Department of Information and Computer Science

Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis

He Zhang

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 127/2014

Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis

He Zhang

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 19 September 2014 at 12.

Aalto University
School of Science
Department of Information and Computer Science

Supervising professor
Aalto Distinguished Professor Erkki Oja

Thesis advisor
Dr. Zhirong Yang

Preliminary examiners
Research Scientist Ph.D. Rafal Zdunek, Wroclaw University of Technology, Poland
Associate Professor Ph.D. Morten Mørup, DTU Compute, Denmark

Opponent
Associate Professor Ali Taylan Cemgil, Bogazici University, Turkey

Aalto University publication series
DOCTORAL DISSERTATIONS 127/2014

ISBN 978-952-60-5828-3
ISBN 978-952-60-5829-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5829-0

Unigrafia Oy
Helsinki 2014

Finland

Author
He Zhang

Name of the doctoral dissertation
Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 127/2014

Field of research Machine Learning

Manuscript submitted 14 May 2014 **Date of the defence** 19 September 2014

Permission to publish granted (date) 27 June 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Nonnegative Matrix Factorization (NMF) has found a wide variety of applications in machine learning and data mining. NMF seeks to approximate a nonnegative data matrix by a product of several low-rank factorizing matrices, some of which are constrained to be nonnegative. Such additive nature often results in parts-based representation of the data, which is a desired property especially for cluster analysis.

This thesis presents advances in NMF with application in cluster analysis. It reviews a class of higher-order NMF methods called Quadratic Nonnegative Matrix Factorization (QNMF). QNMF differs from most existing NMF methods in that some of its factorizing matrices occur twice in the approximation. The thesis also reviews a structural matrix decomposition method based on Data-Cluster-Data (DCD) random walk. DCD goes beyond matrix factorization and has a solid probabilistic interpretation by forming the approximation with cluster assigning probabilities only. Besides, the Kullback-Leibler divergence adopted by DCD is advantageous in handling sparse similarities for cluster analysis.

Multiplicative update algorithms have been commonly used for optimizing NMF objectives, since they naturally maintain the nonnegativity constraint of the factorizing matrix and require no user-specified parameters. In this work, an adaptive multiplicative update algorithm is proposed to increase the convergence speed of QNMF objectives.

Initialization conditions play a key role in cluster analysis. In this thesis, a comprehensive initialization strategy is proposed to improve the clustering performance by combining a set of base clustering methods. The proposed method can better accommodate clustering methods that need a careful initialization such as the DCD.

The proposed methods have been tested on various real-world datasets, such as text documents, face images, protein, etc. In particular, the proposed approach has been applied to the cluster analysis of emotional data.

Keywords Nonnegative Matrix Factorization, Cluster Analysis, Multiplicative Update Rule, Affective Computing, Image Emotion

ISBN (printed) 978-952-60-5828-3 **ISBN (pdf)** 978-952-60-5829-0

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 189 **urn** http://urn.fi/URN:ISBN:978-952-60-5829-0

Preface

The work presented in this thesis has been carried out at the Department of Information and Computer Science in Aalto university School of Science during the years 2008-2014. This work has received funding from the EU FP7 project Personal Information Navigator Adapting Through Viewing (PinView) from 2008 to 2010, and from the Academy of Finland in the project Finnish Center of Excellence in Computational Inference Research (COIN) from 2011 to 2014.

I am indeed indebted to my supervisor, Professor Erkki Oja. Thank you for your guidance and financial support, especially when the project funding ended three years ago. I truly thank my instructor Dr. Zhirong Yang, who is also a good friend of mine. You actually lead me into the world of scientific research, from polishing ideas and deducing formulae to editing papers and writing review responses. Thank you so much for the consistent academic instruction and great patience to me in all these years. Also, I have benefited a lot from the collaboration with Dr. Mehmet Gönen and Professor Timo Honkela. Thank you for those thought-provoking discussions and fruitful advices.

I would like to thank my previous instructor Dr. Jorma Laaksonen for bringing me into the image group. I have learned a lot from you. I would also like to thank other group members, Mats Sjöberg, Dr. Markus Koskela, Dr. Ville Viitaniemi, and Xi Chen, the only female doctoral student in the group. Mats, thank you for always sparing time to help me in solving the various computer problems, and I wish you and Xi a great success in your upcoming Ph.D. defense. Besides, I want to thank secretaries Minna Kauppila, Leila Koivisto, and Tarja Pihamaa for arranging me conference trips and other practical issues of the department in these years. You have made my life here easy and comfortable.

I would like to thank the pre-examiners of this thesis, Research Scientist

Dr. Rafal Zdunek and Associate Professor Dr. Morten Mørup, for their detailed examination and valuable feedbacks that truly helped in making this thesis better.

