An Overview for Non-Negative Matrix Factorization

D

Yu-Jin Zhang
Department of Electronic Engineering, Tsinghua University, China

INTRODUCTION

A basically important problem in Image Engineering (IE), Pattern Recognition (PR), and Computer Vision (CV) is how to find a suitable representation of multivariate data.

In many cases, the primitive data sets or observations are organized as data matrices (or tensors), and described by linear (or multi-linear) combination models. From the algebraic perspective, the formulation of dimensionality reduction can be regarded as decomposing the original data matrix into two factor matrices. The canonical methods, such as principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA), and vector quantization (VQ) et al., are the exemplars of such low-rank approximations. They differ from one another in the statistical properties attributable to the different constraints imposed on the component matrices and their underlying structures; however, they have something in common that there is no constraint in the sign of the elements in the factorized matrices. In other words, the negative component or the subtractive combination is allowed in the representation.

In contrast, a new paradigm of factorization — Non-negative Matrix Factorization (NMF) is quite different in this aspect. NMF is a recently developed, biologically inspired method for nonlinearly finding purely additive, sparse, linear, and low-dimension representations of non-negative multivariate data to consequently make latent structure, feature or pattern in the data clear (Lee, 1999).

NMF makes all representation components non-negative (only purely additive representations are allowable) and nonlinearly implements dimension reduction. Psychological and physiological evidence for NMF is that perception of the whole is based on perception of its parts, which is compatible with the intuitive notion of combining parts to form a whole (Lee, 1999), therefore it is considered to grasp the essence of intelligent or biological data representation in some degree.

Far beyond a mathematical exploration, the philosophy underlying NMF, which tries to formulate a feasible model for learning object parts, is closely relevant to perception mechanism. While the parts-based representation seems intuitive, it is indeed based on physiological and psychological evidence: perception of the whole is based on perception of its parts (Paatero, 1997). In fact there are two complementary connotations in non-negativity — non-negative component and purely additive combination. On the one hand, the negative values of both observations and latent components are physically meaningless in many kinds of real world data, such as image analysis tasks. Meanwhile, the discovered prototypes commonly correspond with certain semantic interpretation.

Besides, NMF usually produces a sparse representation of data, which has been shown to be a useful middle ground between a completely distributed representation and a unary representation (Field 1994). The non-negativity constraint will lead to sort of sparseness naturally (Lee, 1999), which is proved to be a highly effective representation distinguished from both the completely distributed and the solely active component description (Field, 1994).

When NMF is interpreted as a neural-network learning algorithm depicting how the visible variables are generated from the hidden ones, the parts-based representation is obtained from the additive model. A positive number indicates the presence and a zero value represents the absence of some event or component. This conforms nicely to the dualistic properties of neural activity and synaptic strengths in neurophysiology: either excitatory or inhibitory without changing sign (Lee, 1999).

DOI: 10.4018/978-1-4666-5888-2.ch156

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

BACKGROUND

Given an M dimensional random vector $\boldsymbol{x}$ with non-negative elements, whose N observations are denoted as $\boldsymbol{x}_j$, $j = 1, 2, \ldots, N$, let data matrix be $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}_{\geq 0}^{M\times N}$, NMF seeks to decompose $\boldsymbol{X}$ into non-negative $M \times L$ basis matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_L] \in \mathbb{R}_{\geq 0}^{M\times L}$ and non-negative $L \times N$ coefficient matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N] \in \mathbb{R}_{\geq 0}^{L\times N}$, such that $\boldsymbol{X} \approx \boldsymbol{UV}$, where $\mathbb{R}_{\geq 0}^{M\times N}$ stands for the set of $M \times N$ element-wise non-negative matrices. This can also be written as the equivalent vector formula $\mathbf{x}_j \approx \sum_{i=1}^{L} \mathbf{u}_i \mathbf{V}_{ij}$.

It is obvious that $\boldsymbol{vj}$ is the weight coefficient of the observation $\boldsymbol{xj}$ on the columns of $\boldsymbol{U}$, the basis vectors or the latent feature vectors of $\boldsymbol{X}$. Hence, NMF decomposes each data into the linear combination of the basis vectors. Because of the initial condition $L << \min(M, N)$, the obtained basis vectors are incomplete over the original vector space. In other words, this approach tries to represent the high dimensional stochastic pattern with far fewer bases, so the perfect approximation can be achieved successfully only if the intrinsic features are identified in $\boldsymbol{U}$.

In most cases, NMF is viewed as a dimensionality reduction and feature extraction technique with $L << M$, $L << N$; that is, the basis set learnt from NMF model is incomplete, and the energy is compacted. However, in general, L can be smaller, equal or larger than M. However, there are fundamental differences in the decomposition for $L < M$ and $L > M$. It is a sort of sparse coding and compressed sensing with over-complete basis when $L > M$. Hence, L needs not be limited by the dimensionality of the data.

