic distance-based schemes.[1]

The proposed approach is however more general since in a complete scheme, all time points of the window, and not just the center, are considered as potential change points. Moreover, the way of averaging the observations in the respective parts of the window is naturally provided by the mean of the sufficient observations, which geometrically simply corresponds to the arithmetic mean of the observed points in expectation parameters and actually coincides with the corresponding Bregman right-sided centroid. It is also interesting to remark that the two distances are then naturally weighted by the number of observations in the respective parts of the window, which cancels out when considering only the center. Last but not least, common windowing heuristics based on growing and sliding factors can be seen as approximations of a complete scheme, where the incoming observations are accumulated in groups of several observations, and the window is offset by several time frames as soon as it attains its maximal allowed size. As a result, this makes the scheme faster, but incomplete in the sense that not all time points are tested as candidate boundaries, and that the analysis step is larger than a single time frame, hence potentially resulting in detection errors and increased latency.

To go further, some approaches based on kernels methods can also be explained under the umbrella of the proposed framework. Because the sufficient observations in exponential families appear through their scalar product with the natural parameters, we can extend them to a high-dimensional feature space through a reproducing kernel Hilbert space. This has notably been formalized by Canu and Smola [2006] who demonstrated the equivalence between one-class SVM approaches to novelty detection, and approximate GLR statistics computed at the center of the window.

[1]The two schemes actually correspond if we only seek to detect a change point at the center of the window, and if we take the first half of the samples to estimate the unknown parameter before change.

Again, the proposed framework provides statistical and geometrical insights into these methods, and extends them with the introduction of exact GLR statistics.

From a computational perspective, exact GLR statistics have often been replaced with their approximate counterparts, except from normal models, as discussed previously. This is because the standard expression in terms of likelihoods may be intensive to compute, in particular for complete schemes where all potential change points are tested, and where the parameters in the respective hypotheses all need to be estimated to compute the different likelihoods. For normal models, computational optimizations have been proposed to overcome this by taking advantage of specific relations between the parameters in successive windows. We argue that such optimizations can actually be made more generally for all exponential families, and arise naturally from convex duality. Indeed, rewriting the exact GLR statistics in expectation parameters leads to the following expression:

$$\frac{1}{2}\,\widehat{\Lambda}^i(\bar{\mathbf{x}}) = i\,\phi(\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}})) + (n-i)\,\phi(\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})) - n\,\phi(\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}))\ . \tag{4.3}$$

Because maximum likelihood estimates between successive windows are related by simple time shifts or barycentric updates in expectation parameters, this provides a computationally efficient scheme for calculating the statistics in a sequential fashion. For example, if no change has been detected in the previous window, the statistics can then be simply updated as $\widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow \widehat{\boldsymbol{\eta}}^i_{0\,\mathrm{ml}}(\bar{\mathbf{x}})$, $\widehat{\boldsymbol{\eta}}^{n-1}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow \widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})$, $\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow ((n-i-1)\widehat{\boldsymbol{\eta}}^i_{1\,\mathrm{ml}}(\bar{\mathbf{x}})+\mathbf{x}_n)/(n-i)$, $\widehat{\boldsymbol{\eta}}^{n-1}_{1\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow \mathbf{x}_n$, $\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}}) \leftarrow ((n-1)\widehat{\boldsymbol{\eta}}_{0\,\mathrm{ml}}(\bar{\mathbf{x}})+\mathbf{x}_n)/n$, for all $1 \leq i < n-1$. Similar updates can be obtained when a change point has been detected, or when employing growing and sliding window heuristics. Moreover, certain values at which the conjugate ϕ is evaluated actually reappear because of time shifts, and can therefore be stored to facilitate tractability.

4.3. Experimental results

In this section, we report experimental results of the proposed approach on various types of audio signals and of homogeneity criteria. We notably showcase applications based on the energy for segmentation into silence and activity regions, and on timbral or spectral characteristics for segmentation into music and speech, into different speakers, or into polyphonic note slices. The generic audio segmentation framework presented above is capable of controlling information rate changes in real time given that the sound representation is modeled through a member of the ubiquitous exponential families. By unifying and generalizing several reference approaches to audio segmentation, this framework is expected to provide at least equivalent performances than the encompassed baseline methods. Therefore, the goal of this section is not to evaluate systematically the performance of the proposed schemes in comparison to the literature, but rather to illustrate the plurality of the included applications on different problems, by adapting the proposed framework to the signals and homogeneity criteria considered. A quantitative evaluation on a difficult dataset is nonetheless performed for the specific task of musical onset detection with various instruments and music styles, to demonstrate how the proposed approach can leverage modeling on a complex problem.

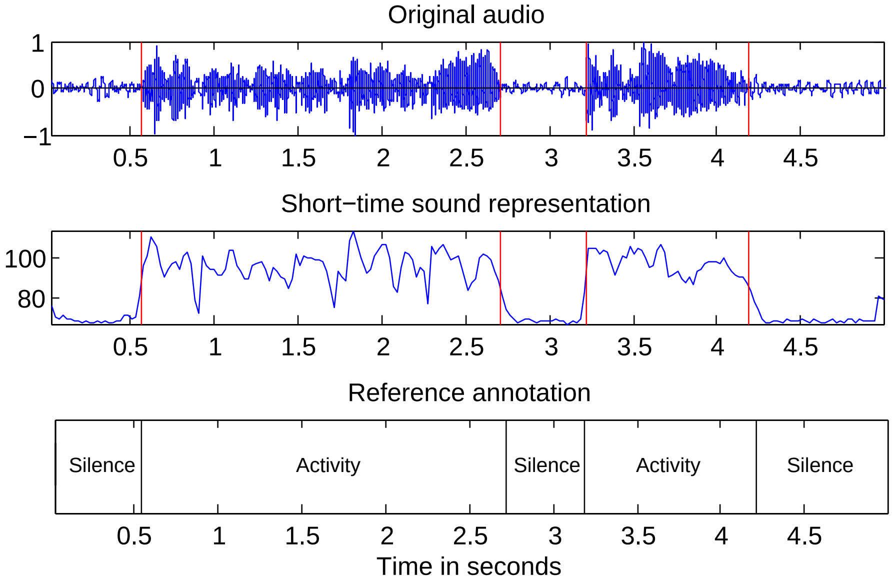

Figure 4.3.: Segmentation into silence and activity. By modeling the energy variations, the system has correctly detected the boundaries between silence and activity regions, despite the presence of background noise.

4.3.1. Segmentation into silence and activity

We begin with the problem of segmenting an audio signal according to variations of the energy. As a simple experiment, we considered the fundamental task of segmentation into silence and activity regions. Solving this basic problem is of primordial interest since a silence detector is the first pre-processing stage in numerous audio engineering systems. To showcase the system on this task, we analyzed a speech utterance containing pauses and background noise. As a sound representation, we chose the short-term energy computed through a standard 40-band filter bank on a Mel-frequency scale, with a frame size of 512 samples and a hop size of 256 samples at a sampling rate of 11025 Hz. We modeled the energy observations with Rayleigh distributions. The threshold was set manually to $\lambda = 1$. The change detection was computed under MATLAB on a 2.40 GHz laptop with 4.00 Go RAM, and was about 10 times faster than real time.

The results are represented in Figure 4.3. The top plot shows the computed segmentation on the audio waveform, while the bottom plot provides the reference annotation. It confirms that the system has reliably detected the boundaries between silence and activity regions despite the background noise. This is because we do not rely on a crude local analysis of the raw signal variations, but on a statistical modeling of the energy, whose distribution changes during silences compared to activity regions, as visible on the middle plot. As such, the system is able to discriminate correctly between silence and activity, but does not classify the segments accordingly. This could however be done easily by using the estimated scalar parameters in the respective segments, which for Rayleigh distributions characterize both the mean and variance of the observations.

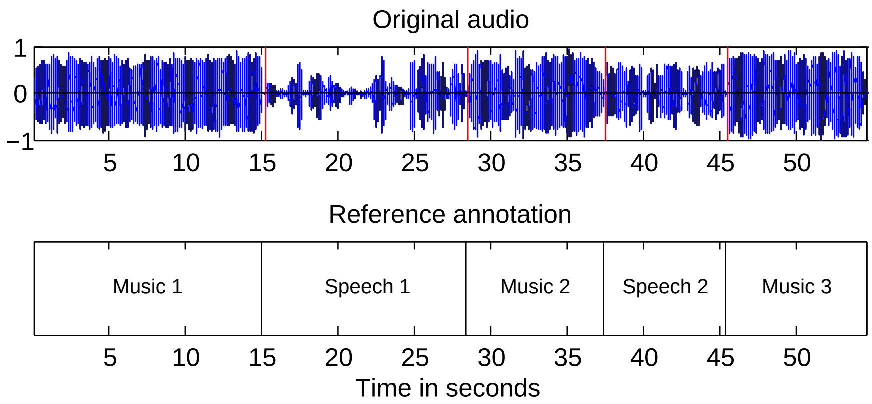

Figure 4.4.: Segmentation into music and speech. By monitoring variations in the timbre, the system has correctly found the different music and speech segments, while staying robust within the two classes.

4.3.2. Segmentation into music and speech

We now turn to the task of segmenting audio into music and speech, which arises in several applications such as the automatic indexing of radio broadcasts. The discrimination between music and speech can be addressed with a timbral homogeneity criterion. In this context, a baseline sound representation is given by the MFCC observations. To illustrate this, we created an audio fragment by concatenating music excerpts and speech utterances. To make the example realistic, we chose three music excerpts with vocals, and two speech passages with utterances from two different speakers in the first passage. This was done to verify whether the system is able to discriminate between the classes of music and speech despite the presence of vocals, while being robust against speaker changes within the class of speech. We computed the first 12 MFCC observations excluding the 0th energy coefficient, with a frame size of 512 samples and a hop size of 256 samples at a sampling rate of 11025 Hz. We modeled the MFCC observations through multivariate spherical normal distributions with a fixed variance of 100. The threshold was set manually to $\lambda = 300$.[2] The change detection was about 10 times faster than real time.

The results are represented in Figure 4.4. The top plot shows the obtained segmentation on the audio waveform, while the bottom plot shows the ground-truth segmentation. This proves that the system has correctly detected the changes between music and speech despite the presence of vocals in the music excerpts. Moreover, there is no over-segmentation within the class of speech even if two different speakers are present in the first speech region. Here for spherical Gaussians, the estimated multivariate parameters in the segments actually correspond to the exact means of the 12 MFCC observations within the respective segments, and could further be employed to classify the obtained regions as music or speech.

[2] There is actually a redundancy between the variance and the threshold for spherical normal distributions, where multiplying the variance by some positive value amounts to dividing the threshold by the same value.

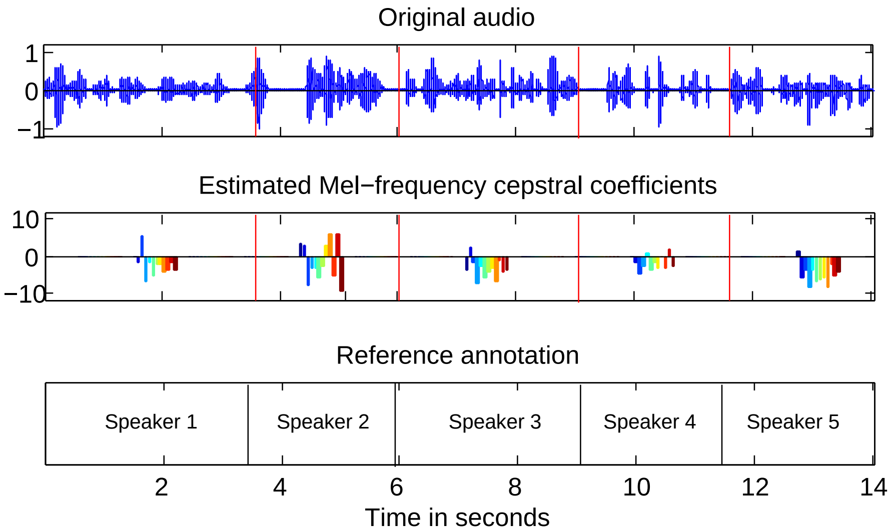

Figure 4.5.: Segmentation into different speakers. By considering characteristic timbre structures within the class of speech, the system has correctly identified the boundaries between the different speakers.

4.3.3. Segmentation into different speakers

To go further, we now experiment the segmentation of a speech signal into different speakers. We constructed a speech fragment by concatenating utterances from 5 different speakers. Because the timbre is also characteristic of speakers at a smaller scale than speech and music, we chose the same analysis parameters as above and computed MFCC observations. The threshold was however refined to $\lambda = 100$, for a finer timbral discrimination at the speaker level within the class of speech. The change detection was about 10 times faster than real time.

The results are represented in Figure 4.5. The top plot shows the obtained segmentation on the audio waveform, while the bottom plot shows the annotated reference segmentation. This proves that the system has correctly detected the different speaker turns given the usual 1-second tolerance in this context. The middle plot depicts the estimated MFCC parameters for the different speakers in the respective segments. These estimated coefficients actually correspond to the mean of the MFCC observations, and are clearly distinctive of the different speakers. As a result, they could be beneficially employed for further applications such as speaker recognition or diarization.

4.3.4. Segmentation into polyphonic note slices

We now turn to the segmentation of audio based on a spectral homogeneity criterion. Spectral characteristics provide complementary information to the timbre, by representing finely the frequency content rather than the global frequency shape, or spectral envelope, of the source. Spectral segmentation therefore applies in general at a shorter-term level, for tasks such as detecting phoneme boundaries or musical notes. We focus on the latter application in the context of slicing polyphonic music into stationary polyphonic chunks. We considered an excerpt of the 4th movement

Les entretiens de la Belle et de la Bête, from *Ma mère l'Oye, Cinq pièces enfantines pour piano à quatre mains* (1908-1910), composed by Maurice Ravel (1875-1937). This excerpt was synthesized from a ground-truth reference with real piano samples. In seeking to obtain stationary polyphonic slices, we actually want to detect note onsets and offsets, so that each segment exhibit a constant combination of note events that differs from its neighbors. As a sound feature, we employed a normalized magnitude spectrum, computed using a simple short-time Fourier transform with a frame size of 512 samples and a hop size of 128 samples at a sampling rate of 11025 Hz. We considered the normalized magnitude spectra as frequency histograms and modeled them with categorical distributions. The threshold was set manually to $\lambda = 10$.[3] The change detection was about 10 times faster than real time.

The results are represented in Figure 4.6. The top plot represents the detected segments on the audio waveform, while the bottom plot explores their relevancy compared to the ground-truth note pitches presented as a piano roll. This is to confirm that the system has successfully detected change points that actually correspond to note onsets and offsets. We insist on the fact that the system has however no knowledge about musical notes, and rather detects changes by monitoring the variations of the spectral information over time. For categorical distributions, the estimated parameters within the respective segments correspond to the probabilities of occurrence of the respective variables. For frequency spectra, it provides spectral information as the distribution of the different frequency bins. Therefore, the segmentation and respective estimated spectral distributions could also be used for further applications in music information retrieval.

4.3.5. Evaluation on musical onset detection

To complement the above examples and discussions, we finally provide a quantitative evaluation of the proposed approach on the specific task of musical onset detection. We notably demonstrate the benefits of the proposed approach compared to the heuristic distance-based method of spectral flux. We considered a well-known and difficult dataset described in [Leveau et al. 2004]. This dataset is composed of 17 heterogeneous music clips recorded in various conditions, and ranging from solo performances of various monophonic instruments and polyphonic instruments, to complex mixes, in different music genres such as rock, classical, pop, techno, jazz.

For online segmentation, we computed a normalized magnitude spectrum with a frame size of 1024 samples and a hop size of 126 samples at a sampling rate of 12600 Hz, for a time resolution of 10 ms. We modeled these features with categorical distributions. The threshold was chosen constant, and was tuned with a step of 0.1 in the range $1 \leq \lambda \leq 10$, so as to achieve optimal results over the dataset.

We compared the results of the proposed algorithm (`GLR`) with the three most common variations of the spectral flux method (`SF`), based respectively on the Kullback-Leibler divergence, the Euclidean distance, and the half-wave rectified difference,

[3] There is actually equivalence between considering categorical distributions with a certain threshold, and the corresponding i.i.d. sampling model of multinomial distributions with a given number of trials where the threshold is divided by the number of trials.

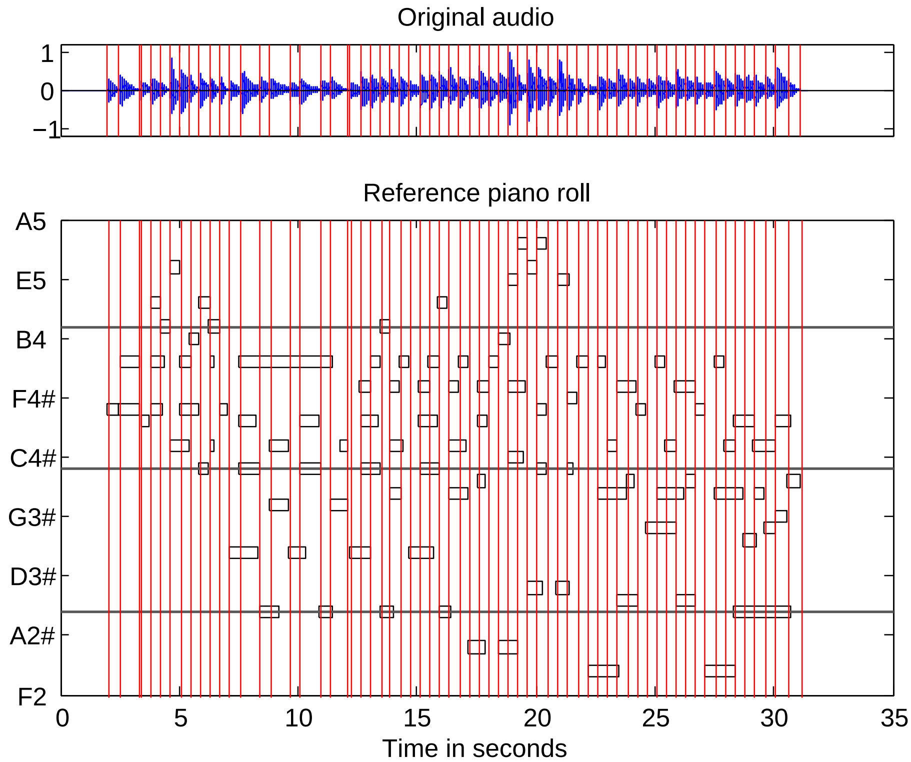

Figure 4.6.: Segmentation into polyphonic note slices. By modeling the spectral variations, the system has correctly sliced the music excerpt into stationary polyphonic chunks.

computed between two successive time frames. The threshold was also chosen constant and was tuned with a step of 0.01 in the range $0.01 \leq \lambda \leq 1$.

