Neurocomputing 141 (2014) 26–36

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A convex formulation for informed source separation in the single channel setting ☆

Augustin Lefèvre [a,b,*], François Glineur [b], P.-A. Absil [b]

[a] Niland, 96 bis boulevard Raspail, 75006 Paris, France
[b] ICTEAM Institute, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form
17 October 2013
Accepted 9 December 2013
Available online 12 April 2014

Keywords:
Source separation
Machine learning
Music signal processing
Nonnegative matrix factorization
Nonsmooth optimization

ABSTRACT

Blind audio source separation is well-suited for the application of unsupervised techniques such as nonnegative matrix factorization (NMF). It has been shown that on simple examples, it retrieves sensible solutions even in the single-channel setting, which is highly ill-posed. However, it is now widely accepted that NMF alone cannot solve single-channel source separation, for real world audio signals. Several proposals have appeared recently for systems that allow the user to control the output of NMF, by specifying additional equality constraints on the coefficients of the sources in the time-frequency domain. In this article, we show that matrix factorization problems involving these constraints can be formulated as convex problems, using the nuclear norm as a low-rank inducing penalty. We propose to solve the resulting nonsmooth convex formulation using a simple subgradient algorithm. Numerical experiments confirm that the nuclear norm penalty allows the recovery of (approximately) low-rank solutions that satisfy the additional user-imposed constraints. Moreover, for a given computational budget, we show that this algorithm matches the performance or even outperforms state-of-the-art NMF methods in terms of the quality of the estimated sources.

1. Introduction

Single-channel source separation is an underdetermined problem, commonly used as a pre-processing technique for higher-level tasks (speech recognition in complex environments, polyphonic music transcription, etc.). While exact source recovery cannot be expected in general, a key ingredient in source separation techniques consists in assuming some form of redundancy in the data, which renders the problem overdetermined. This is typically done by representing audio tracks in the time-frequency domain as low-rank matrices. Nonnegative matrix factorization was first applied to audio signals for polyphonic transcription [24], although it was already used in other fields [21,15].

An important idea underlying matrix factorization techniques for audio signals is that they recover a representation of signals in terms of template signals modulated by location-dependent gains. In the field of music signal processing, this idea was supported by experiments on simple music signals [6]. In computer vision, similar experiments suggested that a part-based representation of visual objects could be retrieved by NMF [15]. The miracle of part-based representation no longer works for real music or speech signals, because they cannot be assumed to satisfy the low-rank hypothesis, but it has spawned several interesting research tracks: parameterized templates were introduced in [27] in order to match the harmonic structure of many musical instruments; probabilistic models and penalty functions to favor smooth time-varying gains in [30,8]; Markov models, to stabilize the recognition of vowels in speech processing [16].

In parallel to these research tracks, linear models for audio signals have also been the subject of many contributions. These models rely on the library approach (or dictionary approach), where audio templates correspond to actual signals stored offline in libraries, each specific to an instrument. The University of Iowa's electronic music studios, for instance, have made available recordings of isolated notes for many popular instruments: violin, piano, cello, more generally instruments belonging to the family of woodwind, brass, or string instruments. Due to the large size of the libraries, there are many ways to represent any audio signals as a linear combination of audio templates. Thus, in the library approach, structured decompositions are introduced, based on simple principles: if an instrument is present in the mix, only a few of its templates should be used at the same time [26]; in the case where the sources are unknown group structures are employed to select the appropriate libraries [4].

More recently, several contributions have been made to take into account prior information specific to the target mix signal: manual

☆This paper presents results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

* Corresponding author.
E-mail addresses: augustin.lefevre@niland.io (A. Lefèvre), francois.glineur@uclouvain.be (F. Glineur), pa.absil@uclouvain.be (P.-A. Absil).

http://dx.doi.org/10.1016/j.neucom.2013.12.053
0925-2312/

segmentation of audio tracks [19], MIDI aligned music scores [10,11], time-aligned pitch estimates for the singing voice [5]. A common trait of these methods is that they are all based on a simple extension of NMF: annotations are used to specify equality constraints in the matrix of activation coefficients in NMF, setting them to known values. Thus, annotations help learn a source specific dictionary on segments of the recording where only that source is active: in this way, manual segmentation of audio signals allows a blind source separation task to be cast as a supervised linear model. In [10], prior information consists in the score that the music follows. Digital music synthesizers are used to provide a rough guess of the sources. All these contributions are now identified as the category of *informed source separation* methods. The formulation proposed in this article belongs to this category.

While time segmentation of audio signals allows us to use supervised learning techniques, it is not always applicable. Instead, one can always rely on a universal property of natural signals: they have a very sparse representation in the time–frequency domain. This property, dubbed W-disjoint orthogonality, is at the heart of several source separation techniques in the multiple microphone setting [31,1].