In this situation, it may benefit from the sparseness due to both non-negativity and redundant representation. One approach to obtain this NMF model is to perform the decomposition on the residue matrix $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{UV}$ repeatedly and sequentially (Gupta, 2010).

Considering NMF as a kind of matrix factorization model, three essential questions need answering: (1) existence, whether the nontrivial NMF solutions exist; (2) uniqueness, under what assumptions NMF is, at least in some sense, unique; (3) effectiveness, under what assumptions NMF is able to recover the "right answer." The existence was showed via the theory of Completely Positive (CP) Factorization for the first time in (Vasiloglou, 2009).

The last two concerns were first mentioned and discussed from a geometric viewpoint in (Donoho, 2004). Complete NMF $\boldsymbol{X} = \boldsymbol{UV}$ is considered firstly for the analysis of existence, convexity, and computational complexity. The trivial solution always exists as $\boldsymbol{U} = \boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{I}N$. By relating NMF to CP Factorization, it is showed that every non-negative matrix has a nontrivial complete NMF (Vasiloglou, 2009). As such, CP Factorization is a special case, where a non-negative matrix $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{M\times M}$ is CP if it can be factored in the form $\boldsymbol{X} = \boldsymbol{UU}^{\mathrm{T}}$, $\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{M\times L}$. The minimum L is called the CP-rank of $\boldsymbol{X}$. When combining that the set of CP matrices forms a convex cone with that the solution to NMF belongs to a CP cone, solving NMF is a convex optimization problem (Vasiloglou, 2009).

Using the bilinear model, complete NMF can be rewritten as linear combination of rank-one nonnegative matrices expressed by

$$X = \sum_{i=1}^{L} \mathbf{U}_{\bullet i} \mathbf{V}_{i\bullet} = \sum_{i=1}^{L} \mathbf{U}_{\bullet i} \circ (\mathbf{V}_{i\bullet})^{\mathrm{T}} \tag{1}$$

where $\boldsymbol{U}_{\bullet i}$ is the $\boldsymbol{i}$-th column vector of $\boldsymbol{U}$ while $\boldsymbol{V}_{i\bullet}$ is the $\boldsymbol{i}$-th row vector of $\boldsymbol{V}$, and ° denotes the outer product of two vectors. The smallest L making the decomposition possible is called the non-negative rank of the non-negative matrix $\boldsymbol{X}$, denoted as $rank_+(\boldsymbol{X})$. Moreover, it satisfies the following trivial bounds:

$$\operatorname{rank}(\mathbf{X}) \leq \operatorname{rank}_+(\mathbf{X}) \leq \min(M, N) \tag{2}$$

While PCA can be solved in polynomial time, the optimization problem of NMF, with respect to determining the non-negative rank and computing the associated factorization, is more difficult to solve than its unconstrained counterpart does. It is in fact NP-hard when requiring both the dimension and the factorization rank of $\boldsymbol{X}$ to increase, which was proved via relating it to NP-hard intermediate simplex problem (Vavasis, 2009). This is also the corollary of CP programming, since the CP cone cannot be described in polynomial time despite its convexity. In the special case when $\operatorname{rank}(\boldsymbol{X}) = 1$, complete NMF can be solved in polynomial time. However, the complexity of complete NMF

9 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/an-overview-for-non-negative-matrix-factorization/112568

Related Content

Diffusion of Innovations Theory: Inconsistency Between Theory and Practice
Francisco Chia Cua and Tony C. Garrett (2009). *Handbook of Research on Contemporary Theoretical Models in Information Systems (pp. 242-276).*
www.irma-international.org/chapter/diffusion-innovations-theory/35834

An Efficient Image Retrieval Based on Fusion of Fast Features and Query Image Classification
Vibhav Prakash Singh, Subodh Srivastava and Rajeev Srivastava (2017). *International Journal of Rough Sets and Data Analysis (pp. 19-37).*
www.irma-international.org/article/an-efficient-image-retrieval-based-on-fusion-of-fast-features-and-query-image-classification/169172

A Comparative Analysis of a Novel Anomaly Detection Algorithm with Neural Networks
Srijan Das, Arpita Dutta, Saurav Sharma and Sangharatna Godboley (2017). *International Journal of Rough Sets and Data Analysis (pp. 1-16).*
www.irma-international.org/article/a-comparative-analysis-of-a-novel-anomaly-detection-algorithm-with-neural-networks/186855

Detection of Shotgun Surgery and Message Chain Code Smells using Machine Learning Techniques
Thirupathi Guggulothu and Salman Abdul Moiz (2019). *International Journal of Rough Sets and Data Analysis (pp. 34-50).*
www.irma-international.org/article/detection-of-shotgun-surgery-and-message-chain-code-smells-using-machine-learning-techniques/233596

Data Mining of Chemogenomics Data Using Activity Landscape and Partial Least Squares
Kiyoshi Hasegawa and Kimito Funatsu (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 1723-1731).*
www.irma-international.org/chapter/data-mining-of-chemogenomics-data-using-activity-landscape-and-partial-least-squares/112577