The performance of the algorithms was measured through the F-measure $\mathcal{F}$ computed as the harmonic mean of the precision and recall, where the precision $\mathcal{P}$ is the percentage of correctly detected onsets over the total number of detected onsets, and thus quantifies the robustness of the algorithm in relation to true negatives and false positives, while the recall $\mathcal{R}$ is the percentage of correctly detected onsets over the total number of ground-truth onsets, and thus quantifies the efficiency of the algorithm in relation to true positives and false negatives. According to the methodological guidelines provided in [Leveau et al. 2004], we assumed an onset to be correctly detected if it is within a time tolerance of ± 50 ms from a ground-truth onset. We also considered doubled and merged onsets to account for them as errors.

We report the evaluation results in Table 4.1. Overall, the results show that the proposed system and `GLR` algorithm perform relatively well on this complex dataset with a maximum F-measure $\mathcal{F} = 64.52$. Moreover, it significantly outperforms all reference `SF` algorithms. Using the simple Euclidean distance, `SF` has the lowest F-measure $\mathcal{F} = 27.08$. The Euclidean distance actually corresponds to spherical Gaussian assumptions on the observations, which is not adapted to model the magnitude spectrum. Concerning the Kullback-Leibler distance function, it exactly corresponds to the divergence related to the categorical model. The corresponding `SF` is thus a crude approximation of `GLR`, with search for a change point in a sliding

Algorithm	λ	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Distance function
GLR	5.00	60.93	68.55	**64.52**	Kullback-Leibler
SF	0.06	22.56	33.87	27.08	Euclidean
SF	0.10	34.42	41.26	37.53	Kullback-Leibler
SF	0.17	40.20	42.74	**41.43**	Half-wave rectified difference

Table 4.1.: Evaluation results for musical onset detection. The results show that the proposed approach performs relatively well, with a maximum F-measure $\mathcal{F} = 64.52$, compared to the baseline spectral flux methods.

window of two observations, and with estimation of the unknown parameter before change using the first observation. It is thus expected to perform poorly compared to `GLR`, and exhibits a F-measure $\mathcal{F} = 37.53$. It is nonetheless higher than for `SF` with the Euclidean distance, thanks to the more reliable modeling assumptions underlying the Kullback-Leibler divergence. This is confirmed by the precision and recall which are both higher too, meaning that the effect is not due to one algorithm detecting more ground-truth onsets while also making more detection errors.

The best results for `SF` are actually obtained for the heuristic half-wave rectified difference with a F-measure $\mathcal{F} = 41.43$. This distance function does not correspond to a Bregman divergence, and hence does not have a rigorous statistical interpretation within the proposed framework. Nevertheless, the half-wave rectification is often used in onset detection based on the spectral flux, in order to account only for the significant variations of the spectra that correspond to positive contributions of energy, and thus to onsets rather than offsets. The other `SF` and `GLR` algorithms employed here do not account for this and therefore also detect spectral changes corresponding to note offsets. Nonetheless, the proposed `GLR` outperforms the heuristic `SF` method with the half-wave rectified difference. This is again confirmed by the precision and recall which are also both higher for `GLR`.

To compare the obtained results, we also performed the evaluation for two systems at the state-of-the-art. The first one (`LFSF`) was developed by Böck et al. [2012], and is a variant of spectral flux with several pre- and post-processing stages, including the application of a filter bank with a logarithmic frequency scale for computing the spectra. The second one (`TSPC`) was developed by Röbel [2011], and relies on the classification of the spectral peaks into transients and non-transients.

The comparative results are presented in Table 4.2. Overall, they show that `GLR` does not reach state-of-the-art results in terms of F-measure $\mathcal{F}$ when using the simple spectral representation discussed above. This is confirmed by the precision and recall $\mathcal{P}, \mathcal{R}$, which are both less for `GLR` than for `LFSF` and `TSPC`. It means that even if `GLR` outperforms the different variants of `SF` when sharing a simple spectral representation, the enhanced spectral representation and other heuristics introduced in `LFSF` significantly improves the results over `GLR`. This points out the fact that more elaborate representations should be investigated in combination with `GLR` in future work, as discussed later. We notice, however, that by encompassing general spectral flux methods, the proposed approach is likely to reach competitive results with the state-of-the-art when considering similar enhanced representations and heuristics.

Algorithm	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
`GLR`	60.93	68.55	**64.52**
`LFSF`	79.60	86.02	**82.69**
`TSPC`	76.46	84.27	80.17

Table 4.2.: Comparative results for musical onset detection. The results show that the proposed system and algorithm do not reach state-of-the-art results for the simple spectral representation considered here, although it would likely be competitive with the state-of-the-art approaches when employing an enhanced spectral representation as well as pre- and post-processing heuristics, since it encompasses general spectral flux methods such as the one evaluated here.

4.4. Discussion

In this chapter, we discussed the problem of real-time audio segmentation. We addressed this problem by proposing a generic and unifying framework capable of handling various types of signals and of homogeneity criteria. The proposed approach is centered around a core scheme for sequential change detection with exponential families, where the audio stream is segmented as it unfolds in time, by modeling the distribution of the sound features and monitoring structural changes in these distributions. We also discussed how the framework unifies and generalizes statistical and distance-based approaches to segmentation through the dually flat geometry of exponential families. We finally showcased various applications to illustrate the generality of the proposed approach, and notably performed a quantitative evaluation for musical onset detection to demonstrate how it can leverage modeling in a complex task. The obtained results are encouraging for further developments from both theoretical and applicative perspectives.

First of all, we want to tackle the assumption that the sound features are statistically independent along time. This has not been a serious issue for the audio features considered here, but it may hinder the correct modeling of certain features that exhibit more complex temporal profiles with clear temporal dependence. A potential direction is to leave the generality of the framework aside, and to consider change detection within specific statistical models, such as the autoregressive models employed for speech segmentation in [André-Obrecht 1988]. More generally, we could consider linear or even non-linear systems. Online schemes based on particle filtering have been proposed to detect changes in non-linear systems for instance in [Fearnhead and Liu 2007], but such schemes suffer from computational burdens when properly considering unknown parameters and exact inference. An alternative based on CUSUM-like test statistics has recently been proposed in [Vaswani 2007].

On another perspective, we would like to improve the robustness of the system. A first possibility is to employ more elaborate post-processing techniques. Here, following a statistically-grounded viewpoint, we considered the simple thresholding of the test statistics along time as a decision rule. Yet we could take advantage of application-specific heuristics to tailor the detection rule by smoothing the test statistics and adapting the threshold, or by considering sliding and growing heuristics for windowing. More interestingly, it would be a major step to assess such heuristics in relation to statistical considerations. This has been recently investigated in [Lai and Xing 2010], where a scanning window decision rule is analyzed from the standpoint of asymptotic optimality. This should be investigated further.

A complementary idea has been discussed here, by clarifying the relations between the GLR statistics and the AIC or BIC methods from model selection, hence providing a way to automatically tune the threshold, not necessarily constant across the windows. A more specific criterion than AIC and BIC has also been proposed for estimating the number of change points in an offline setup with exponential families in [Lee 1997], and could beneficially be employed. The issue of these criteria, however, is that they rely on the large sample theory, which is not always relevant for certain signals such as audio where the time intervals between segments may be low. This is why an additional threshold parameter is for example often used to define penalized BIC methods as discussed previously. Moreover, when computing exact statistics, we also test for change points at the window extremities, and relevant criteria should account for this, by correcting the statistics for small sample sizes. It has been noticed and addressed for offline contexts early in [Deshayes and Picard 1986], or more recently in [Chen et al. 2006], and should be adapted to online contexts.

Besides the non-Bayesian framework employed here, we may also formulate online change detection in a Bayesian framework. This becomes interesting when we possess prior knowledge on the distributions of the parameters in the respective segments, and on the run length distribution between change points, either by expert knowledge or by learning these distributions on a training dataset. In this context, several frameworks have already been proposed, for example in [Adams and MacKay 2007, Turner et al. 2009, Lai et al. 2009, Lai and Xing 2010, Fearnhead and Liu 2011], certain dealing notably with exponential families. The inference schemes, however, are in general more demanding than for non-Bayesian approaches. This avenue merits further considerations and adaptations to real-time processing.

In addition to improving the core scheme of change detection, a parallel effort could be pursued on considering more elaborate representations and modeling of sounds. In particular, the framework of exponential families permits the combination of various features with different topologies. Combined with kernel methods, it provides an interesting approach to enhance the standard sound representations employed in the present work. In this context, it would become pertinent to also provide methods for automatic feature selection and model selection, in order to fit the representational front-end and its statistical modeling to the problem at hand, either in an unsupervised setup with no input from the user, or in a semi-supervised approach.

More specifically, employing enhanced representations tailored to specific tasks may improve the results compared to the simple energy, cepstral or spectral representations considered here. It would notably be worth investigating enhanced spectral representations through logarithmic or perceptual frequency scales, by using filter banks as well as pre- and post-processing heuristics, to make the detection of onsets in music signals more robust and reach competitive results with the state-of-the-art for this task. On another line, discrete musical representations may also be modeled with exponential families, for example, through Zipf laws, to detect changes in symbolic musical data, represented as a list of notes, intervals, or patterns, for a high-level segmentation into chords, harmony, or tonality, among other possibilities.

From the standpoint of applications, we now want to evaluate the proposed framework thoroughly in a variety of direct audio applications. We have showcased here several such applications, including segmentation of audio into silence and activity, or into speech and music, and segmentation of speech into different speakers, or of polyphonic music into polyphonic note slices. Other applications are conceivable, for example, music structural segmentation based on timbral characteristics, as well as segmentation of speech in phonemes or syllables, among others. We have demonstrated, on the specific task of musical onset detection, how the proposed framework can improve segmentation compared to heuristic baseline methods. This should be extended on a case-by-case study to the various applications mentioned. Because the proposed framework encompasses several baseline approaches in these respective applications, it is expected that it would provide benefits and further intuitions on the choice of models and distance functions considered for improving the results.

Finally, the proposed audio segmentation framework could be beneficially employed for further applications as briefly touched upon here. We can not only use the computed segmentation as a first pre-processing stage in a variety of applications, but also employ the estimated statistical prototypes characterizing the respective segments, hence opening up the field for further processing in audio analysis, indexing and information retrieval. Relevant examples include real-time speaker diarization, or real-time music structural analysis, among others. Last but not least, we believe that the proposed approach can also provide benefits for the analysis of time series in other domains within the realm of signal processing, including geophysics, econometrics, medicine, or image analysis.

5. Real-Time Polyphonic Music Transcription

In this chapter, we investigate a second problem of major interest in audio signal processing, and more particularly in music information retrieval, namely the automatic transcription of polyphonic music. We consider a supervised setup based on non-negative decomposition, where the music signal arrives in real time to the system, and is projected onto a dictionary of note spectral templates that are learned offline prior to the decomposition. An important drawback of existing approaches in this context is the lack of controls on the decomposition, resulting in practical shortcomings regarding the correct detection of notes. This issue is addressed by employing the methods developed for non-negative matrix factorization with convex-concave divergences in Chapter 3. In particular, we focus on the parametric family of (α, β)-divergences, and explicitly interpret their relevancy as a way of controlling the frequency compromise during the decomposition. The proposed system is then evaluated through a methodological series of experiments, and is shown to outperform two state-of-the-art offline systems while maintaining low computational costs that are suitable to real-time constraints.

5.1. Context

In this section, we first provide some background information on the problem of music transcription, in particular for polyphonic music signals and approaches based on non-negative matrix factorization techniques. We then discuss the motivations of our approach to the task of polyphonic music transcription with real-time constraints. We finally sum up our main contributions in this context.

5.1.1. Background

The task of *music transcription* consists in converting a raw music signal into a symbolic representation such as a score, as sketched in Figure 5.1. In more details, the purpose of music transcription is to analyze the low-level information of a music signal given as a simple audio waveform, in order to extract some high-level symbolic information that describes its musical content. Several types of musical information are of interest for music transcription, for example, when considering pitched instruments, we may want to find the respective pitches and onsets of the notes played, as well as their durations or their offsets. Other information may be of complementary interest, such as the dynamics, the tempi and rhythms, the instruments, the unpitched events from percussions, among others.

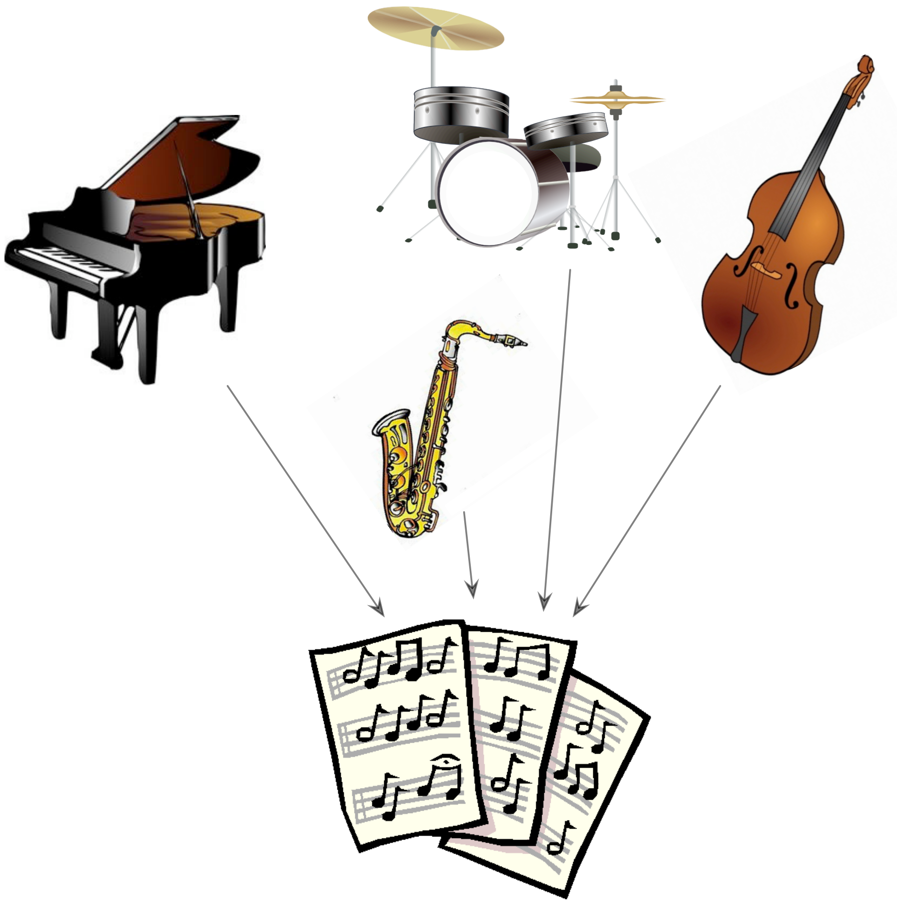

Figure 5.1.: Schematic view of the music transcription task. Starting from the low-level information of a raw music signal given as a simple audio waveform, the goal is to extract some high-level information that describes its musical content as a symbolic music score does.

Because of the different types of relevant musical information, the approaches to music transcription are in general aimed at extracting a particular type of information, and we are still far today from a single system capable of transcribing music like a human expert would. A recent review of the literature is provided in the book edited by Klapuri and Davy [2006], which gathers contributed chapters of leading researchers for the main approaches. We notably consider here the important and difficult subtask of transcribing the note information from polyphonic music. This is closely related to the problem of *multiple fundamental frequency estimation*, or *multiple pitch estimation*, which consists in determining the fundamental frequencies that are present in the signal as a function of time, in the scenario where there are potentially several overlapping pitched sound sources. This problem has been largely investigated for music as well as speech signals, and a wide variety of methods have been proposed as exposed in the comprehensive book chapter of de Cheveigné [2006].

In particular, non-negative matrix factorization (NMF) has been widely used since the pioneering work of Smaragdis and Brown [2003], Abdallah and Plumbley [2004].

In this context, the observation matrix $\mathbf{Y}$ is in general a time-frequency representation of the sound excerpt to analyze. The rows and columns represent respectively the different frequency bins and the successive time frames of the analysis. The factorization $\mathbf{Y} \approx \mathbf{AX}$ can then be interpreted as follows. The observation vectors $\mathbf{y}_j$ provide the spectral distribution of the signal at the respective time frames. The dictionary matrix $\mathbf{A}$ contains basis vectors as characteristic templates of the spectral distribution of the signal. The encoding vectors $\mathbf{x}_j$ represent the activations of the respective spectral templates at the corresponding time frames.

The main issue of the early NMF approaches to polyphonic music transcription stems from the lack of controls on the factorization. Notably, there is no guarantee that the obtained dictionary is made of spectral templates that actually correspond to musical notes. As a result, the activations of these templates along time do not necessarily correspond to the exact occurrence of notes. Moreover, there is a priori no structure in the learned dictionary, and ad hoc techniques are required to associate each spectral template to a musical pitch, for instance, by using hand grouping and labeling, automatic classification, or single-pitch estimation.

To circumvent these issues, several specific extensions have been developed, such as a source-filter model [Virtanen and Klapuri 2006], a pure harmonic constraint [Raczyński et al. 2007], a dictionary constrained to note spectral templates [Niedermayer 2008], a selective sparsity regularization [Marolt 2009], or a subspace model of basis instruments [Grindlay and Ellis 2009]. Considering such models ensures that the dictionary in the factorization is structured in sound entities that actually correspond to musical notes whose pitch is known beforehand. This makes the interpretation of the factorization easier and permits to directly process the activations to find the note pitches that are present in the music excerpt.

