Chapter 11
Itakura–Saito Nonnegative Factorizations of the Power Spectrogram for Music Signal Decomposition

Cédric Févotte
TELECOM ParisTech, France

ABSTRACT
Nonnegative matrix factorization (NMF) is a popular linear regression technique in the fields of machine learning and signal/image processing. Much research about this topic has been driven by applications in audio. NMF has been for example applied with success to automatic music transcription and audio source separation, where the data is usually taken as the magnitude spectrogram of the sound signal, and the Euclidean distance or Kullback-Leibler divergence are used as measures of fit between the original spectrogram and its approximate factorization. In this chapter the authors give evidence of the relevance of considering factorization of the power spectrogram, with the Itakura-Saito (IS) divergence. Indeed, IS-NMF is shown to be connected to maximum likelihood inference of variance parameters in a well-defined statistical model of superimposed Gaussian components and this model is in turn shown to be well suited to audio. Furthermore, the statistical setting opens doors to Bayesian approaches and to a variety of computational inference techniques. They discuss in particular model order selection strategies and Markov regularization of the activation matrix, to account for time-persistence in audio. This chapter also discusses extensions of NMF to the multichannel case, in both instantaneous or convolutive recordings, possibly underdetermined. The authors present in particular audio source separation results of a real stereo musical excerpt.

DOI: 10.4018/978-1-61520-919-4.ch011
INTRODUCTION

Nonnegative matrix factorization (NMF) is a linear regression technique, employed for nonsubtractive, part-based representation of nonnegative data. Given a data matrix \( V \) of dimensions \( F \times N \) with nonnegative entries, NMF is the problem of finding a factorization

\[ V \approx WH \]  

(1)

where \( W \) and \( H \) are nonnegative matrices of dimensions \( F \times K \) and \( K \times N \), respectively. \( K \) is usually chosen such that \( FK + KN << FN \), hence reducing the data dimension. Early works about NMF include (Paatero, 1997) and (Lee and Seung, 1999), the latter in particular prove very influential. NMF has been applied to diverse problems (such as pattern recognition, clustering, data mining, source separation, collaborative filtering) in many areas (such as text processing, bioinformatics, signal/image processing, finance). Much research about NMF has been driven by applications in audio, namely automatic music transcription (Smaragdis and Brown, 2003; Abdallah and Plumbley, 2004) and source separation (Virtanen, 2007; Smaragdis, 2007), where the data \( V \) is usually taken as the magnitude spectrogram of the audio signal.

Along Vector Quantization (VQ), Principal Component Analysis (PCA) or Independent Component Analysis (ICA), NMF provides an unsupervised linear representation of data, in the sense that a data point \( v_n \) (\( n \)th column of \( V \)) is approximated as a linear combination of salient features:

A distinctive feature of NMF with respect to VQ, PCA or ICA is that it keeps \( W \) and \( h_n \) nonnegative, hence improving the interpretability of the learnt dictionary and of the activation coefficients when the data is nonnegative. More precisely, the nonnegativity restriction on \( W \) allows the learnt features to belong to the same space than data, while the nonnegative restriction on \( H \) favors part-based decomposition as subtractive combination are forbidden, i.e, the data has to be “assembled” from the elementary building blocks in \( W \). As such, in their seminal paper Lee and Seung (1999) show how parts of faces (noise, eyes, cheeks, etc.) can be learnt from a training set composed of faces, when the bases returned by PCA or VQ are more “holistic” in the sense that each feature attempts to generalize as much as possible the entire dataset. The effect of NMF onto an audio spectrogram is illustrated on Figure 2. It can be seen that the NMF model is well suited to the composite structure of music in the sense that the factorization can be expected to separate mingled patterns in the spectrogram; the patterns may correspond to spectra of elementary musical objects such as notes or percussions or, as we shall see later, higher level structures.

In the literature, the factorization (1) is usually achieved through minimization of a measure of fit defined by

\[ D(V \mid WH) = \sum_{j=1}^{F} \sum_{n=1}^{N} d([V]_{jn}, [WH]_{jn}) \]  

(2)

where \( d(x|y) \) is a scalar cost function, typically a positive function with a single minimum 0 for \( x = y \). The minimization, with respect to \( W \) and \( H \), is subject to nonnegativity constraints on the coefficients of both factors. Popular cost functions are the Euclidean distance, here defined as

\[ d_{EUC}(x \mid y) = \frac{1}{2}(x - y)^2 \]  

(3)