Digital Signal Processing 47 (2015) 178–191

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Non-negative tensor factorization models for Bayesian audio processing

Umut Şimşekli [a,*], Tuomas Virtanen [b], Ali Taylan Cemgil [a]

[a] Department of Computer Engineering, Boğaziçi University, 34342, Bebek, İstanbul, Turkey
[b] Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland

ARTICLE INFO

Article history:
Available online 25 March 2015

Keywords:
Nonnegative matrix and tensor factorization
Coupled factorization
Bayesian audio modeling
Bayesian inference

ABSTRACT

We provide an overview of matrix and tensor factorization methods from a Bayesian perspective, giving emphasis on both the inference methods and modeling techniques. Factorization based models and their many extensions such as tensor factorizations have proved useful in a broad range of applications, supporting a practical and computationally tractable framework for modeling. Especially in audio processing, tensor models help in a unified manner the use of prior knowledge about signals, the data generation processes as well as available data from different modalities. After a general review of tensor models, we describe the general statistical framework, give examples of several audio applications and describe modeling strategies for key problems such as deconvolution, source separation, and transcription.

1. Introduction

With the recent technological advances of sensor and communication technologies, the cost of data acquisition and storage is significantly reduced. Consequently, the last decade has witnessed the dramatic increase in the amount of data that can be easily collected. One important facet of data processing is extracting meaningful information from highly structured datasets that can be of interest for scientific, financial, or technological purposes.

The key to exploiting the potential of large datasets lies in developing computational techniques that can efficiently extract meaningful information. These computational methods must be scalable and tailored for the specifics of an application, but still be versatile enough to be useful in several scenarios. In this paper, we will focus on audio processing and review one particular class of such models, that provide a favorable balance between high modeling accuracy, ease of implementation and ease of management of required computational resources. This class of models, coined under the name of tensor factorization models along with their Bayesian interpretations, will be the focus of this tutorial paper. The mathematical setup may look somewhat abstract at a first sight, but the generic nature of the approach makes tensors suitable for a broad range of applications where complicated structured datasets need to be analyzed. In particular, we will show examples in the domain of audio processing where significant progress has been achieved using tensor methods. While the modeling and inference strategies can be applied in the broader context of general audio and other non-stationary time series analysis, the hierarchical Bayesian nature of the framework makes the approach particularly suitable for the analysis of acoustical signals.

In audio processing, an increasing number of applications are developed that can handle challenging acoustical conditions and highly variable sound sources. Here, one needs to exploit the inherent structure of acoustic signals to address some of the key problems such as denoising, restoration, interpolation, source separation, transcription, bandwidth extension, upmixing, coding, event recognition and classification. Not surprisingly, many different modeling techniques have been developed for those purposes. However, as is the case for computational modeling of all physical phenomena, we face here with a trade off: accuracy versus computational tractability – a physically realistic and accurate model may be too complex to meet the demands of a given application to be useful in practice.

Typically, there is a lot of *a-priori* knowledge available for acoustic signals. This includes knowledge of the physical or cognitive mechanisms by which sounds are generated or perceived, as well as the hierarchical nature by which they are organized in an acoustical scene. In more specific domains, such as music transcription or audio event recognition, more specialized assumptions about the hierarchical organization are needed. Yet, the

* Corresponding author.
E-mail addresses: umut.simsekli@boun.edu.tr (U. Şimşekli), tuomas.virtanen@tut.fi (T. Virtanen), taylan.cemgil@boun.edu.tr (A.T. Cemgil).

http://dx.doi.org/10.1016/j.dsp.2015.03.011
1051-2004/

resulting models often possess complex statistical structure and highly adaptive and powerful computational techniques are needed to perform inference.

Factorization-based modeling has been useful in addressing the modeling accuracy versus computational requirement trade off in various domains beyond audio signal processing [1], with prominent examples such as text processing [2], bioinformatics [3], computer vision [4], social media analysis [5], and network traffic analysis [6]. The aim in such modeling strategies is to decompose an observed matrix or tensor (multidimensional array) into semantically meaningful factors in order to obtain useful predictions. Meanwhile, the factors themselves also provide a useful feature representation about the specifics of the domain.

In this paper, we review tensor based statistical models and associated inference methods developed recently for audio and music processing and describe various extensions and applications of these models. In Section 2, we illustrate the ideas of factorization based modeling, and then in Section 3 we describe a probabilistic interpretation of these models. The probabilistic interpretation opens up the way for a full Bayesian treatment via Bayesian hierarchical modeling. This leads to a very natural means for unification, allowing the formulation of highly structured probabilistic models for audio data at the various levels of abstraction, as we will illustrate in Section 6. The paper concludes with remarks on future research directions.

2. Factorization-based data modeling

In this section, we will describe the basics of factorization based modeling, and describe extensions such as coupled tensor factorizations and nonnegative decompositions. This section will describe the main structure and the notation.