Similar approaches in the framework of probabilistic models with latent variables also share common perspectives with NMF techniques and have been employed for polyphonic music transcription. In particular, in the framework of probabilistic latent component analysis (PLCA), the non-negative data are considered as a discrete distribution and are factorized into a mixture model where each latent component represents a source [Smaragdis and Raj 2007, Shashanka et al. 2008]. It can then be shown that maximum likelihood estimation of the mixture parameters amounts to NMF with the Kullback-Leibler divergence, and that the expectation-maximization algorithm is equivalent to the classic multiplicative updates scheme. Considering the problem in a probabilistic framework is however convenient for enhancing the standard model, and for adding regularization terms through priors and maximum a posteriori estimation instead of maximum likelihood estimation.

In particular, PLCA has been employed in polyphonic music transcription to include shift-invariance and sparsity [Smaragdis et al. 2008]. Recent works have extended the latter model to include a temporal smoothing and a unimodal prior for the impulse distributions [Mysore and Smaragdis 2009], a hierarchical subspace model for representing instrument families [Grindlay and Ellis 2011], a scale-invariance [Hennequin et al. 2011b], a time-varying harmonic structure [Fuentes et al. 2011], and multiple spectral templates [Benetos and Dixon 2011a].

The above approaches consider either the standard Euclidean distance or the Kullback-Leibler divergence as a cost function. Recent works on music analysis have also investigated the relevance of other cost functions such as the Itakura-Saito divergence [Bertin et al. 2009, 2010, Févotte et al. 2009, Févotte 2011, Dikmen and Févotte 2011, Lefèvre et al. 2011a], with various extensions, including penalties for sparsity or temporal smoothness, and harmonic models. The more general parametric β-divergences have also been employed successfully for music and sound analysis. Relevant examples include audio systems for speech analysis [O'Grady and Pearlmutter 2008], musical source separation [FitzGerald et al. 2009], polyphonic music transcription [Vincent et al. 2010], and non-stationarity modeling of audio with a parametric model for spectral templates [Hennequin et al. 2010], or with a source-filter model for time-varying activations [Hennequin et al. 2011a].

5.1.2. Motivations

Our main goal is to devise a robust real-time system for the transcription of polyphonic music. Since several note events may overlap, we cannot use single-source detection techniques such as a simple correlation between spectral note templates and the audio stream. Instead, we rely on NMF techniques that permit to cope with the simultaneity of the detected sound events. Nonetheless, most NMF techniques are inherently suitable only for offline processing of the data, and this is the case for all music analysis systems discussed above. We are thus interested in adapting such systems to real-time constraints, where we cannot learn the factorization on a whole excerpt of music, but rather need to determine the occurrence of musical pitches rapidly as the audio stream unfolds in time.

A direct approach to achieve this is to rely on the supervised NMF problem of non-negative decomposition. In this approach, the sound sources are represented with a fixed dictionary of event templates, which is learned offline prior to the decomposition, and onto which the audio stream is projected incrementally as it unfolds in time to provide the activations of the respective templates. The occurrence of the different events can then be determined by a local analysis of their activations around the current time. As a result, we do not need to store a long audio excerpt for learning the factorization. Moreover, because we can a priori structure the dictionary of templates before the decomposition, we no longer need to perform classification nor optimize elaborate structured models during the real-time processing of the audio stream for decoding the nature of the events. This makes the approach potentially suitable to real-time constraints.

This approach has already been investigated by several authors. For example, a real-time system to identify the presence and determine the pitch of one or more voices was proposed by Sha and Saul [2005], and was further adapted to sight-reading evaluation of solo instrument by Cheng et al. [2008]. Cont [2006] also designed a real-time system for polyphonic music transcription, which was further enhanced by Cont et al. [2007] for real-time coupled multiple-pitch and multiple-instrument recognition. An important drawback of these approaches, however, is the lack of relevant controls on the robustness and efficiency of the decomposition.

Indeed, during the decomposition of a signal, the price to pay for the simplicity of a standard scheme is the misuse of event templates to reconstruct this signal. For the specific task of polyphonic music transcription, this amounts to common note insertions and substitutions such as octave or harmonic errors. The issue is almost as serious in a general pattern recognition setting where different classes of events are allowed to overlap in time with the presence of noise. In such realistic cases, providing controls on the decomposition can improve the detection of the different events. Yet in the literature, little attention has been payed to providing such controls. To the best of our knowledge, we are only aware of the system of Cont [2006], Cont et al. [2007], where a control on the decomposition is provided by enforcing the solutions to have a fixed desired sparsity.

In our context, controlling the sparsity of the solutions can help with reducing the space of plausible results and increasing the economy of class usage during decomposition. In most applications, however, the user does not know in advance the sparsity of the solutions and cannot estimate it easily. Moreover, the sparsity may also change along the signal. For example, in the problem of polyphonic music transcription, sparsity is highly correlated to the number of notes played simultaneously at each instant, which is both unknown and variable. In the system of Cont [2006], Cont et al. [2007], sparsity is nonetheless considered as fixed over the whole signal. We investigated in prior work the use of more flexible controls on sparsity by considering both bound constraints and penalties, rather than the hard constraint of the latter approach. In our experiments, it helped to improve the recognition of acoustic events in complex environmental auditory scenes with background noise, but it did not improve systematically the transcription of polyphonic music, where the same effect as sparsity could be obtained by controlling a simple threshold on the activations. Therefore, we do not discuss this approach further, and rather concentrate on other types of control which reveal relevant for polyphonic music transcription.

In particular, we seek here to introduce controls on the frequency trade-off of the decomposition between the different frequency components. Such a control may permit to better balance the consideration of certain important frequencies in the decomposition, such as partials for musical notes, and thus improve detection. Since the important frequencies to emphasize for polyphonic music transcription vary along time directly with the notes played, we cannot rely on a simple global weighting of the frequency distribution. We are thus interested in more adaptable techniques, where the control is flexible enough so as to fit the weighting to the current dynamic of the signal.

Such an approach has been addressed by using the parametric β-divergences in several offline systems discussed above. All these systems however employ heuristic multiplicative updates and lack convergence guarantees, notably concerning the monotonic descent of the cost function during the iterations.[1] This is of course undesirable for designing a robust real-time system, where we cannot accept unpredictable and unstable behavior of the core procedure during the iterations. This is alleviated in the real-time systems mentioned by using either the standard Euclidean distance or the Kullback-Leibler divergence as cost functions, for which monotonically decreasing multiplicative updates are well-known. Yet some recent advances on monotonically decreasing updates for the β-divergences presented by Nakano et al.

[1]Nonetheless, some of these heuristic multiplicative updates can a posteriori be proved to make the cost function decrease monotonically by using appropriate auxiliary functions, as discussed by Févotte and Idier [2011].

[2010a], Févotte and Idier [2011], would permit to design a real-time system based on these cost functions while maintaining a certain safety in the quality of the output solutions and in the behavior of the system. Last but not least, the more general (α, β)-divergences proposed by Cichocki et al. [2011] encompass both α-divergences and β-divergences, and could beneficially be employed. Yet they are quite recent and we are not aware of previous work that studies their relevancy in audio analysis.

5.1.3. Contributions

We follow an approach to polyphonic music transcription via non-negative matrix factorization. To sum up the above discussion, the main downside of most existing systems in this context is that of their intrinsic offline nature. A couple of real-time systems have yet been developed by employing the non-negative decomposition framework. Nonetheless, because they simply rely on the Euclidean distance or on the Kullback-Leibler divergence to construct the cost function, these systems do not feature any suitable controls on the decomposition. Our contributions in this context can be discussed as follows.

We develop a real-time system for polyphonic music transcription. The proposed system relies on the supervised setup of non-negative decomposition. In this setup, the music signal arrives in real time to the system and is represented through a short-time frequency transform. The spectral representations are then projected as they arrive in time onto a dictionary of note spectral templates that are learned offline prior to the decomposition. On the contrary to previous similar approaches, we propose to use the parametric family of (α, β)-divergences for decomposition. In addition to the known robustness properties of these divergences for statistical estimation, we provide other insights into their relevancy for audio analysis, by interpreting the effect of the parameters as a way to introduce a flexible control on the frequency compromise during the decomposition. Moreover, this is the first time to our knowledge that the recently introduced (α, β)-divergences are applied and interpreted in the framework of audio analysis. We also discuss some computational issues of the non-negative decomposition with (α, β)-divergences. We notably expand multiplicative updates tailored to real time by taking advantage of the decomposition framework where the dictionary of note spectral templates is kept fixed. Under mild assumptions, which can be obtained after some basic pre-processing of the observations, these multiplicative updates ensure the monotonic decrease of the cost function and thus its convergence. It guarantees the stable behavior of the system with regards to the quality of the output activations. The proposed system is finally evaluated through a methodological series of experiments. In these two tasks, the proposed approach is shown to outperform two state-of-the-art offline systems, while maintaining low computational costs that are suitable to real-time constraints. The remainder of this chapter is organized as follows.

In Section 5.2, we present the proposed approach in more detail. In Subsection 5.2.1, we describe the architecture of the proposed real-time system. In Subsection 5.2.2, we focus on the non-negative decomposition algorithm with the parametric family of (α, β)-divergences. In particular, we discuss some implementation issues regarding complexity and convergence guarantees. In Section 5.3, we evaluate the system through a methodological series of experiments. In Subsection 5.3.1, we notably showcase a sample example of piano music to illustrate the discussion and provide qualitative insights into the effect of the parameters in the (α, β)-divergences. This highlights that a region of parameters is of particular interest for polyphonic music transcription, in concordance with the interpretation of the parameters as a way to control the frequency compromise in the decomposition. In Subsections 5.3.2 and 5.3.3, this parameter region is further explored through quantitative experiments in a standard evaluation framework, for the respective tasks of multiple fundamental frequency estimation at the frame level, and multiple fundamental frequency tracking at the note level.

5.2. Proposed approach

In this section, we present the proposed approach to polyphonic music transcription. We first outline the general architecture of the real-time system designed, which relies on the non-negative decomposition of the incoming music stream into note events provided a priori to the system as a dictionary of note templates. We then elaborate on the non-negative decomposition scheme by considering the (α, β)-divergences as a cost function, and we interpret the relevancy of these information divergences in the context of audio analysis as a way to control the frequency compromise during the analysis.

5.2.1. System architecture

The general architecture of the system is depicted in Figure 5.2. The system can be divided into two distinct modules. On the right side of the figure, the main module performs the online non-negative decomposition of the incoming audio stream into note events, and outputs their respective activations. These note events are provided by the secondary module as a dictionary of note templates, which is learned offline prior to the decomposition by using non-negative matrix factorization, as shown on the left side of the figure. We describe the two general modules hereafter.

The offline learning module aims at building a dictionary matrix $\mathbf{A}$ of basis vectors that represent characteristic and discriminative templates of the r note events to detect. We suppose that the user possesses isolated sound exemplars of the r note events, from which the system learns the desired templates. For each sound exemplar, we learn exactly one template as follows. We first compute a short-time sound feature of dimension m to represent the sound exemplar, such as a short-term magnitude or power frequency spectrum.[2] These representations along time provide the non-negative observation matrix $\mathbf{Y}^{(k)}$, where each column $\mathbf{y}_j^{(k)}$ is the representation of the k-th sound exemplar at the j-th time frame. We then solve a simple NMF problem $\mathbf{Y}^{(k)} \approx \mathbf{a}^{(k)}\mathbf{x}^{(k)\top}$, with a rank of factorization equal to 1, and a normalization of the activations along time. This learning scheme simply gives one note event template $\mathbf{a}^{(k)}$, while the information from the activations $\mathbf{x}^{(k)\top}$ is discarded.

[2] For the framework of non-negative matrix factorization to apply, we suppose that the short-time sound representation is non-negative and approximatively additive, since both non-negativity and additivity cannot be obtained together.

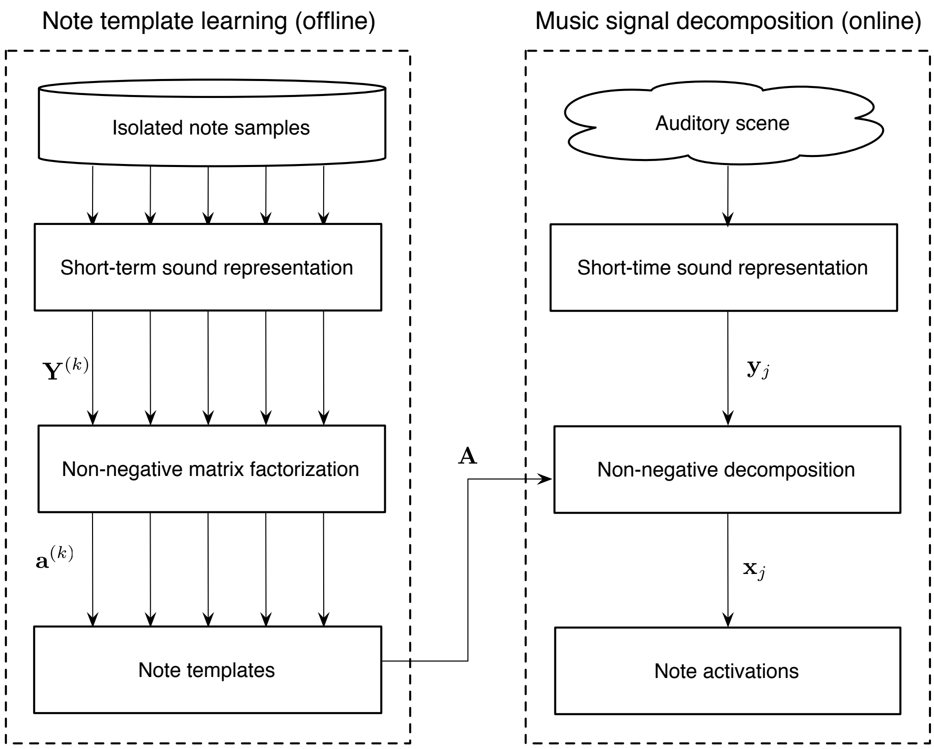

Figure 5.2.: Architecture of the proposed real-time system. The music signal arrives online to the system, and is decomposed onto note events whose descriptions are provided as a dictionary of note templates learned offline prior to the decomposition.

Having learned one template for each sound exemplar, we construct the dictionary matrix $\mathbf{A}$ by stacking the r note templates in columns. The problem of real-time decomposition of an audio stream then amounts to projecting the incoming signal $\mathbf{y}_j$ onto $\mathbf{A}$, where $\mathbf{y}_j$ share the same representational front-end as the templates. The problem is thus equivalent to a non-negative decomposition $\mathbf{y}_j \approx \mathbf{A}\mathbf{x}_j$, where the observation vector $\mathbf{y}_j$ is of length m, the dictionary matrix $\mathbf{A}$ is of size $m \times r$ and is kept fixed, and the encoding vector $\mathbf{x}_j$ is of length r and is learned online. The learned encoding vectors $\mathbf{x}_j$ then provide the activations of the different note events potentially present in the polyphonic music signal along time. To simplify the notations in the sequel, we restrict without loss of generality to the case where there is only one vector $\mathbf{y}$ to decompose as $\mathbf{y} \approx \mathbf{A}\mathbf{x}$.

As such, the system reports only a frame-level activity of the different note events. Depending on the final application, some post-processing is thus needed to extract more information about the presence of each note event at the frame level or at a longer-term level. For example, we can simply threshold the activations in order to assess the presence of note events at a given time frame, or use more elaborate methods for onset detection. To extract some information at the note level, we can further smooth the activations or model their temporal evolution.

In preliminary experiments, we tried several post-processing methods on the activations, such as thresholding, median filtering, minimum duration pruning, temporal smoothing with a Hidden Markov model. The best detection results were obtained when coupling the thresholding and pruning heuristics, which are also quite inexpensive from a computational perspective, and thus suitable to our real-time constraints. For the sake of conciseness, we therefore discuss only these two types of post-processing in the sequel.

Concerning non-negative decomposition, the similar systems in the literature consider either the Euclidean distance or the Kullback-Leibler divergence as cost functions. Moreover, there is in general no control on the decomposition, except from the system of Cont [2006], Cont et al. [2007], where the sparsity of the solutions is regularized but considered as fixed over the whole signal. As discussed previously, we tried in prior work several approaches to control sparsity more flexibly, but it did not succeed to improve systematically the detection of note events in polyphonic music, where the thresholding heuristics were shown to perform similarly. In the sequel, we rather investigate the use of a certain parametric family of information divergences as cost functions, a with flexible control on the frequency compromise during the decomposition.

5.2.2. Non-negative decomposition

In the present work, we consider the parametric family of (α, β)-divergences to construct the cost function for non-negative decomposition. To the best of our knowledge, this is the first time that these recently introduced divergences are employed in audio analysis. The (α, β)-divergences nonetheless encompass several well-known divergences as particular cases, and notably the β-divergences that have already been proved beneficial in several offline systems for audio analysis, as discussed previously. The relevance of the β-divergences for audio analysis is partly due to its interesting scaling property as noted by Févotte et al. [2009]. It appears that the (α, β)-divergences enjoy a similar scaling property which can be discussed as follows.

For any parameters $\alpha, \beta \in \mathbb{R}$, and for any positive scaling factor $\gamma > 0$, the (α, β)-divergence $d^{(ab)}_{\alpha,\beta}$ verifies the following relation:

$$d^{(ab)}_{\alpha,\beta}(\gamma y \| \gamma y') = \gamma^{\alpha+\beta} \, d^{(ab)}_{\alpha,\beta}(y \| y') \ . \tag{5.1}$$

This scaling property provides insights into the relevancy of the (α, β)-divergences in the context of audio analysis. For $\alpha + \beta = 0$, including the Itakura-Saito divergence as a special case, the (α, β)-divergence is scale-invariant. This means that the corresponding NMF problem gives the exact same relative weight to all coefficients, and thus penalizes equally a bad fit of factorization for small and large values. For other values of $\alpha+\beta$, however, the scaling property implies that a different emphasis is put on the coefficients depending on their magnitude. When $\alpha + \beta > 0$, more emphasis is put on the higher magnitude coefficients, and the emphasis augments with $\alpha + \beta$. When $\alpha+\beta < 0$, the effect is the converse. In particular, all α-divergences, including the Kullback-Leibler divergence, are such that $\alpha+\beta = 1$, and are thus homogeneous to the scale. On the contrary, the Euclidean distance responds quadratically to a rescaling and therefore considerably emphasizes the high-energy coefficients in the factorization.