In a previous contribution [14], we formulated a problem of nonnegative matrix factorization (NMF) with additional equality constraints, consistently with the strong tradition in audio source separation. Results on the SISEC database showed that we can obtain state-of-the-art results while annotating only a fraction of the spectrogram; since user annotation is difficult and time-consuming, we also experimented with automatic annotation methods, relying on supervised learning. Interaction with the user has been further explored in [3,7]. Although based on a slightly different technique, called probabilistic latent component analysis (PLCA), the formulation used in [3] can be viewed as NMF where dissimilarity between observations and the model is measured with a Kullback–Leibler divergence.

While it gives satisfactory results, NMF is hard to solve: for typical values of the "rank" parameter used in audio, algorithms cannot be guaranteed to converge to globally optimal solutions, and there is no alternative but to resort to algorithms that converge to local minima. In practice, this means that several initial points should be tried and the best be selected on a principled basis. One would be tempted to replace the strict low-rank constraint by a convex penalty function favoring low-rank solutions.

The main contribution of this paper is to show that we can replace NMF by a matrix approximation problem involving nonnegativity constraints, low-rank inducing penalty functions and constraints on the coefficients of the solutions to model additional information provided by the user, i.e. annotations. The main advantage of such a formulation is that one can borrow tools from the field of convex optimization to construct algorithms that retrieve source estimates of similar if not better quality, for a comparable computational budget, as shown in preliminary results [13]. In this paper, we give a detailed presentation of a subgradient algorithm used to solve the proposed formulation, and show that it has the desired effect of finding solutions that are (approximately) low-rank. Our second contribution, which we detail in Section 5.1, is related to the way we let the user specify annotations: by restricting the set of annotated time-frequency coefficients to those whose target values are zero, we show that our formulation can gain robustness, at a small sacrifice in terms of generality.

The rest of this paper is organized as follows: in Section 2, we review of well-established techniques for single-channel source separation: time–frequency transforms, filtering techniques for source estimates recovery, and evaluation metrics. In Section 3, we introduce a formulation of informed source separation using nonnegative matrix factorization which was previously proposed [14]. In Section 3.2, we discuss a convex formulation of annotation-informed source separation, dubbed `AISS_lownuc`, in the form of a low-rank matrix approximation problem with a low-rank inducing penalty term, and equality constraints. User-provided annotations are encoded as equality constraints, and those are key to the success of our formulation. After presenting in Section 4 our algorithm for `AISS_lownuc`, we investigate in Section 5 the impact of various choices of annotations, and demonstrate the benefits of our convex formulation compared with NMF.

2. Time–frequency analysis and audio source separation

This section is a brief introduction to audio source separation. In Section 2.1, we present time–frequency transforms, which allow to transform a one-dimensional audio signal into a two-dimensional object, frequency and time being now the dimensions of variation. The matrix factorization problem that we introduce is indeed posed in the time–frequency domain, so that an input time–frequency matrix is separated as a sum of matrices, which are interpreted as source terms (as illustrated in Fig. 3). We refer the reader to textbooks such as [20] for a complete presentation of time–frequency transforms and their many applications, such as modifying the duration of an audio signal of modifying its pitch.

Next, we explain in Section 2.2 how to transform those source estimates back as audio signals, using *time–frequency masking*: early proposals for source separation recognized filtering as the best way to avoid artifacts due to inexact solutions [24,6]. In Section 2.3, we summarize evaluation metrics for audio source separation [28], and define the notion of oracle estimates, in controlled experiments where the true source signals are known in advance.

2.1. Time–frequency representation of audio signals

Single-channel source separation consists in recovering a certain number of unknown source signals from measurements of their sum. The first step in single-channel source separation consist in finding a representation of the source signals that enhances their redundancy. As we shall explain in this section, this is done by computing their spectrogram, which is a time–frequency representation. Time–frequency representations of audio signals are sparse and redundant, which is key to the success of blind source separation.

The computation of spectrograms is illustrated in Fig. 1: short time segments are extracted from the signal and multiplied coefficientwise by a window function. Successive windows overlap by a fraction of their length, which is usually taken as 50%. On each of these segments, a Fourier transform is computed. Thus, from a one-dimensional signal $x \in \mathbb{R}^T$, we obtain a complex matrix C of size $F \times N$ where $FN \simeq 2T$ (because of the 50% overlap between windows). These preliminary steps correspond to computing the short time Fourier transform (STFT):

$$C_{fn} = \sum_{t=1}^{F} x_{t+(n-1)H} w_t \exp\left(-\frac{2(f-1)\pi(t-1)}{F}\right)$$

for all $f \in \{1 \ldots F\}$, and $n \in \{1 \ldots N\}$. The so-called hop size H determines the overlap between successive windows, $w \in \mathbb{R}^F$ is a window function, and N is chosen to match the size of the signal. To make this possible, the signal should be appropriately zero-padded beforehand. We refer the reader to textbooks such as [20] for more explanations. Finally, we take $Y_{fn} = |C_{fn}|^2$, in order to obtain approximate invariance to translations of the signal. Coefficient Y_{fn} measures the amount of energy of the signal at frequency f and time index n in the time–frequency plane. This magnitude is represented as a color code in Fig. 1: blue for small coefficients, and red for high coefficients.