In many applications, data can be represented as a matrix, for example, the spectrogram of an audio signal (frequency vs time), a dataset of images (pixel coordinates vs instances), word frequencies among different documents (words vs documents), and the adjacency structure of a graph (nodes vs nodes) to name a few. Here the indices of the matrix correspond to the entities, and the matrix elements describe a relation between the two entities. Matrix Factorization (MF) models are one of the most widely used methods for analyzing the data that involve two entities [7–10]. The goal in these models is to calculate a factorization of the form:

$$X_1(i,j) \approx \hat{X}_1(i,j) = \sum_k Z_1(i,k) Z_2(k,j) \tag{1}$$

where X_1 is the given data matrix, $\hat{X}_1$ is an approximation to X_1, and Z_1, and Z_2 are factor matrices to be estimated. Even though we have a single observed matrix in this model, we use a subscript in X_1 since we will consider factorization models that involve more than one observed matrix or tensor, later in this section. Here, X_1 is expressed as the product of Z_1 and Z_2, where Z_1 is considered as the *dictionary* matrix and Z_2 contains the corresponding *weights*. From another perspective, X_1 is approximated as the sum of inner products of the columns of Z_1 and the rows of Z_2, as illustrated at the top of Fig. 2. Note that, if Z_1 would have been fixed, the problem would have been equivalent to basis regression where the weights (expansion coefficients) Z_2 are estimated [11]. In contrast, in matrix factorization the dictionary (the set of basis vectors) is estimated along with the coefficients. This modeling strategy has been shown to be successful in various fields including signal processing, finance, bioinformatics, and natural language processing [8].

Matrix factorization models are applicable when the observed data encapsulates the relation of two different entities (e.g., i and j in Eq. (1)). However, when the data involves multiple entities

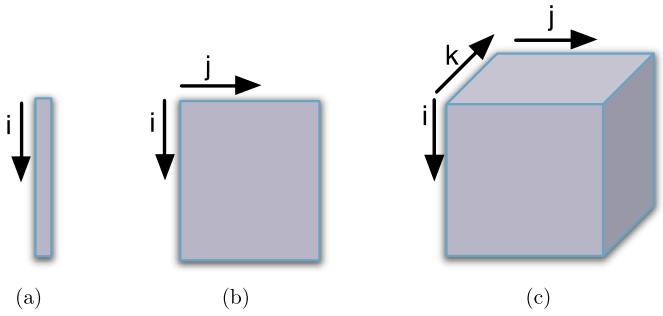

Fig. 1. Illustration of a) a vector $X(i)$: an array with one index, b) a matrix $X(i,j)$ an array with two indices, c) a tensor $X(i,j,k)$: an array with three or more indices. In this study, we refer vectors as tensors with one mode and matrices as tensors with two modes.

of interest, such as ternary or higher order relations it cannot be represented without loss of structure by using matrices. For example a multichannel sound library of several instances may be represented in the time-frequency domain conveniently as an object with several entities, say the power at each (frequency, time, channel, instance). One could in principle 'concatenate' each spectrogram across time and instances to obtain a big matrix, say (frequency × channel, time × instance) but this representation would obscure important structural information – compare simply with representing a matrix with a column vector. Hence one needs naturally multiway tables, the so-called *tensors*, where each element is denoted by $T(i,j,k,\ldots)$. Here, T is the tensor and the indices $i,j,k,\ldots$ are the entities. The number of distinct entities dictates the mode of a tensor. Hence a vector and a matrix are tensors of mode one and two respectively. Tensors are illustrated in Fig. 1 and we will give a more precise and compact definition in Section 3.

For modeling multiway arrays with more than two entities the canonical polyadic decomposition [12,13] (also referred as, CP, PARAFAC, or CANDECOMP) is one of the most popular factorization models. The model, for three entities, is defined as follows:

$$X_2(i,m,r) \approx \hat{X}_2(i,m,r) = \sum_k Z_1(i,k) Z_3(m,k) Z_4(r,k) \tag{2}$$

where the observed tensor X_2 is decomposed as a product of three different matrices. Analogous to MF models, this model approximates X_2 as the sum of 'inner products' of the columns of Z_1, Z_3, and Z_4 as illustrated at the bottom of Fig. 2. This model has been shown to be useful in chemometrics [14], psychometrics [12], and signal processing [8].

Tucker model [15] is another important model for analyzing tensors with three modes, which is a generalization of the PARAFAC model. The model is defined as follows:

$$\begin{aligned} X_3(i,j,k) &\approx \hat{X}_3(i,j,k) \\ &= \sum_p \sum_q \sum_r Z_1(i,p) Z_2(j,q) Z_3(k,r) Z_4(p,q,r) \end{aligned} \tag{3}$$

where X_3 is expressed as the product of three matrices ($Z_{1:3}$) and a 'core tensor' (Z_4). When the core tensor Z_4 is chosen as super diagonal ($Z_4(p,q,r) \neq 0$ only if $p = q = r$), Tucker decomposition reduces to PARAFAC.

2.1. Coupled factorization models

In certain applications, information from different sources are available and need to be combined for obtaining more accurate predictions [16–20]. In musical audio processing, one example is having a large collection of annotated audio data and a collection of symbolic music scores as side information. Similarly, in product recommendation systems, a customer–product rating matrix