Considering audio signals that are represented with a short-time frequency spectrum, this amounts to giving different importance to high and low-energy frequency components in the spectra. In the context of polyphonic music decomposition, we try to reconstruct an incoming signal by addition of note spectral templates. In order to avoid common octave and harmonic errors, a good reconstruction would have to find a compromise between focusing on the fundamental frequencies, the first partials, and the higher partials. This compromise should also be achieved in an adaptable way, independent of the fundamental frequencies, similarly to a compression rather than a global weighting of the different components. The parameters α, β, can thus help with controlling this trade-off. This interpretation comes in addition to the zooming and weighting effects of the parameters α, β, as a way to improve the robustness compared to maximum likelihood estimation, as discussed by Cichocki et al. [2011].

In the present work, we consider the right-sided non-negative decomposition problem with the (α, β)-divergences.[3] We assume that the observation vector $\mathbf{y}$ is positive, and that the dictionary matrix $\mathbf{A}$ has no null row nor column. We can then solve the non-negative decomposition by initializing the encoding vector $\mathbf{x}$ with positive values and by updating it iteratively with the multiplicative updates derived in Chapter 3.

For $\alpha \neq 0$, rewriting the updates in vector form leads to the following expression:

$$\mathbf{x} \leftarrow \mathbf{x} \otimes \left(\frac{\left(\mathbf{A}^\top \otimes \left(\mathbf{y}^{\odot\,\alpha}\ \mathbf{e}_r{}^\top\right)\right)^\top (\mathbf{A}\mathbf{x})^{\odot\,\beta-1}}{\mathbf{A}^\top (\mathbf{A}\mathbf{x})^{\odot\,\alpha+\beta-1}} \right)^{\odot\, p_{\alpha,\beta}} , \tag{5.2}$$

where the circled times $\otimes$, the fraction bar $/$, and the circled dot $\odot$, denote respectively the element-wise multiplication, division, and power, $\mathbf{e}_r$ is a vector of length r full of ones, and the exponent $p_{\alpha,\beta}$ depends on the parameter values as follows:

$$p_{\alpha,\beta} = \begin{cases} 1/(\alpha+\beta-1) & \text{if } \beta/\alpha \geq 1/\alpha \ ; \\ 1/(1-\beta) & \text{if } \beta/\alpha \leq 1/\alpha - 1 \ ; \\ 1/\alpha & \text{if } 1/\alpha - 1 \leq \beta/\alpha \leq 1/\alpha \ . \end{cases} \tag{5.3}$$

Concerning implementation, we can take advantage of the dictionary matrix $\mathbf{A}$ being fixed to tailor the multiplicative updates to real-time processing. This helps with reducing the computational cost of the updates since the matrix $\mathbf{A}^\top$ can be computed offline beforehand, and the matrix $\left(\mathbf{A}^\top \otimes \left(\mathbf{y}^{\odot\,\alpha}\ \mathbf{e}_r{}^\top\right)\right)^\top$ can be computed only once per time frame. Moreover, the vector $\mathbf{A}\mathbf{x}$ can be computed only once and exponentiated twice per iteration. In a tailored implementation, the update thus amounts to computing a maximum of three matrix-vector multiplications, one element-wise vector multiplication, one element-wise vector division, and three element-wise vector powers per iteration, as well as one additional element-wise matrix multiplication, and one vector transposed replication per time frame.

[3] Since the parametric family of (α, β)-divergences is stable under swapping the arguments, which is equivalent to swapping the parameters α and β, there is no loss of generality in considering only right-sided problems and not left-sided problems.

Now for $\alpha = 0$ and $\beta = 1$, corresponding to the dual Kullback-Leibler divergence, the non-negative decomposition can be solved with the following specific multiplicative updates in vector form:

$$\mathbf{x} \leftarrow \mathbf{x} \otimes \exp\left(\frac{\mathbf{A}^\top (\log \mathbf{y} - \log(\mathbf{A}\mathbf{x}))}{\mathbf{A}^\top \mathbf{e}_m}\right) \quad , \tag{5.4}$$

where the exponential exp is considered element-wise. We can again take advantage of the dictionary matrix $\mathbf{A}$ being fixed to implement multiplicative updates tailored to real time. Indeed, the matrix $\mathbf{A}^\top$ and the vector of row sums $\mathbf{A}^\top \mathbf{e}_m$ can be computed offline beforehand, while the vector $\log \mathbf{y}$ can be computed only once per time frame.[4] In a tailored implementation, the update thus amounts to computing a maximum of two matrix-vector multiplications, one element-wise vector multiplication, one element-wise vector division, one element-wise vector subtraction, one element-wise vector logarithm per iteration, and one element-wise vector exponential per iteration, as well as one additional element-wise vector logarithm per frame.

For $\alpha = 0$ and $\beta \neq 1$, we do not have yet such multiplicative updates. The two above updates ensure monotonic decrease, and thus convergence of the cost function, as soon as $\alpha \neq 0$ or $\beta = 1$. This guarantees at least a certain quality of the output solutions. Nevertheless, they a priori do not the ensure convergence of the output solutions, even if we observed this in practice in general. Moreover, the solution vector $\mathbf{x}$ at each time frame can be directly initialized with the output solution of the previous frame to speed up the convergence of the cost function. The only restrictive assumption we made here for deriving the updates is that the observation vector $\mathbf{y}$ is positive, which is in general the case for real-world audio signals, except in certain chunks where there is a null entry gain, or for simplistic signals with sinusoidal components and no noise. To guarantee robustness in all cases, we can simply perform a pre-whitening of the frequency spectra, or set the zero coefficients to small values $\varepsilon > 0$.

5.3. Experimental results

In this section, we report a methodological series of experiments to assess the relevancy of the proposed approach. We begin with a sample example of piano music to illustrate the devised system, and to provide qualitative insights into the effect of the parametric (α, β)-divergences. This highlights that a region of parameters α, β, is of particular interest for polyphonic music transcription. This region is thus studied through a quantitative evaluation in a standard evaluation framework, for the task of multiple pitch estimation at the frame level. We select and fix the best couple of values for parameters α, β, in terms of transcription quality. We then evaluate quantitatively the system on a second task, for multiple pitch tracking at the note level. In the two evaluations, the proposed real-time system is shown to outperform two state-of-the-art offline systems.

[4] We can consider alternatives by either separating or grouping the difference of logarithms, which may become interesting depending on the number of iterations compared to the dimensions of the data, and to the implementations of the elementary operations.

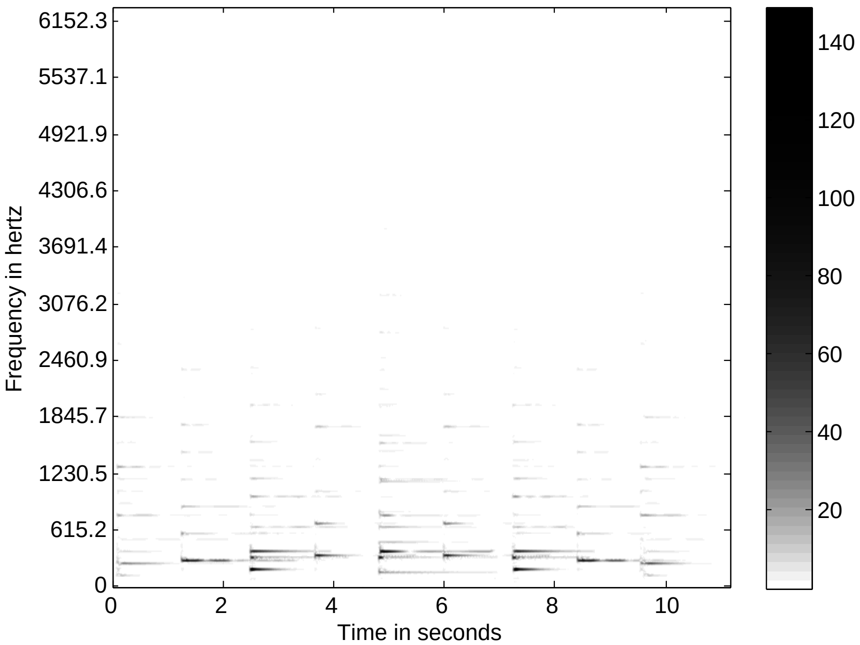

Figure 5.3.: Spectrogram of the piano excerpt. The excerpt is composed of the notes C-D-E-F-G-F-E-D-C played successively on the right hand, while the left hand plays the notes C-G-E-G-C at half the speed.

5.3.1. Sample example of piano music

We consider here a simple piano music signal as a sample example to illustrate the proposed approach and to assess the effect of the design parameters α, β. We synthesized a short piano excerpt whose spectrogram is shown in Figure 5.3. The audio synthesis was done by using real piano samples from the MIDI-Aligned Piano Sounds (MAPS) database [Emiya et al. 2010].

We also learned one spectral template for each of the $r = 88$ notes of the piano, using the respective isolated samples from MAPS. As a representation front-end, we considered a simple short-time magnitude spectrum computed with a Fourier transform, using half-overlapping frames of 1024 samples windowed with a Hamming function at a sampling rate of 12600 Hz, leading to observations of $m = 513$ dimensions after removing the redundant negative frequency bins. We employed the standard Euclidean NMF for learning the respective templates, and 10 iterations of the multiplicative updates were sufficient to stabilize the learned factors. The learned dictionary matrix $\mathbf{A}$ containing the different spectral templates is represented in Figure 5.4, where relatively discriminative and characteristic spectral templates of the respective note pitches have been learned successfully.

For the online decomposition of the successive frequency spectra from the piano excerpt onto the learned dictionary of note spectral templates, we refined the hop size to 126 samples to obtain an analysis step of 10 ms. We considered values of the divergence parameters in a standard range $\alpha, \beta \in \{-1, 0, 1, 2\}$, leading to 16 different couples of values. We chose these values to include some well-known α-divergences and β-divergences, namely, the Euclidean distance, the Kullback-Leibler and dual Kullback-Leibler divergences, the Itakura-Saito divergence, the Pearson's and Ney-

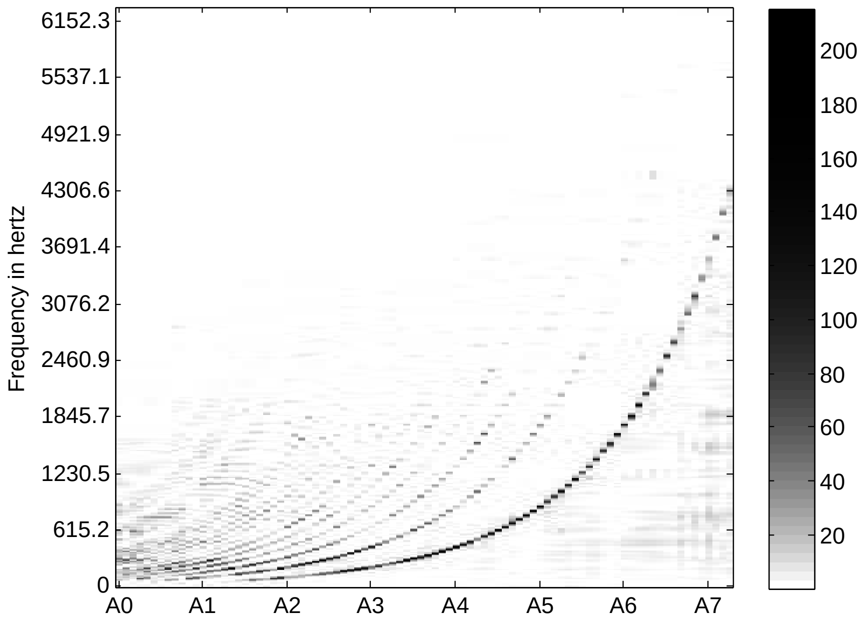

Figure 5.4.: Spectral templates of the piano notes. Characteristic and discriminative spectral templates are learned for each note on the piano, by applying non-negative matrix factorization on the corresponding isolated samples.

man's χ^2 distances. The number of iterations was set to 100 for decomposition so as to ensure the stabilization of the updates, even if we observed in practice that 10 to 20 iterations, and even less during note sustains due to the initialization with previous activations, are in general sufficient to obtain relatively good activations.

The results of the decomposition for the different couples (α, β) are shown in Figure 5.5 in terms of activations of the respective note spectral templates along time. There are actually three couples in the figure for which the multiplicative updates do not apply, and which are nonetheless left with empty activations for better visualization. Overall, this demonstrates that the system has succeeded in matching the note templates to the incoming piano signal. We notice however that for $\alpha + \beta = 2$, such as for the Euclidean distance, the system tends to use more templates than present notes, in particular templates from octaves and other harmonics of the actual notes. This effect diminishes and the activations get cleaner for $\alpha + \beta = 1$, corresponding to the α-divergences, and in particular here to the Kullback-Leibler and dual Kullback-Leibler divergences, as well as the Pearson's and Neyman's χ^2 distances. For $\alpha + \beta = 0$, such as for the Itakura-Saito divergence, the activations get sparser, and even the present notes are sometimes activated less strongly. The same effect also seems to appear for small values of α.

These results prove that a compromise is to be found between the different frequency components when decomposing the spectral distributions. We recall that for $\alpha + \beta = 2$, the scaling of the observations is quadratic, while it is linear for $\alpha + \beta = 1$. As a result, the partials are emphasized more when $\alpha + \beta$ decreases. For $\alpha + \beta = 0$, they are equally weighted compared to the fundamental frequencies. A good frequency trade-off seems to be obtained for a scale dependence between linearity and invariance. We thus focus in the sequel on the parameter region $0 \leq \alpha + \beta \leq 1$.

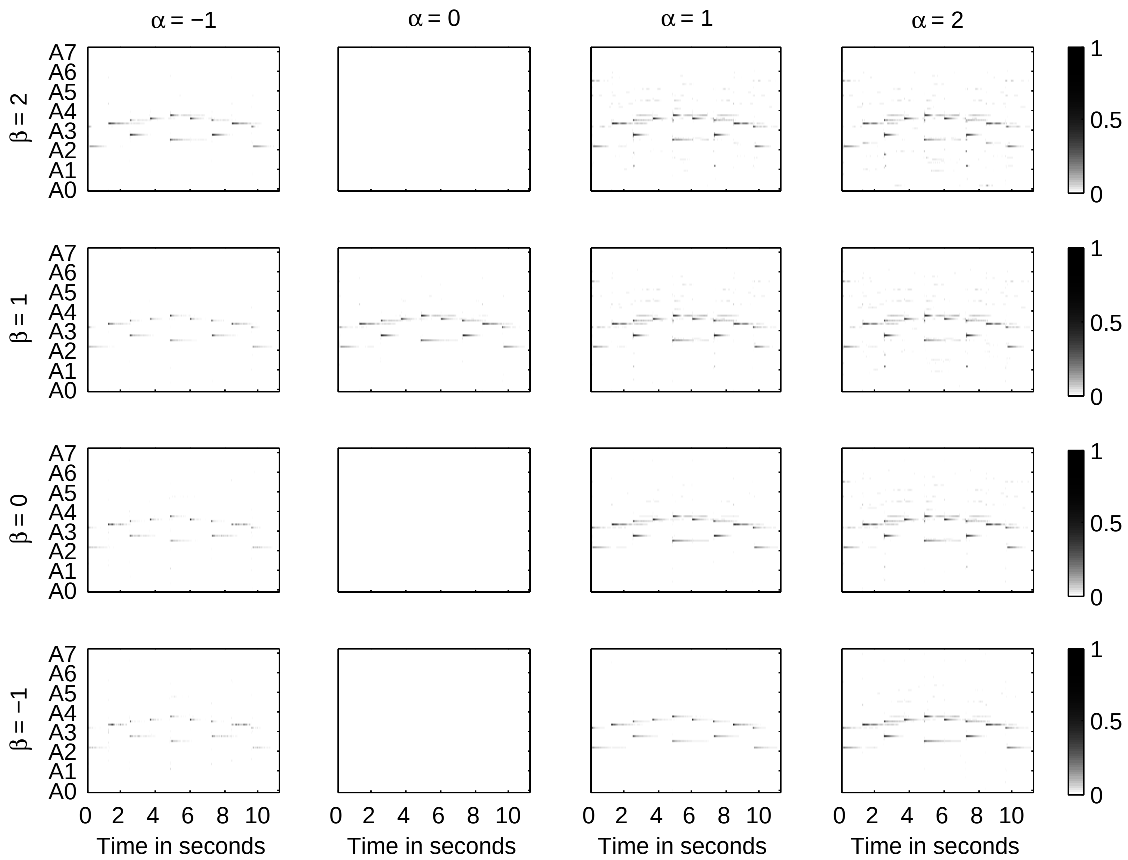

Figure 5.5.: Activations of the spectral templates. The activations are cleaner for $0 \leq \alpha + \beta \leq 1$, corresponding to a scale dependence between linearity and invariance, than for $\alpha+\beta = 2$, corresponding to a quadratic dependence.

5.3.2. Evaluation on multiple fundamental frequency estimation

We now evaluate quantitatively the proposed approach in a framework with widely accepted evaluation metrics, namely the Music Information Retrieval Evaluation eXchange (MIREX) [Bay et al. 2009]. We first consider the task of multiple pitch estimation at the frame level according to the MIREX metrics.

For the evaluation dataset, we considered the MAPS database as above. In addition to isolated samples of piano notes, MAPS contains real audio recordings of piano pieces with ground-truth references given as MIDI files. We selected 25 pieces recorded with the Yamaha Disklavier Mark III in close recording conditions, and truncated each of these pieces to 30 s.

For the dictionary matrix, one spectral template was learned for each note of the piano, from an audio fragment created by concatenating the three corresponding samples in MAPS at dynamics piano, mezzo-forte and forte. We employed the exact same analysis parameters as above.

For online decomposition, we refined the hop size to 10 ms and set the number of iterations to 100. We focused on the parameter region $0 \leq \alpha + \beta \leq 1$, and sampled it with a step of 0.5, for both $-1 \leq \alpha \leq 5$, and $-5 \leq \beta \leq 2$, leading to 39 different couples of parameter values. These values notably include the Kullback-Leibler and dual Kullback-Leibler divergences, the Hellinger distance, the Itakura-Saito divergence, the Neyman's and Pearson's χ^2 distances, as well as the β-divergence for $\beta = 0.5$ which was shown optimal among all β-divergences in experiments on polyphonic music transcription by Vincent et al. [2010].

The output activations of the algorithms were all post-processed with a simple threshold λ to detect the presence of note events at the frame level. The threshold was tuned for each algorithm with a step of 0.001 in the range $0.001 \leq \lambda \leq 0.1$, so as to achieve optimal results over the database. We did not use any further post-processing at this point in order to really compare the quality of the activations output by the different algorithms at the frame level.

The performance of the algorithms was measured through the F-measure $\mathcal{F}$ computed as the harmonic mean of the precision and recall, where the precision $\mathcal{P}$ is the percentage of correctly detected note events over the total number of detected note events, and thus quantifies the robustness of the algorithm in relation to true negatives and false positives, while the recall $\mathcal{R}$ is the percentage of correctly detected note events over the total number of ground-truth note events, and thus quantifies the efficiency of the algorithm in relation to true positives and false negatives.

We report the evaluation results per algorithm in Table 5.1. There are actually two couples of parameter values for which the multiplicative updates do not apply, and which are nonetheless left with empty results for better visualization. Overall, the results show that the proposed system and algorithm perform relatively well, with a maximum F-measure $\mathcal{F} = 70.39$. The best results are clearly obtained for $\alpha + \beta = 0.5$. This corroborates that we need to find a compromise between the different frequency components in the decomposition, and that this trade-off is optimal for a scale dependence between linearity with $\alpha+\beta = 1$, and invariance with $\alpha+\beta = 0$. Interestingly, the optimal results are not obtained for a α-divergence nor for a β-divergence, but for a (α, β)-divergence with $\alpha = 2.5$, $\beta = -2$.

We also notice that the optimal threshold λ decreases as the parameter α decreases on the lines of constant scale dependence, which corroborates previous remarks. For the three such lines, the optimal threshold is around $\lambda = 0.022$. We believe that this effect is due to a necessary rescaling of the threshold to adapt it to the normalization of the spectral templates. Indeed, the spectral templates were computed and normalized to obtain unitary maximum activations for the Euclidean distance during learning. In prior experiments, we tried to learn and normalize the spectral templates directly according to the divergence used for decomposition. Nevertheless, it did not improve systematically the results compared to a simple learning and normalization with the Euclidean distance. Further considerations are needed on this line to provide insights into this effect.

Last but not least, for the linear scale dependence corresponding to α-divergences, we notice that the best results are obtained for the Hellinger distance which is widely used in statistical inference. For the other scale dependence, the best results are however obtained for statistical divergences outside the range of common distance measures. This highlights that other statistical divergences than the standard ones may be beneficial in some applications, and confirms the interest in using parametric families of divergences with this in prospect.

To compare the results of the proposed scheme (`ABND`), we also performed the

$\alpha+\beta$	α	β	λ	$\mathcal{F}$	Distance function
0.0	-1.0	$+1.0$	0.007	55.62	
	-0.5	$+0.5$	0.008	60.13	
	∓ 0.0	± 0.0			
	$+0.5$	-0.5	0.011	64.00	
	$+1.0$	-1.0	0.013	64.92	Itakura-Saito
	$+1.5$	-1.5	0.015	65.67	
	$+2.0$	-2.0	0.016	66.19	
	$+2.5$	-2.5	0.018	66.51	
	$+3.0$	-3.0	0.019	66.75	
	$+3.5$	-3.5	0.021	66.86	
	$+4.0$	-4.0	0.022	**66.94**	
	$+4.5$	-4.5	0.023	66.91	
	$+5.0$	-5.0	0.024	66.87	
0.5	-1.0	$+1.5$	0.009	61.13	
	-0.5	$+1.0$	0.011	65.52	
	∓ 0.0	$+0.5$			
	$+0.5$	± 0.0	0.015	69.19	
	$+1.0$	-0.5	0.017	69.92	β-divergence with $\beta = 0.5$
	$+1.5$	-1.0	0.018	70.27	
	$+2.0$	-1.5	0.020	70.37	
	$+2.5$	-2.0	0.022	**70.39**	
	$+3.0$	-2.5	0.023	70.35	
	$+3.5$	-3.0	0.025	70.27	
	$+4.0$	-3.5	0.026	70.17	
	$+4.5$	-4.0	0.027	70.04	
	$+5.0$	-4.5	0.028	69.92	
1.0	-1.0	$+2.0$	0.013	62.89	Neyman's χ^2
	-0.5	$+1.5$	0.016	65.76	α-divergence with $\alpha = -0.5$
	∓ 0.0	$+1.0$	0.018	66.92	Dual Kullback-Leibler
	$+0.5$	$+0.5$	0.021	**67.19**	Hellinger
	$+1.0$	± 0.0	0.023	67.19	Kullback-Leibler
	$+1.5$	-0.5	0.024	67.09	α-divergence with $\alpha = 1.5$
	$+2.0$	-1.0	0.026	66.94	Pearson's χ^2
	$+2.5$	-1.5	0.028	66.78	α-divergence with $\alpha = 2.5$
	$+3.0$	-2.0	0.028	66.58	α-divergence with $\alpha = 3$
	$+3.5$	-2.5	0.030	66.37	α-divergence with $\alpha = 3.5$
	$+4.0$	-3.0	0.031	66.19	α-divergence with $\alpha = 4$
	$+4.5$	-3.5	0.031	66.00	α-divergence with $\alpha = 4.5$
	$+5.0$	-4.0	0.032	65.78	α-divergence with $\alpha = 5$

Table 5.1.: Evaluation results for multiple fundamental frequency estimation. The results show that the proposed system and algorithm perform relatively well, with a maximum F-measure $\mathcal{F} = 70.39$, obtained for $\alpha = 2.5$, $\beta = -2$, corresponding to a scale dependence between linearity and invariance.

Algorithm	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{A}$	$\mathcal{E}_{\text{sub}}$	$\mathcal{E}_{\text{mis}}$	$\mathcal{E}_{\text{fal}}$	$\mathcal{E}_{\text{tot}}$
`ABND`	67.23	73.85	**70.39**	**54.31**	6.24	19.91	29.76	**55.91**
`BHNMF`	61.00	66.74	63.74	46.78	10.38	22.88	32.30	65.56
`SACS`	60.03	70.84	**64.99**	**48.13**	16.35	12.81	30.83	**59.99**

Table 5.2.: Comparative results for multiple fundamental frequency estimation. The results confirm that the proposed real-time system and algorithm perform well at the frame level, and even outperform the two offline state-of-the-art systems considered.

evaluation for two offline systems at the state-of-the-art. The first one (`BHNMF`) was developed by Vincent et al. [2010], and is based on unsupervised NMF with the β-divergence for $\beta = 0.5$, and a harmonic model with spectral smoothness to ensure a structured dictionary matrix into spectral templates that correspond to musical notes with a known pitch. The second one (`SACS`) was developed by Yeh et al. [2010], and is based on a sinusoidal analysis with a candidate selection exploiting spectral features. For the sake of completeness, we also computed complementary metrics from the MIREX framework, namely the accuracy $\mathcal{A}$, total error $\mathcal{E}_{\text{tot}}$, substitution error $\mathcal{E}_{\text{sub}}$, missed error $\mathcal{E}_{\text{mis}}$, false alarm error $\mathcal{E}_{\text{fal}}$, as defined in [Bay et al. 2009].

The comparative results are presented in Table 5.2. They confirm that the performance of the system and algorithm is relatively good and competitive with the literature. For the optimal parameter values, the online algorithm `ABND` outperforms the two offline algorithms `BHNMF` and `SACS` for the different global metrics $\mathcal{F}$, $\mathcal{A}$, $\mathcal{E}_{\text{tot}}$. Going into details of the metrics, we see that the precision $\mathcal{P}$ and recall $\mathcal{R}$ of `ABND` are both higher than that for `BHNMF` and `SACS`. It means that the good performance of `ABND` compared to the reference algorithms is not due to one algorithm detecting more ground-truth note events while also making more detection errors.

This is corroborated by the respective error metrics, where we see in general that the three complementary errors $\mathcal{E}_{\text{sub}}$, $\mathcal{E}_{\text{mis}}$, $\mathcal{E}_{\text{fal}}$, are less than for the two reference algorithms. In particular, `ABND` makes significantly less substitution errors, meaning that for one note event detected while a note event is indeed present, there is in general no error on the musical pitch detected. This is due to the control on the frequency compromise which allows a better balance of the fundamental frequencies and partials in the analysis to achieve a good detection. We also remark that the false alarm errors are relatively similar for the three algorithms, so that that they all tend to detect to many note events to the same extent. Interestingly, there are less missed note events with `SACS`, meaning that when ground-truth note events are present, associated note events are more often detected. Yet it is undermined by the fact that these detected note events may not have the correct pitch, as suggested by the substitution error.

Last but not least, the fact that `ABND` outperforms `BHNMF` is due both to the improvement of the (α, β)-divergences compared to the β-divergences, and to the note spectral templates being learned on isolated samples for `ABND` whereas they are learned directly on the respective music pieces for `BHNMF`, which is obviously harder. Nonetheless, `BHNMF` uses more elaborate pre- and post-processing to improve the transcription compared to ABND, such as using a perceptually-motivated frequency scale, or a global normalization of the activations along time which is not possible in real time.

5.3.3. Evaluation on multiple fundamental frequency tracking

We now fix the obtained optimal parameter values $\alpha = 2.5$, $\beta = -2$, and consider further experiments by focusing on the task of multiple pitch tracking at the note level according to the MIREX metrics. We employed the exact same database, dictionary matrix of note templates, and number of iterations for decomposition.

The output activations of the algorithm were post-processed with both a threshold λ, and a minimum duration pruning δ, in order to detect the presence of musical notes. More precisely, a note is detected as soon as its activations exceed the threshold λ during at least δ time frames. The note onset is then determined as the first such time frame, while the note offset is simply detected as the first time frame such that the activations fall under the threshold. A new note onset can then be detected. The threshold was tuned with a step of 0.001 in the range $0.01 \leq \lambda \leq 0.1$, so as to achieve optimal results over the database. The minimum duration pruning was tested for the values $\delta \in \{1, 2, 3, 4, 5\}$.

The performance of the algorithm was measured through the F-measure $\mathcal{F}_1$ computed as the harmonic mean of the precision and recall, where the precision $\mathcal{P}_1$ is the percentage of correctly detected notes over the total number of detected notes, while the recall $\mathcal{R}_1$ is the percentage of correctly detected notes over the total number of ground-truth notes. According to the MIREX metrics, a note is assumed to be correctly detected in this scenario if the detected note onset is within a time tolerance of ± 50 ms from a ground-truth note onset with the same musical pitch. To avoid indeterminacies, it is verified that no two reference notes with the same musical pitch are separated by less than 100 ms. To assess the performance in the detection of note offsets, we also computed the mean overlap ratio $\mathcal{M}_1$ defined as the average of the overlap ratios among the correctly detect notes, where the overlap ratio is the ratio between the length of the intersection and union of the temporal widths of the correctly detected and corresponding reference notes.

We report the evaluation results in Table 5.3. Overall, the results show that the proposed system and chosen algorithm also perform relatively well at the note level, with a maximum F-measure $\mathcal{F}_1 = 77.82$. The best results are obtained for a minimum duration pruning $\delta = 5$, meaning that this simple form of temporal smoothing is relatively efficient. This is confirmed by the mean overlap ratio $\mathcal{M}_1$ which augments with the minimum duration pruning, and reaches its highest value for a maximum time smoothing $\delta = 5$. Increasing the minimum duration pruning above 5 frames did not however improve the transcription results in our experiments, and would yet augment the latency of the system. Reducing the minimum duration pruning, the transcription results are still relatively good while the latency decreases, so that the user may choose a desired compromise without a too severe degradation of the results if the latency needs to be kept as low as possible.

To compare the results of ABND, we again performed the evaluation for the two offline algorithms BHNMF and SACS. We also computed additional metrics with the

$\alpha + \beta$	α	β	δ	λ	$\mathcal{F}_1$	$\mathcal{M}_1$
0.5	+2.5	−2.0	1	0.069	68.80	53.36
			2	0.065	71.84	54.19
			3	0.046	74.47	56.42
			4	0.046	76.43	56.68
			5	0.038	**77.82**	**57.15**

Table 5.3.: Evaluation results for multiple fundamental frequency tracking. The results show that the proposed system and algorithm perform relatively well, with a maximum F-measure $\mathcal{F}_1 = 77.82$, obtained for $\delta = 5$, corresponding to a temporal smoothing of 50 ms.

Algorithm	$\mathcal{P}_1$	$\mathcal{R}_1$	$\mathcal{F}_1$	$\mathcal{M}_1$	$\mathcal{P}_2$	$\mathcal{R}_2$	$\mathcal{F}_2$	$\mathcal{M}_2$
`ABND`	77.73	77.90	**77.82**	**57.15**	28.93	28.99	**28.96**	**77.08**
`BHNMF`	58.09	73.71	**64.98**	**57.66**	20.72	26.29	**23.17**	78.64
`SACS`	33.00	58.83	42.29	55.10	11.59	20.67	14.86	**82.17**

Table 5.4.: Comparative results for multiple fundamental frequency tracking. The results confirm that the proposed real-time system and algorithm perform well at the note level, and even outperform the two offline state-of-the-art systems considered.

F-measure $\mathcal{F}_2$, precision $\mathcal{P}_2$, recall $\mathcal{R}_2$, mean overlap ratio $\mathcal{M}_2$, in a second transcription scenario from the MIREX framework. In addition to the onset tolerance of the first scenario, a note is assumed to be correctly detected in this second scenario, if its offset is also within a time tolerance of either ±50 ms, or ±20% of the ground-truth note duration, whichever is larger, from the ground-truth note offset. This second scenario provides complementary information to the simple mean overlap ratio regarding the correct detection of note offsets.

The comparative results are presented in Table 5.4. They confirm that the performance of the system with the chosen algorithm is relatively good and competitive with the literature. The online algorithm `ABND` again outperforms the two offline algorithms `BHNMF` and `SACS` with respect to both F-measures $\mathcal{F}_1$, $\mathcal{F}_2$. This is corroborated by the precisions $\mathcal{P}_1, \mathcal{P}_2$, and recalls $\mathcal{R}_1, \mathcal{R}_2$, which are both higher for `ABND` than for `BHNMF` and `SACS`. We notice however that the two mean overlap ratios $\mathcal{M}_1, \mathcal{M}_2$, are not better than for the other systems. It might be because some notes that are more difficult to detect are correctly transcribed by `ABND`, and not by `BHNMF` or `SACS`. As a result, the detection of the offsets of these notes is less precise but is taken into account in the mean overlap ratios for `ABND` and not for `BHNMF` or `SACS`. Moreover, `BHNMF` uses more elaborate post-processing with a double thresholding and minimum duration pruning for both note onset and offset detection, hence improving the mean overlap ratio scores.

5.4. Discussion

In this chapter, we discussed the problem of real-time polyphonic music transcription. To address this problem, we designed a system based on non-negative decomposition, where the music signal arrives in real time to the system, and is projected onto a dictionary of note spectral templates that are learned offline prior to the decomposition. We notably investigated the parametric family of (α, β)-divergences for decomposition with tailored multiplicative updates, and interpreted their relevancy as a way to provide controls on the frequency compromise during the decomposition, while maintaining monotonic cost decrease for critical safety, and computational efficiency for real-time processing. We applied the proposed approach in a methodological series of experiments, and discussed the benefits in using such controls to improve transcription. The proposed system was notably shown to outperform two state-of-the-art offline systems for multiple pitch estimation and tracking in a standard evaluation framework. Last but not least, the system has been implemented in the Max/MSP real-time computer music environment, and is now being employed for live interaction in music creation and improvisation. These results are encouraging for further improvement of the proposed approach.

To begin with, we would like to overcome the implicit assumption that the spectral templates are stationary. The rigorous consideration of non-stationarity is likely to improve the detection of notes through a better modeling of their spectro-temporal characteristics. To address this, it is possible to consider front-end representations that capture variability over a short time span, such as the modulation spectrum used in [Cont et al. 2007]. We believe however that a more elaborate approach is necessary to account for the full temporal profiles of note spectra, by considering the temporality of the templates directly within the model. For instance, we could consider extended NMF models that explicitly deal with time-varying objects, such as those proposed recently by [Hennequin et al. 2010, 2011a, Badeau 2011]. Another potential approach is to combine NMF with a state representation of sounds through hidden Markov models as in [Mysore et al. 2010, Nakano et al. 2010b, Benetos and Dixon 2011b]. These two approaches should be investigated further.

Besides modeling the temporality of the note events, the template learning phase may also be improved. We employed here a simple rank-one NMF with the Euclidean distance. An advantage in formulating the learning phase in an NMF framework is that of the variety of extended schemes available to learn one or more templates for each note event, and thus better account for variability between different instances of the same note. As discussed previously, we tried to employ the corresponding (α, β)-divergences directly for template learning. It did not however improve systematically the results in our experience. Further considerations are also needed in this direction to understand the theoretical or practical causes of that.

In addition, we could employ other representations that the simple magnitude spectrum considered here. For polyphonic music transcription, it may be beneficial to consider non-linear frequency scales, for example, a logarithmic scale with the constant-Q transform, or perceptually-motivated frequency scales such as the Mel, Bark, or ERB scales. In a more general setup, we could also use a wavelet transform, maybe coupled with a modulation spectrum representation, to provide a multi-scale analysis of the spectro-temporal features of the sounds. The extension of NMF to tensors may further enhance the system, permitting for instance to use multi-channel information in the sound representations.

We would like also to improve the robustness of the system. A first direction is to employ more elaborate post-processing than the thresholding and minimum duration pruning used here. As discussed previously, these simple techniques, in addition to being computationally inexpensive, have yet provided good results and have outperformed, in our preliminary experiments, other smoothing techniques based on more demanding techniques such as hidden Markov models. A possibility to enhance the system in this direction, is to model the information from the encoding coefficients during template learning, so as to improve the detection during the decomposition. We could also investigate the use of non-fixed updated basis vectors to absorb noise and other undesirable sound components. Alternatively, using other robust divergences for decomposition may be a solution. Nonetheless, we tried the more general skew (α, β)-divergences and the complementary skew Jensen β-divergences in preliminary experiments, but they did not permit to improve the transcription performances compared to the (α, β)-divergences. This should be investigated further, maybe in relation to the use of other cost functions for template learning.

Finally, we did not discuss here the generalization capacities of the proposed system. In real-time contexts such as environments for live interaction in music creation or improvisation, and softwares for music tutoring or computer-aided composition, we can realistically assume that the note spectral templates are learned directly from the corresponding instrument. In other contexts, one may want to apply directly the system to transcribe general music, without available data for learning the templates. To assess the generality in such situations, we submitted a preliminary version of the proposed system to the 2010 MIREX evaluation campaign, where it was evaluated and compared to other algorithms on different tasks of polyphonic music transcription for various instruments and kinds of music. Even if the submitted system contained only simple piano note templates, and was the only real-time system, it did however performed competitively with the other systems submitted.[5] We believe nonetheless that the system could be enhanced in this direction, for example, by employing a hierarchical instrument basis as in [Grindlay and Ellis 2011], or more generally by addressing the use of adaptive templates [Le Roux et al. 2009, Lee et al. 2012], or online dictionary learning techniques [Mairal et al. 2010, Lefèvre et al. 2011b, Szabó et al. 2011]. Future work should address the adaptation of these approaches to the proposed algorithms.

[5]The results of the 2010 MIREX evaluation campaign for multiple fundamental frequency estimation and tracking are available online: http://www.music-ir.org/mirex/wiki/2010:Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results.

Conclusion

This thesis aimed at providing novel computational methods within the framework of information geometry, with a focus on real-time applications in audio signal processing. In this context, we proposed two independent algorithmic schemes that fall within the realm of statistical machine learning and signal processing. On the one hand, we developed a general and unifying framework for change detection with exponential families, and applied it to address the problem of real-time audio segmentation. On the other hand, we elaborated a generic and unifying framework for non-negative matrix factorization with convex-concave divergences, and employed it to address the problem of real-time polyphonic music transcription. In the sequel, we summarize the main contributions of the present work and discuss several potential perspectives for future work.

Summary of the present work

Hereafter, we summarize our main contributions to sequential change detection and of its application to audio segmentation. We then sum up our principal contributions to non-negative matrix factorization and of its application to polyphonic music transcription.

From sequential change detection to audio segmentation

In Chapter 2, we elaborated novel computational methods for sequential change detection with exponential families. We followed a standard non-Bayesian approach and formulated change detection as a statistical decision problem with multiple hypotheses, where the decision relies on the computation of generalized likelihood ratio test statistics. We also introduced exact generalized likelihood ratios with arbitrary estimators. Applying this to exponential families, we developed a generic scheme for change detection under common scenarios with known or unknown parameters, and arbitrary estimators, in close relation to maximum likelihood estimation. We also interpreted this scheme within the dually flat geometry of exponential families, hence providing both statistical and geometrical intuitions, and bridging the gap between statistical and distance-based approaches to change detection. We finally revisited this scheme through convex duality, and derived an attractive scheme with closed-form sequential updates for the exact generalized likelihood ratio statistics, when the parameters before and after change are both unknown, and are estimated by maximum likelihood.

In Chapter 4, we discussed the application of this scheme to the problem of real-time audio segmentation. We addressed this problem by proposing a generic and unifying framework capable of handling various types of signals and of homogeneity criteria. The proposed approach is centered around the scheme for sequential change detection with exponential families, and the audio stream is segmented as it unfolds in time by modeling the distribution of the sound features, and monitoring structural changes in these distributions. We also discussed how the framework unifies and generalizes statistical and distance-based approaches to segmentation through the dually flat geometry of exponential families. We finally showcased various applications to illustrate the generality of the proposed approach, and notably performed a quantitative evaluation for musical onset detection to demonstrate how it can leverage modeling in a complex task.

From non-negative decomposition to polyphonic music transcription

In Chapter 3, we elaborated novel computational methods for non-negative matrix factorization with convex-concave divergences. We developed a general optimization scheme based on variational bounding with auxiliary functions that works for almost arbitrary convex-concave divergences. We obtained monotonically decreasing updates under mild conditions by minimization of the auxiliary function. We also considered symmetrized and skew divergences for the cost function. In particular, we specialized the generic updates to provide updates for Csiszár divergences, certain skew Jeffreys-Bregman divergences, skew Jensen-Bregman divergences, leading to several known multiplicative updates, as well as novel multiplicative updates, for α-divergences, β-divergences, and their symmetrized or skew versions. We also generalized this by considering the family of skew (α, β, λ)-divergences.

In Chapter 5, we discussed the application of these updates for real-time polyphonic music transcription. To address this problem, we designed a system based on non-negative decomposition, where the music signal arrives in real time to the system, and is projected onto a dictionary of note spectral templates that are learned offline prior to the decomposition. We notably investigated the parametric family of (α, β)-divergences for decomposition with tailored multiplicative updates, and interpreted their relevancy as a way to provide controls on the frequency compromise during the decomposition, while maintaining monotonic cost decrease for critical safety, and computational efficiency for real-time processing. We applied the proposed approach in a methodological series of experiments, and discussed the benefits in using such controls to improve transcription. The proposed system was notably shown to outperform two state-of-the-art offline systems for multiple pitch estimation and tracking in a standard evaluation framework.

Perspectives for future work

Several interesting perspectives arose from the proposed approaches and were left out for future work. These perspectives were discussed in detail in the respective chapters. We summarize some of them hereon, focusing on the connections between the theoretical and applicative parts.

Change detection and audio segmentation

To begin with, extensions of the proposed framework for change detection with exponential families could beneficially be employed for audio segmentation. This includes the generalization to non-full families. It would provide new possibilities to model certain audio features more reliably by constraining the parameter space to encode some relevant sound structures, for example, spectral profiles or harmonicity.

The generalization of the proposed approach to address statistical dependence between the random variables would also provide more accurate models for certain audio features with complex temporal profiles and clear dependence, and thus permit the consideration of more elaborate representations. In this context, methods for automatic feature and model selection would find interests in adapting the input observations of change detection to the segmentation problem at hand.

In addition, other estimators than the maximum likelihood could also be employed to account for a priori knowledge on the signals through maximum a posteriori estimation, or to improve robustness in case of model misspecification through quasi likelihoods. More generally, a Bayesian framework may improve the detection of changes, when relevant a priori information is known or can be learned from on a training dataset. Robustness could also be addressed with more elaborate post-processing techniques in link with statistical considerations, for example, by developing relevant model selection criteria to account for small sample sizes, and to adapt the thresholding and windowing heuristics.

Non-negative matrix factorization and music transcription

Concerning non-negative matrix factorization, direct extensions could be handled in the proposed framework. For example, using convex models would permit to consider a structured dictionary with several atoms per note, or even a full hierarchical model of notes and instruments. On another line, tensors could be employed to account for multi-channel information, therefore helping the segregation of notes to improve transcription. Employing convolutive models may also leverage the modeling of the temporal profiles of notes to overcome the stationarity of the spectral templates.

Furthermore, we could employ more elaborate cost functions to provide alternative controls on the audio decomposition. With this respect, the effect of skewing on the divergences needs thorough investigations, from both theoretical and applicative standpoints, to understand their relevancy or irrelevancy for audio analysis, and more generally for pattern analysis. A complementary approach to enhance the cost functions would be to consider penalty terms. This may help the regularization of the solutions, for example, by ensuring temporal smoothness, or by imposing adequate sparsity structures.

Last but not least, alternative optimization schemes may accelerate the convergence of the algorithms and reduce the computational loads. For example, the consideration of equalization instead of minimization of the auxiliary functions is worth investigating. Conditional updates that depend on the observations and regions of the solution space are another promising direction.

General directions of investigation

In this thesis, we have developed computational methods starting either from statistical models or from information divergences. For change detection, we have succeeded in clarifying the relations between the statistical models considered and their canonical divergences, within the dually flat information geometry of exponential families. Yet such relations disappear for non-negative matrix factorization as soon as we employ convex-concave divergences that differ from the canonical ones. It was justified in the present work by robustness considerations. The other side of the coin, however, is that we loose the statistical insights in understanding why a given divergence is adequate or not. It would therefore be interesting to gain further intuitions on the relations between statistical models and divergences, besides the exponential families and the Kullback-Leibler or Bregman divergences.

On a different perspective, we believe that the two computational methods proposed in the present work may provide benefits in broader applications and domains than those discussed. In parallel, other novel and existing methods in the framework of computational information geometry, as those exposed in the introduction, could be developed and employed for addressing problems in audio signal processing. The two schemes proposed in this thesis are just teardrops in the ocean. Nonetheless, it is our hope that the presented contributions will bring interesting insights and directions for future research in the realm of audio signal processing, and more generally of machine learning and signal processing, in the emerging but yet prolific research field of computational information geometry.

Bibliography

S. A. Abdallah and M. D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. In *5th International Conference on Music Information Retrieval (ISMIR)*, pages 318–325, Barcelona, Spain, October 2004. *Cited on page* 100.

R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. Technical report, Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge, UK, October 2007. *Cited on pages* 41, 96.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966. *Cited on page* 15.

S.-i. Amari. Differential geometry of curved exponential families—curvatures and information loss. *The Annals of Statistics*, 10(2):357–385, June 1982. *Cited on pages* xxii, 16.

S.-i. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*, J. Berger, S. Fienberg, J. Gani, K. Krickeberg, I. Olkin, and B. Singer, editors. Springer, 1985. *Cited on page* xxii.

S.-i. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995. *Cited on page* xxiii.

S.-i. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10 (2):251–276, February 1998. *Cited on page* xxiii.

S.-i. Amari. Superefficiency in blind source separation. *IEEE Transactions on Signal Processing*, 47(4):936–944, April 1999. *Cited on page* xxiii.

S.-i. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, July 2001. *Cited on page* xxiii.

S.-i. Amari. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, November 2009. *Cited on page* 16.

S.-i. Amari and J.-F. Cardoso. Blind source separation—semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, November 1997. *Cited on page* xxiii.

S.-i. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, March 2010. *Cited on page* xxii.

S.-i. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*, S. Kobayashi and M. Takesaki, editors. American Mathematical Society, Providence, RI, USA, 2000. *Cited on pages* xxii, 9.

S.-i. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao. *Differential Geometry in Statistical Inference*, volume 10 of *Lecture Notes–Monograph Series*, S. S. Gupta, editor. Institute of Mathematical Statistics, Hayward, CA, USA, 1987. *Cited on page* xxii.

S.-i. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, 3(2):260–271, March 1992. *Cited on page* xxiii.

R. André-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1):29–40, January 1988. *Cited on pages* 24, 41, 78, 95.

X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, February 2012. *Cited on page* 78.

K. A. Arwini and C. T. J. Dodson. *Information Geometry: Near Randomness and Near Independence*, volume 1953 of *Lecture Notes in Mathematics*, J.-M. Morel, F. Takens, and B. Teissier, editors. Springer, Berlin/Heidelberg, Germany, 2008. *Cited on page* xxii.

R. Badeau. Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF). In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 253–256, New Paltz, NY, USA, October 2011. *Cited on page* 118.

R. Badeau, N. Bertin, and E. Vincent. Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(12):1869–1881, December 2010. *Cited on page* 73.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, October 2005. *Cited on page* xxiii.

O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*, R. A. Bradley, D. G. Kendall, J. S. Hunter, and G. S. Watson, editors. Wiley Series in Probability and Mathematical Statistics. Wiley, Chichester, UK, 1978. *Cited on pages* 4, 41, 46.

O. E. Barndorff-Nielsen. Likelihood and observed geometries. *The Annals of Statistics*, 14(3):856–873, September 1986. *Cited on page* xxii.

O. E. Barndorff-Nielsen. Differential geometry and statistics: Some mathematical aspects. *Indian Journal of Mathematics*, 29(3):335–350, 1987. *Cited on page* xxii.

O. E. Barndorff-Nielsen and P. E. Jupp. Yokes and symplectic structures. *Journal of Statistical Planning and Inference*, 63(2):133–146, October 1997. *Cited on page* xxii.

M. Basseville. Detecting changes in signals and systems—A survey. *Automatica*, 24 (3):309–326, May 1988. *Cited on page* 23.

M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, December 1989. *Cited on pages* xxiv, 15.

M. Basseville. Divergence measures for statistical data processing—An annotated bibliography. *Signal Processing*, 93(4):621–633, April 2013. *Cited on pages* xxiv, 15.

M. Basseville and A. Benveniste. Design and comparative study of some sequential jump detection algorithms for digital signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(3):521–535, June 1983a. *Cited on pages* 24, 78.

M. Basseville and A. Benveniste. Sequential detection of abrupt changes in spectral characteristics of digital signals. *IEEE Transactions on Information Theory*, 29 (5):709–724, September 1983b. *Cited on pages* 23, 78.

M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. *Cited on pages* 22, 81.

M. Basseville, I. V. Nikiforov, and A. G. Tartakovsky. *Sequential Analysis: Hypothesis Testing and Change-Point Detection*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL, USA, 2013. *Cited on page* 23.

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, September 1998. *Cited on pages* xxiii, 17, 41, 47.

A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*, volume 120 of *Monographs on Statistics and Applied Probability*, F. Bunea, V. Isham, N. Keiding, T. Louis, R. L. Smith, and H. Tong, editors. Chapman & Hall/CRC, Boca Raton, FL, USA, 2011. *Cited on pages* 41, 47.

M. Bay, A. F. Ehmann, and J. S. Downie. Evaluation of multiple-F0 estimation and tracking systems. In *10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 315–320, Kobe, Japan, October 2009. *Cited on pages* 112, 115.

J. P. Bello and M. B. Sandler. Phase-based note onset detection for music signals. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 441–444, Hong Kong, China, April 2003. *Cited on page* 79.

J. P. Bello, C. Duxbury, M. Davies, and M. B. Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556, June 2004. *Cited on page* 79.

J. P. Bello, L. Daudet, S. A. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, September 2005. *Cited on page* 79.

E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a convolutive probabilistic model. In *8th Sound and Music Computing Conference (SMC)*, pages 19–24, Padova, Italy, July 2011a. *Cited on page* 101.

E. Benetos and S. Dixon. A temporally-constrained convolutive probabilistic model for pitch detection. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 133–136, New Paltz, NY, USA, October 2011b. *Cited on page* 118.

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, September 2007. *Cited on page* 45.

N. Bertin, C. Févotte, and R. Badeau. A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1545–1548, Taipei, Taiwan, April 2009. *Cited on page* 102.

N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3): 538–549, March 2010. *Cited on pages* 45, 102.

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35(1):99–109110, 1943. *Cited on page* 18.

S. Böck, F. Krebs, and M. Schedl. Evaluating the online capabilities of onset detection methods. In *13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 49–54, Porto, Portugal, October 2012. *Cited on page* 94.

J.-D. Boissonnat, F. Nielsen, and R. Nock. Bregman Voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281–307, September 2010. *Cited on page* xxiv.

J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, and C. J. Wellekens. A speaker tracking system based on speaker turn detection for NIST evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1177–1180, Istanbul, Turkey, June 2000. *Cited on page* 80.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA, USA, 1990. *Cited on page* 80.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. *Cited on pages* 11, 16.

B. E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change-Point Problems*, volume 243 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, Netherlands, 1993. *Cited on page* 23.

L. D. Broemeling and H. Tsurumi. *Econometrics and Structural Change*, volume 74 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York, NY, USA, 1987. *Cited on page* 24.

M. Broniatowski and A. Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, January 2009. *Cited on pages* 41, 47.

L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9 of *Lecture Notes–Monograph Series*, S. S. Gupta, editor. Institute of Mathematical Statistics, Hayward, CA, USA, 1986. *Cited on pages* 4, 41, 46.

J. Burbea and C. R. Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, 28(3):489–495, May 1982a. *Cited on page* 18.

J. Burbea and C. R. Rao. On the convexity of higher order Jensen differences based on entropy functions. *IEEE Transactions on Information Theory*, 28(6):961–963, November 1982b. *Cited on page* 18.

S. Canu and A. Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7–9):714–720, March 2006. *Cited on pages* 24, 87.

K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero. FINE: Fisher information nonparametric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2093–2098, November 2009. *Cited on page* xxiii.

K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero. Information-geometric dimensionality reduction. *IEEE Signal Processing Magazine*, 28(2):89–99, 2011. *Cited on page* xxiii.

L. Cayton. Fast nearest neighbor retrieval for Bregman divergences. In A. McCallum and S. Roweis, editors, *25th International Conference on Machine Learning (ICML)*, pages 112–119, Helsinki, Finland, July 2008. *Cited on page* xxiv.

L. Cayton. Efficient Bregman range search. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 22, pages 243–251. NIPS Foundation, La Jolla, CA, USA, 2009. *Cited on page* xxiv.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009: 17 pages, 2009. *Cited on page* 45.

A. Cena and G. Pistone. Exponential statistical manifold. *Annals of the Institute of Statistical Mathematics*, 59(1):27–56, March 2007. *Cited on page* xxii.

M. Cettolo and M. Vescovi. Efficient audio segmentation algorithms based on the BIC. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 6, pages 537–540, Hong Kong, China, April 2003. *Cited on page* 80.

J. Chen and A. K. Gupta. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Birkhäuser, Boston, MA, USA, second edition, 2012. *Cited on page* 22.

J. Chen, A. K. Gupta, and P. Jianmin. Information criterion and change point problem for regular models. *Sankhyā: The Indian Journal of Statistics*, 68(2): 252–282, May 2006. *Cited on page* 96.

S. S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, Lansdowne, VA, USA, February 1998. *Cited on page* 80.

C.-C. Cheng, D. J. Hu, and L. K. Saul. Nonnegative matrix factorization for real time musical analysis and sight-reading evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2017–2020, Las Vegas, NV, USA, March/April 2008. *Cited on page* 102.

N. N. Chentsov. A systematic theory of exponential families of probability distributions. *Theory of Probability and Its Applications*, 11(3):425–435, 1966. *Cited on page* 4.

N. N. Chentsov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*, L. J. Leifman, editor. American Mathematical Society, Providence, RI, USA, 1982. *Cited on page* xxii.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, December 1952. *Cited on page* 16.

A. Cichocki and S.-i. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, June 2010. *Cited on page* 15.

A. Cichocki, R. Zdunek, and S.-i. Amari. Csiszár's divergences for non-negative matrix factorization: Family of new algorithms. In J. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation: 6th International Conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Proceedings*, volume 3889 of *Lecture Notes in Computer Science*, pages 32–39. Springer, Berlin/Heidelberg, Germany, 2006. *Cited on page* 47.

A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9):1433–1440, July 2008. *Cited on pages* 47, 59.

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*. Wiley, Chichester, UK, 2009. *Cited on page* 45.

A. Cichocki, S. Cruces, and S.-i. Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, January 2011. *Cited on pages* 19, 20, 47, 67, 68, 73, 104, 108.

M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 617–624. MIT Press, Cambridge, MA, USA, 2002. *Cited on pages* xxiii, 46.

A. Cont. Realtime multiple pitch observation using sparse non-negative constraints. In *7th International Conference on Music Information Retrieval (ISMIR)*, pages 206–211, Victoria, Canada, October 2006. *Cited on pages* 102, 103, 107.

A. Cont and A. Dessein. Applications de la géométrie de l'information au traitement des flux audio temps réel. In *5e Biennale Française des Mathématiques Appliquées, Congrès de la Société de Mathématiques Appliquées et Industrielles (SMAI)*, Guidel, France, May 2011. *Cited on page* xxx.

A. Cont, S. Dubnov, and D. Wessel. Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints. In *10th International Conference on Digital Audio Effects (DAFx)*, pages 85–92, Bordeaux, France, September 2007. *Cited on pages* 102, 103, 107, 118.

A. Cont, S. Dubnov, and G. Assayag. On the information geometry of audio streams with applications to similarity computing. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):837–846, May 2011. *Cited on page* 80.

F. Critchley, P. Marriott, and M. Salmon. Preferred point geometry and statistical manifolds. *The Annals of Statistics*, 21(3):1197–1224, September 1993. *Cited on page* xxii.

I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tudományos Akadémia Matematikai Kutató Intézeténk Közleményei*, 8:85–108, 1963. *Cited on page* 15.

I. Csiszár. Information measures: A critical survey. In J. Kozesnik, editor, *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians held at Prague, from August 18 to 23, 1974*, volume B, pages 73–86. D. Reidel, Dordrecht, Netherlands, 1978. *Cited on page* 15.

I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3): 261–273, September 2008. *Cited on page* 15.

I. Csiszár and F. Matúš. Closures of exponential families. *The Annals of Probability*, 33(2):582–600, March 2005. *Cited on page* 41.

M. Csörgő and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley Series in Probability and Statistics. Wiley, Chichester, UK, 1997. *Cited on page* 23.

G. Darmois. Sur les lois de probabilités à estimation exhaustive. In *Comptes Rendus des Séances Hebdomadaires de l'Académie des Sciences*, volume 200, pages 1265–1266. Gauthier-Villars, Paris, France, 1935. *Cited on page* 3.

M. Davy and S. Godsill. Detection of abrupt spectral changes using support vector machines. An application to audio signal segmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1313–1316, Orlando, FL, USA, May 2002. *Cited on page* 81.

A. de Cheveigné. Multiple $F0$ estimation. In D. Wang and G. J. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, chapter 2, pages 45–79. Wiley-IEEE Press, Hoboken, NJ, USA, 2006. *Cited on page* 100.

P. Delacourt and C. J. Wellekens. DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1–2):111–126, September 2000. *Cited on page* 80.

J. Deshayes and D. Picard. Off-line statistical analysis of change-point models using non parametric and likelihood methods. In M. Basseville and A. Benveniste, editors, *Detection of Abrupt Changes in Signals and Dynamical Systems*, volume 77 of *Lecture Notes in Control and Information Sciences*, pages 103–168. Springer, Berlin/Heidelberg, Germany, 1986. *Cited on page* 96.

F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, August 2005. *Cited on pages* 24, 81.

A. Dessein. Introduction to the tools of information geometry for the manipulation of audio streams. In *Séminaire Mathématiques, Musique et relations avec d'autres disciplines (MaMuX)*, Institut de Recherche et Coordination Acoustique/Musique, Paris, France, October 2009a. *Cited on page* xxix.

A. Dessein. La géométrie de l'information pour l'analyse et le traitement des flux audio. In *5es Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio (JJCAAS)*, Marseille, France, November 2009b. *Cited on page* xxix.

A. Dessein. Some applications of non-negative matrix factorization and of information geometry in audio signal processing. In *Research Seminar*, RIKEN Brain Science Institute, Tokyo, Japan, October 2010a. *Cited on page* xxix.

A. Dessein. Information geometry for real-time processing of audio signals. In *Research Seminar*, Japanese-French Laboratory for Informatics, Tokyo, Japan, November 2010b. *Cited on page* xxix.

A. Dessein. Music information geometry. In *Journée interdisciplinaire Mathématiques et Musique (JMM)*, Institut de Recherche Mathématique Avancée, Strasbourg, France, April 2011. *Cited on page* xxx.

A. Dessein. Computational information geometry for audio signal processing. In *Symposium on Optimal Transport and Information Geometry*, Institut Henri Poincaré, Paris, France, February 2012. *Cited on page* xxx.

A. Dessein and A. Cont. Segmentation statistique de flux audio en temps-réel dans le cadre de la géométrie de l'information. In *23e Colloque du Groupe de Recherche et d'Études du Traitement du Signal (GRETSI)*, Bordeaux, France, September 2011a. *Cited on page* xxvii.

A. Dessein and A. Cont. Applications of information geometry to audio signal processing. In *14th International Conference on Digital Audio Effects (DAFx)*, Paris, France, September 2011b. *Cited on page* xxix.

A. Dessein and A. Cont. Applications of information geometry to audio signal processing. In *Indo-French Workshop on Matrix Information Geometry (MIG)*, pages 16–17, École Polytechnique, Palaiseau, France, February 2011c. *Cited on page* xxx.

A. Dessein and A. Cont. An information-geometric approach to real-time audio segmentation. *IEEE Signal Processing Letters*, 20(4):331–334, April 2013a. *Cited on page* xxvii.

A. Dessein and A. Cont. Online change detection in exponential families with unknown parameters before and after change. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information: First International Conference, GSI 2013, Paris, France, August 28-30, 2013, Proceedings*, volume 8085 of *Lecture Notes in Computer Science*, pages 633–640. Springer, Berlin/Heidelberg, Germany, 2013b. *Cited on page* xxviii.

A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 489–494, Utrecht, Netherlands, August 2010a. *Cited on page* xxviii.

A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *6th Music Information Retrieval Evaluation eXchange (MIREX)*, Utrecht, Netherlands, August 2010b. *Cited on page* xxviii.

A. Dessein, A. Cont, and G. Lemaitre. Real-time detection of overlapping sound events with non-negative matrix factorization. In F. Nielsen and R. Bhatia, editors, *Matrix Information Geometry*, chapter 14, pages 341–371. Springer, Berlin/Heidelberg, Germany, 2013. *Cited on page* xxviii.

I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 18, pages 283–290. MIT Press, Cambridge, MA, USA, 2006. *Cited on pages* xxiii, 47, 59, 61.

I. S. Dhillon and J. A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2008. *Cited on pages* xxiii, 51.

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 (NIPS)*, volume 24, pages 2267–2275. NIPS Foundation, La Jolla, CA, USA, 2011. *Cited on pages* 45, 102.

S. Dixon. Onset detection revisited. In *9th International Conference on Digital Audio Effects (DAFx)*, pages 133–137, Montreal, Canada, September 2006. *Cited on page* 79.

C. Duxbury, M. B. Sandler, and M. Davies. A hybrid approach to musical note onset detection. In *5th International Conference on Digital Audio Effects (DAFx)*, pages 33–38, Hamburg, Germany, September 2002. *Cited on page* 79.

C. Duxbury, J. P. Bello, M. Davies, and M. B. Sandler. A combined phase and amplitude based approach to onset detection for audio segmentation. In *4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 275–280, London, UK, April 2003. *Cited on page* 79.

B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242, November 1975. *Cited on page* xxi.

S. Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, 11(3):793–803, September 1983. *Cited on pages* xxi, 41, 47.

S. Eguchi. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Mathematical Journal*, 15(2):341–391, 1985. *Cited on page* xxii.

S. Eguchi. Geometry of minimum contrast. *Hiroshima Mathematical Journal*, 22 (3):631–647, 1992. *Cited on page* xxii.

S. Eguchi. Information divergence geometry and the application to statistical machine learning. In F. Emmert-Streib and M. Dehmer, editors, *Information Theory and Statistical Learning*, chapter 13, pages 309–332. Springer, New York, NY, USA, 2009. *Cited on pages* xxiii, 41, 47.

S. Eguchi and Y. Kano. Robustifying maximum likelihood estimation. Technical report, The Institute of Statistical Mathematics, Tokyo, Japan, 2001. *Cited on pages* xxiii, 17, 41, 47.

V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, August 2010. *Cited on page* 110.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 589–605, April 2007. *Cited on pages* 41, 95.

P. Fearnhead and Z. Liu. Efficient Bayesian analysis of multiple changepoint models with dependence across segments. *Statistics and Computing*, 21(2):217–229, 2011. *Cited on pages* 41, 96.

B. Fergani, M. Davy, and A. Houacine. Unsupervised speaker indexing using one-class support vector machines. In *14th European Signal Processing Conference (EUSIPCO)*, Florence, Italy, September 2006. *Cited on page* 81.

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, pages 266–296. IGI Global Press, Hershey, PA, USA, 2011. *Cited on pages* 45, 63, 102.

C. Févotte and A. T. Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. In *17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917, Glasgow, UK, August 2009. *Cited on page* 45.

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, September 2011. *Cited on pages* 47, 62, 71, 72, 103, 104.

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, March 2009. *Cited on pages* 45, 102, 107.

R. A. Fisher. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London: Series A, Containing Papers of a Mathematical and Physical Character*, 144(852):285–307, March 1934. *Cited on page* 3.

D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *20th IET Irish Signals and Systems Conference (ISSC)*, Dublin, Ireland, June 2009. *Cited on page* 102.

J. Foote. Visualizing music and audio using self-similarity. In J. F. Buford, S. M. Stevens, D. C. A. Bulterman, K. Jeffay, and H. Zhang, editors, *7th ACM International Conference on Multimedia (MM)*, volume 1, pages 77–80, Orlando, FL, USA, October/November 1999. *Cited on page* 78.

J. Foote. Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 452–455, New York, NY, USA, July/August 2000. *Cited on page* 77.

B. Fuentes, R. Badeau, and G. Richard. Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 401–404, Prague, Czech Republic, May 2011. *Cited on page* 101.

Y. Fujimoto and N. Murata. A modified EM algorithm for mixture models based on Bregman divergence. *Annals of the Institute of Statistical Mathematics*, 59(1): 3–25, March 2007. *Cited on page* xxiii.

V. Garcia and F. Nielsen. Simplification and hierarchical representations of mixtures of exponential families. *Signal Processing*, 90(12):3197–3212, December 2010. *Cited on page* xxiii.

E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *28th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 601–602, Salvador, Brazil, August 2005. *Cited on page* 46.

P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 1(2):325–347, April 1998. *Cited on page* xxii.

M. A. Girshick and H. Rubin. A Bayes approach to a quality control model. *The Annals of Mathematical Statistics*, 23(1):114–125, March 1952. *Cited on page* 22.

G. Grindlay and D. P. W. Ellis. Multi-voice polyphonic music transcription using eigeninstruments. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 53–56, New Paltz, NY, USA, October 2009. *Cited on page* 101.

G. Grindlay and D. P. W. Ellis. Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1159–1169, October 2011. *Cited on pages* 101, 119.

F. Gustafsson. *Adaptive Filtering and Change Detection*. Wiley, Chichester, UK, 2000. *Cited on page* 23.

Z. Harchaoui and C. Lévy-Leduc. Catching change-points with lasso. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 617–624. MIT Press, Cambridge, MA, USA, 2008. *Cited on page* 24.

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, December 2010. *Cited on page* 24.

Z. Harchaoui, F. Bach, and E. Moulines. Kernel change-point analysis. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 609–616. NIPS Foundation, La Jolla, CA, USA, 2009a. *Cited on page* 24.

Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappé. A regularized kernel-based approach to unsupervised audio segmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1665–1668, Taipei, Taiwan, April 2009b. *Cited on page* 81.

P. Harremoës and N. Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *IEEE International Symposium on Information Theory (ISIT)*, pages 566–570, Nice, France, June 2007. *Cited on page* 13.

J. Havrda and F. Charvát. Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika*, 3(1):30–35, 1967. *Cited on page* 16.

R. Hennequin, R. Badeau, and B. David. Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *13th International Conference On Digital Audio Effects (DAFx)*, pages 246–253, Graz, Austria, September 2010. *Cited on pages* 102, 118.

R. Hennequin, R. Badeau, and B. David. NMF with time-frequency activations to model nonstationary audio events. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):744–753, May 2011a. *Cited on pages* 102, 118.

R. Hennequin, R. Badeau, and B. David. Scale-invariant probabilistic latent component analysis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 129–132, New Paltz, NY, USA, October 2011b. *Cited on page* 101.

A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval (revised 2). Technical report, Communications and Signal Processing Laboratory, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, December 2002. *Cited on page* xxiii.

M. D. Hoffman, D. M. Blei, and P. R. Cook. Bayesian nonparametric matrix factorization for recorded music. In J. Fürnkranz and T. Joachims, editors, *27th International Conference on Machine Learning (ICML)*, pages 439–446, Haifa, Israel, June 2010. *Cited on page* 45.

T. Hofmann. Probabilistic latent semantic indexing. In *22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57, Berkeley, CA, USA, August 1999. *Cited on page* 46.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, February 2004. *Cited on page* 52.

S. Ikeda, T. Tanaka, and S.-i. Amari. Information geometry of turbo and low-density parity-check codes. *IEEE Transactions on Information Theory*, 50(6):1097–1114, June 2004. *Cited on page* xxiii.

H. Kadri, M. Davy, A. Rabaoui, Z. Lachiri, and N. Ellouze. Robust audio speaker segmentation using one class SVMs. In *16th European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, August 2008. *Cited on page* 81.

R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*, volume 125 of *Wiley Series in Probability and Statistics*. Wiley, New York, NY, USA, 1997. *Cited on page* xxii.

T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1423–1426, Istanbul, Turkey, June 2000. *Cited on page* 79.

A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 6, pages 3089–3092, Phoenix, AZ, USA, March 1999. *Cited on page* 79.

A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer, New York, NY, USA, 2006. *Cited on page* 100.

R. Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural Computation*, 19(3):780–791, March 2007. *Cited on page* 47.

B. O. Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(3):399–409, May 1936. *Cited on page* 3.

M. Kotti, V. Moschou, and C. Kotropoulos. Speaker segmentation and clustering. *Signal Processing*, 88(5):1091–1124, May 2008. *Cited on page* 79.

B. Kulis, M. A. Sustik, and I. S. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research*, 10:341–376, February 2009. *Cited on page* xxiii.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. *Cited on page* 11.

J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163, January 2005. *Cited on page* xxiii.

T. L. Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4): 613–658, 1995. *Cited on page* 22.

T. L. Lai and H. Xing. Sequential change-point detection when the pre- and post-change parameters are unknown. *Sequential Analysis: Design Methods and Applications*, 29(2):162–175, April 2010. *Cited on pages* 26, 41, 82, 96.

T. L. Lai, T. Liu, and H. Xing. A Bayesian approach to sequential surveillance in exponential families. *Communications in Statistics: Theory and Methods*, 38 (16–17):2958–2968, August 2009. *Cited on pages* 41, 96.

J. Le Roux, A. de Cheveigné, and L. C. Parra. Adaptive template matching with shift-invariant semi-NMF. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, pages 921–928. NIPS Foundation, La Jolla,CA, USA, 2009. *Cited on page* 119.

C.-B. Lee. Estimating the number of change points in exponential families distributions. *Scandinavian Journal of Statistics*, 24(2):201–210, June 1997. *Cited on page* 96.

C.-T. Lee, Y.-H. Yang, and H. H. Chen. Multipitch estimation of piano music by exemplar-based sparse representation. *IEEE Transactions on Multimedia*, 14(3): 608–618, June 2012. *Cited on page* 119.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999. *Cited on page* 45.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 556–562. MIT Press, Cambridge, MA, USA, 2001. *Cited on page* 45.

A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 21–24, Prague, Czech Republic, May 2011a. *Cited on pages* 45, 72, 102.

A. Lefèvre, F. Bach, and C. Févotte. Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 313–316, New Paltz, NY, USA, October 2011b. *Cited on page* 119.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*, G. Casella, S. Fienberg, and I. Olkin, editors. Springer Texts in Statistics. Springer, New York, NY, USA, second edition, 1998. *Cited on page* 4.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*, G. Casella, S. Fienberg, and I. Olkin, editors. Springer Texts in Statistics. Springer, New York, NY, USA, third edition, 2005. *Cited on page* 4.

P. Leveau, L. Daudet, and G. Richard. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *5th International Conference on Music Information Retrieval (ISMIR)*, pages 72–75, Barcelona, Spain, October 2004. *Cited on pages* 92, 93.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, October 2006. *Cited on page* 15.

C.-J. Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6):1589–1596, November 2007. *Cited on page* 73.

P.-C. Lin, J.-C. Wang, J.-F. Wang, and H.-C. Sung. Unsupervised speaker change detection using SVM training misclassification rate. *IEEE Transactions on Computers*, 56(9):1234–1244, September 2007. *Cited on page* 81.

M. Liu, B. C. Vemuri, S.-i. Amari, and F. Nielsen. Total Bregman divergence and its applications to shape retrieval. In *23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3463–3468, San Francisco, CA, USA, June 2010. *Cited on page* xxiii.

M. Liu, B. C. Vemuri, S.-i. Amari, and F. Nielsen. Shape retrieval using hierarchical total Bregman soft clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2407–2419, December 2012. *Cited on page* xxiii.

M. Liuni, A. Röbel, M. Romito, and X. Rodet. Rényi information measures for spectral change detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3824–3827, Prague, Czech Republic, May 2011. *Cited on page* 80.

G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, December 1971. *Cited on page* 22.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, January 2010. *Cited on page* 119.

M. Marolt. Non-negative matrix factorization with selective sparsity constraints for transcription of bell chiming recordings. In *6th Sound and Music Computing Conference (SMC)*, pages 137–142, Porto, Portugal, July 2009. *Cited on page* 101.

T. Matumoto. Any statistical manifold has a contrast function—On the C^3-functions taking the minimum at the diagonal of the product manifold. *Hiroshima Mathematical Journal*, 23(2):327–332, 1993. *Cited on page* xxii.

Y. Mei. Sequential change-point detection when unknown parameters are present in the pre-change distribution. *The Annals of Statistics*, 34(1):92–122, February 2006. *Cited on page* 26.

M. Mihoko and S. Eguchi. Robust blind source separation by beta divergence. *Neural Computation*, 14(8):1859–1886, August 2002. *Cited on pages* xxiii, 41, 47.

S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian exponential family PCA. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 1089–1096. NIPS Foundation, La Jolla, CA, USA, 2009. *Cited on page* 46.

S. Mohamed, K. Heller, and Z. Ghahramani. Evaluating Bayesian and L_1 approaches for sparse unsupervised learning. In J. Langford and J. Pineau, editors, *29th International Conference on Machine Learning (ICML)*, pages 751–758, Edinburgh, UK, June/July 2012. *Cited on page* 46.

T. Morimoto. Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, March 1963. *Cited on page* 15.

G. V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, December 1986. *Cited on page* 22.

N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, July 2004. *Cited on page* xxiii.

M. K. Murray and J. W. Rice. *Differential Geometry and Statistics*, volume 48 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, UK, 1993. *Cited on page* xxii.

G. J. Mysore and P. Smaragdis. Relative pitch estimation of multiple instruments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 313–316, Taipei, Taiwan, April 2009. *Cited on page* 101.

G. J. Mysore, P. Smaragdis, and B. Raj. Non-negative hidden Markov modeling of audio with application to source separation. In V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, editors, *Latent Variable Analysis and Signal Separation: 9th International Conference, LVA/ICA 2010, St. Malo, France, September 27-30, 2010, Proceedings*, volume 6365 of *Lecture Notes in Computer Science*, pages 140–148. Springer, Berlin/Heidelberg, Germany, 2010. *Cited on page* 118.

H. Nagaoka and S.-i. Amari. Differential geometry of smooth families of probability distributions. Technical report, Department of Mathematical Engineering and Instrumentation Physics, Faculty of Engineering, University of Tokyo, Tokyo, Japan, October 1982. *Cited on page* xxii.

M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 283–288, Kittilä, Finland, August/September 2010a. *Cited on pages* 47, 62, 103.

M. Nakano, J. Le Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama. Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms. In V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, editors, *Latent Variable Analysis and Signal Separation: 9th International Conference, LVA/ICA 2010, St. Malo, France, September 27-30, 2010, Proceedings*, volume 6365 of *Lecture Notes in Computer Science*, pages 149–156. Springer, Berlin/Heidelberg, Germany, 2010b. *Cited on page* 118.

B. Niedermayer. Non-negative matrix division for the automatic transcription of polyphonic music. In *9th International Conference on Music Information Retrieval (ISMIR)*, pages 544–549, Philadelphia, PA, USA, September 2008. *Cited on page* 101.

F. Nielsen and S. Boltz. The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, August 2011. *Cited on pages* xxiv, 18.

F. Nielsen and R. Nock. A fast deterministic smallest enclosing disk approximation algorithm. *Information Processing Letters*, 93(6):263–268, March 2005. *Cited on page* xxiv.

F. Nielsen and R. Nock. On the smallest enclosing information disk. *Information Processing Letters*, 105(3):93–97, January 2008. *Cited on page* xxiv.

F. Nielsen and R. Nock. Approximating smallest enclosing balls with applications to machine learning. *International Journal of Computational Geometry & Applications*, 19(5):389–414, October 2009a. *Cited on page* xxiv.

F. Nielsen and R. Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2882–2904, June 2009b. *Cited on pages* xxiv, 17.

F. Nielsen and R. Nock. Skew Jensen-Bregman Voronoi diagrams. In M. L. Gavrilova, C. J. K. Tan, and M. A. Mostafavi, editors, *Transactions on Computational Science XIV: Special Issue on Voronoi Diagrams and Delaunay Triangulation*, volume 6970 of *Lecture Notes in Computer Science*, pages 102–128. Springer, Berlin/Heidelberg, Germany, 2011. *Cited on page* xxiv.

F. Nielsen, J.-D. Boissonnat, and R. Nock. On Bregman Voronoi diagrams. In N. Bansal, K. Pruhs, and C. Stein, editors, *18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 746–755, New Orleans, LA, USA, January 2007. *Cited on page* xxiv.

F. Nielsen, P. Piro, and M. Barlaud. Tailored Bregman ball trees for effective nearest neighbors. In *25th European Workshop on Computational Geometry (EuroCG)*, pages 29–32, Brussels, Belgium, March 2009a. *Cited on page* xxiv.

F. Nielsen, P. Piro, and M. Barlaud. Bregman vantage point trees for efficient nearest neighbor queries. In *IEEE International Conference on Multimedia and*

Expo (ICME), pages 878–881, New York, NY, USA, June/July 2009b. *Cited on page* xxiv.

R. Nock and F. Nielsen. Fitting the smallest enclosing Bregman ball. In J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo, editors, *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, volume 3720 of *Lecture Notes in Computer Science*, pages 649–656. Springer, Berlin/Heidelberg, Germany, 2005. *Cited on page* xxiv.

R. Nock and F. Nielsen. On weighting clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1223–1235, August 2006. *Cited on page* xxiii.

R. Nock and F. Nielsen. Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2048–2059, November 2009. *Cited on page* xxiii.

R. Nock, P. Luosto, and J. Kivinen. Mixed Bregman clustering with approximation guarantees. In W. Daelemans, B. Goethals, and K. Morik, editors, *Machine Learning and Knowledge Discovery in Databases: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II*, volume 5212 of *Lecture Notes in Computer Science*, pages 154–169. Springer, Berlin/Heidelberg, Germany, 2008. *Cited on page* xxiii.

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, December 2008. *Cited on page* 102.

M. K. Omar, U. Chaudhari, and G. Ramaswamy. Blind change detection for audio segmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 501–504, Philadelphia, PA, USA, March 2005. *Cited on page* 80.

P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1):23–35, May 1997. *Cited on page* 45.

P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, June 1994. *Cited on page* 45.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, June 1954. *Cited on page* 22.

L. Pardo. *Statistical Inference Based on Divergence Measures*. Statistics: A Series of Textbooks and Monographs. Chapman & Hall/CRC, Boca Raton, FL, USA, 2006. *Cited on pages* 41, 47.

J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636, Utrecht, Netherlands, August 2010. *Cited on page* 78.

A. M. Peter and A. Rangarajan. Shape analysis using the Fisher-Rao Riemannian metric: Unifying shape representation and deformation. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro (ISBI)*, pages 1164–1167, Arlington, VA, USA, April 2006. *Cited on page* xxiii.

A. M. Peter and A. Rangarajan. Information geometry for landmark shape analysis: Unifying shape representation and deformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):337–350, February 2009. *Cited on page* xxiii.

G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 23(5):1543–1561, October 1995. *Cited on page* xxii.

E. J. G. Pitman. Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(4):567–579, December 1936. *Cited on page* 3.

M. Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, 13(1):206–227, March 1985. *Cited on page* 22.

M. Pollak. Average run lengths of an optimal method of detecting a change in distribution. *The Annals of Statistics*, 15(2):749–779, June 1987. *Cited on page* 22.

M. Pollak and D. Siegmund. Approximations to the expected sample size of certain sequential tests. *The Annals of Statistics*, 3(6):1267–1282, November 1975. *Cited on page* 22.

A. S. Polunchenko and A. G. Tartakovsky. State-of-the-art in sequential change-point detection. *Methodology and Computing in Applied Probability*, 14(3):649–684, September 2012. *Cited on page* 23.

V. H. Poor and O. Hadjiliadis. *Quickest Detection*. Cambridge University Press, New York, NY, USA, 2009. *Cited on page* 22.

I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon. Overlapping community detection using Bayesian non-negative matrix factorization. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, 83(6): 9 pages, June 2011. *Cited on page* 46.

S. A. Raczyński, N. Ono, and S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *8th International Conference on Music Information Retrieval (ISMIR)*, pages 381–386, Vienna, Austria, September 2007. *Cited on page* 101.

R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3): 294–307, March 2005. *Cited on page* 24.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945. *Cited on page* xxi.

A. Rényi. On measures of entropy and information. In J. Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561. University of California Press, Berkeley/Los Angeles, CA, USA, 1961. *Cited on page* 16.

Y. Ritov. Decision theoretic optimality of the Cusum procedure. *The Annals of Statistics*, 18(3):1464–1469, September 1990. *Cited on page* 22.

A. Röbel. Onset detection by means of transient peak classification. In *Music Information Retrieval Evaluation eXchange (MIREX)*, Miami, FL, USA, October 2011. *Cited on page* 94.

S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, August 1959. *Cited on page* 22.

S. W. Roberts. A comparison of some control charts procedures. *Technometrics*, 8 (3):411–430, August 1966. *Cited on page* 22.

R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*, P. A. Griffiths, M. Morse, and E. M. Stein, editors. Princeton University Press, Princeton, NJ, USA, 1970. *Cited on page* 6.

J. S. Rustagi. *Variational Methods in Statistics*, volume 121 of *Mathematics in Science and Engineering*. Academic Press, Inc., New York, NY, USA, 1976. *Cited on page* 52.

S. O. Sadjadi and J. H. L. Hansen. A scanning window scheme based on SVM training error rate for unsupervised audio segmentation. In *18th European Signal Processing Conference (EUSIPCO)*, pages 1262–1266, Aalborg, Denmark, August 2010. *Cited on page* 81.

M. N. Schmidt and H. Laurberg. Nonnegative matrix factorization with Gaussian process priors. *Computational Intelligence and Neuroscience*, 2008: 10 pages, 2008. *Cited on page* 45.

M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, editors, *Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009, Proceedings*, volume 5441 of *Lecture Notes in Computer Science*, pages 540–547. Springer, Berlin/Heidelberg, Germany, 2009. *Cited on page* 45.

O. Schwander and F. Nielsen. Reranking with contextual dissimilarity measures from representational Bregman k-means. In P. Richard and J. Braz, editors, *VISAPP 2010: Proceedings of the Fifth International Conference on Computer Vision Theory and Applications, Angers, France, May 17-21, 2010*, volume 1, pages 118–123. INSTICC Press, Setubal, Portugal, 2010. *Cited on page* xxiii.

O. Schwander and F. Nielsen. Non-flat clustering with alpha-divergences. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2100–2103, Prague, Czech Republic, May 2011. *Cited on page* xxiii.

F. Sha and L. K. Saul. Real-time pitch determination of one or more voices by nonnegative matrix factorization. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 17, pages 1233–1240. MIT Press, Cambridge, MA, USA, 2005. *Cited on page* 102.

F. Sha, Y. Lin, L. K. Saul, and D. D. Lee. Multiplicative updates for nonnegative quadratic programming. *Neural Computation*, 19(8):2004–2031, August 2007. *Cited on page* 73.

M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008: 8 pages, 2008. *Cited on pages* 46, 101.

W. A. Shewhart. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association*, 20(152): 546–548, December 1925. *Cited on page* 22.

W. A. Shewhart. *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company, Inc., New York, NY, USA, 1931. *Cited on page* 22.

A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8(1):22–46, 1963. *Cited on page* 22.

A. N. Shiryaev. *Optimal Stopping Rules*, volume 8 of *Applications of Mathematics*. Springer, Berlin/Heidelberg, Germany, 1978. *Cited on page* 22.

M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation, classification, and clustering of broadcast news audio. In *DARPA Speech Recognition Workshop*, pages 97–99, Chantilly, VA, USA, February 1997. *Cited on page* 79.

D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, 23(1):255–271, February 1995. *Cited on pages* 26, 82.

A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In W. Daelemans, B. Goethals, and K. Morik, editors, *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II*, volume 5212 of *Lecture Notes in Computer Science*, pages 358–373. Springer, Berlin/Heidelberg, Germany, 2008. *Cited on page* 46.

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In C. G. Puntonet and A. Prieto, editors, *Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004, Granada, Spain, September 22-24, 2004, Proceedings*, volume 3195 of *Lecture Notes in Computer Science*, pages 494–499. Springer, Berlin/Heidelberg, Germany, 2004. *Cited on page* 71.

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)*, pages 177–180, New Paltz, NY, USA, October 2003. *Cited on page* 100.

P. Smaragdis and B. Raj. Shift-invariant probabilistic latent component analysis. Technical report, Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, December 2007. *Cited on pages* 46, 101.

P. Smaragdis, B. Raj, and M. Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2069–2072, Las Vegas, NV, USA, March/April 2008. *Cited on page* 101.

H. Snoussi. The geometry of prior selection. *Neurocomputing*, 67:214–244, August 2005. *Cited on page* xxiii.

C. Sonesson and D. Bock. A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 166(1):5–21, 2003. *Cited on page* 24.

B. K. Sriperumbudur and G. R. G. Lanckriet. A proof of convergence of the concave-convex procedure using Zangwill's theory. *Neural Computation*, 24(6):1391–1407, June 2012. *Cited on page* 73.

Z. Szabó, B. Póczos, and A. Lőrincz. Online group-structured dictionary learning. In *24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2865–2872, Colorado Springs, CO, USA, June 2011. *Cited on page* 119.

V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization. In *Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS)*, Saint-Malo, France, April 2009. *Cited on page* 46.

T. Tanaka. Information geometry of mean-field approximation. *Neural Computation*, 12(8):1951–1968, August 2000. *Cited on page* xxiii.

A. G. Tartakovsky, B. L. Rozovskii, R. B. Blǎzek, and H. Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54(9):3372–3382, September 2006. *Cited on page* 23.

S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5): 1557–1565, September 2006. *Cited on page* 78.

A. Tritschler and R. A. Gopinath. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In *6th European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 2, pages 679–682, Budapest, Hungary, September 1999. *Cited on page* 80.

C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1–2):479–487, July 1988. *Cited on page* 16.

R. Turner, Y. Saatci, and C. E. Rasmussen. Adaptive sequential Bayesian change point detection. In *NIPS Workshop on Temporal Segmentation*, Whistler, Canada, December 2009. *Cited on pages* 41, 96.

G. Tzanetakis and P. R. Cook. Multifeature audio segmentation for browsing and annotation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 103–106, New Paltz, NY, USA, October 1999. *Cited on pages* 77, 81.

N. Vaswani. Additive change detection in nonlinear systems with unknown change parameters. *IEEE Transactions on Signal Processing*, 55(3):859–872, March 2007. *Cited on pages* 41, 95.

B. C. Vemuri, M. Liu, S.-i. Amari, and F. Nielsen. Total Bregman divergence and its applications to DTI analysis. *IEEE Transactions on Medical Imaging*, 30(2): 475–483, February 2011. *Cited on page* xxiv.

J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 23, pages 2343–2351. NIPS Foundation, La Jolla, CA, USA, 2010. *Cited on page* 24.

E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, March 2010. *Cited on pages* 102, 113, 115.

T. Virtanen and A. T. Cemgil. Mixtures of gamma priors for non-negative matrix factorization based speech separation. In T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, editors, *Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009, Proceedings*, volume 5441 of *Lecture Notes in Computer Science*, pages 646–653. Springer, Berlin/Heidelberg, Germany, 2009. *Cited on page* 45.

T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *NIPS Workshop on Advances in Models for Acoustic Processing*, Whistler, Canada, December 2006. *Cited on page* 101.

T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1825–1828, Las Vegas, NV, USA, March/April 2008. *Cited on page* 45.

P. W. Vos. Geometry of f-divergence. *Annals of the Institute of Statistical Mathematics*, 43(3):515–537, September 1991. *Cited on page* xxii.

M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*, volume 1(1–2) of *Foundation and Trends in Machine Learning*. now Publishers Inc., Hanover, MA, USA, 2008. *Cited on page* xxiii.

G. B. Wetherill and D. W. Brown. *Statistical Process Control: Theory and Practice*. Chapman & Hall, London, UK, 1991. *Cited on page* 23.

A. S. Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12(6):601–611, November 1976. *Cited on page* 23.

L. Wu, S. C.-H. Hoi, R. Jin, J. Zhu, and N. Yu. Learning Bregman distance functions for semi-supervised clustering. *IEEE Transactions on Knowledge and Data Engineering*, 24(3):478–491, March 2012. *Cited on page* xxiv.

S. Yang and M. Ye. Multistability of α-divergence based NMF algorithms. *Computers & Mathematics with Applications*, 64(2):73–88, July 2012. *Cited on page* 73.

C. Yeh, A. Röbel, and X. Rodet. Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1116–1126, August 2010. *Cited on page* 115.

A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, April 2003. *Cited on page* 57.

W. I. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice-Hall International Series in Management. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1969. *Cited on page* 73.

J. Zhang. Divergence function, duality, and convex analysis. *Neural Computation*, 16(1):159–195, January 2004. *Cited on page* xxii.

M. Zhong and M. Girolami. Reversible jump MCMC for non-negative matrix factorization. In D. van Dyk and M. Welling, editors, *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR Workshop and Conference Proceedings*, pages 663–670, Clearwater Beach, FL, USA, April 2009. *Cited on page* 45.