My family has given me strong and endless support for all these years in Finland. I sincerely thank my father Lianju Zhang and my mother Lili He, for encouraging me seeing the world outside, and for your visions on my study and career. I also thank my parents and my parents-in-law, Desheng Li and Shuxia Shi, for buying us a very nice apartment here. I really owe you for that. My biggest thank goes to my wife, Lina. You quit being the flight attendant in Korea and came to Finland for me. Thank you for the countless happy moments you have brought to me, for your selfless love and understanding during my difficult times, and for giving birth to our dearest son, Taile, who has made my life truly meaningful!

Espoo, Finland, September 11, 2014,

He Zhang

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Zhirong Yang, He Zhang, Zhijian Yuan, and Erkki Oja. Kullback-Leibler divergence for nonnegative matrix factorization. In *Proceedings of 21st International Conference on Artificial Neural Networks (ICANN)*, pages 250–257, Espoo, Finland, June 2011.

II He Zhang, Tele Hao, Zhirong Yang, and Erkki Oja. Pairwise clustering with t-PLSI. In *Proceedings of 22nd International Conference on Artificial Neural Networks (ICANN)*, pages 411–418, Lausanne, Switzerland, September 2012.

III Zhirong Yang, He Zhang, and Erkki Oja. Online Projective Nonnegative Matrix Factorization for large datasets. In *Proceedings of 19th International Conference on Neural Information Processing (ICONIP)*, pages 285–290, Doha, Qatar, November 2012.

IV He Zhang, Zhirong Yang, and Erkki Oja. Adaptive multiplicative updates for projective nonnegative matrix factorization. In *Proceedings of 19th International Conference on Neural Information Processing (ICONIP)*, pages 277–284, Doha, Qatar, November 2012.

V He Zhang, Zhirong Yang, and Erkki Oja. Adaptive Multiplicative Updates for quadratic nonnegative matrix factorization. *Neurocomputing*, 134: 206–213, 2014.

VI He Zhang, Zhirong Yang, and Erkki Oja. Improving cluster analysis by co-initializations. *Pattern Recognition Letters*, 45: 71–77, 2014.

VII He Zhang, Teemu Ruokolainen, Jorma Laaksonen, Christina Hochleitner, and Rudolf Traunmüller. Gaze-and speech-enhanced content-based image retrieval in image tagging. In *Proceedings of 21st International Conference on Artificial Neural Networks (ICANN)*, pages 373–380, Espoo, Finland, June 2011.

VIII He Zhang, Eimontas Augilius, Timo Honkela, Jorma Laaksonen, Hannes Gamper, and Henok Alene. Analyzing emotional semantics of abstract art using low-level image features. In *Proceedings of 10th International Symposium on Intelligent Data Analysis (IDA)*, pages 413–423, Porto, Portugal, October 2011.

IX He Zhang, Zhirong Yang, Mehmet Gönen, Markus Koskela, Jorma Laaksonen, and Erkki Oja. Affective abstract image classification and retrieval using multiple kernel learning. In *Proceedings of 20th International Conference on Neural Information Processing (ICONIP)*, pages 737–746, Daegu, South Korea, November 2013.

List of Symbols and Abbreviations

Symbols

α, β	regularization parameters
ϵ	step size parameter
η	exponent parameter
$\nabla, \nabla^+, \nabla^-$	gradient, positive gradient, negative gradient
λ	tradeoff parameter
$\mathbf{A}$	similarity matrix
$\mathbf{H}$	coefficient matrix
$\mathbf{I}$	identity matrix
$\mathbf{U}$	binary cluster indicator matrix
$\mathbf{W}$	basis matrix or cluster indicator matrix
$\mathbf{X}, \mathbf{x}_i$	data matrix, data vector
$\odot$	Kronecker (element-wise) product
$\oslash$	element-wise division
$\widehat{\mathbf{X}}$	approximation matrice
G, V, E	graph, node, edge
J	objective (cost) function
m, n	dimensionality, number of samples
r	matrix rank or number of clusters
G(-,-)	auxiliary function

Abbreviations

1-SPEC	1-Spectral Ratio Cheeger Cut
AIC	Akaike's Information Criterion
ALS	Alternating Least Squares
ANLS	Alternating Nonnegative Least Squares
ANN	Artificial Neural Networks
ART	Adaptive Resonance Theory
BIC	Bayes Information Criterion
BIRCH	Balanced Iterative Reduction and Clustering using Hierarchies
CBIR	Content-Based Image Retrieval
CURE	Clustering Using Representatives
DCD	Data-Cluster-Data random walk
EEG	Electroencephalogram
EM	Expectation-Maximization
ESIR	Emotional Semantic Image Retrieval
EU	Euclidean distance
GA	Genetic Algorithms
GGA	Genetically Guided Algorithm
GKA	Genetic K-means Algorithm
IAPS	International Affective Picture System
IS	Itakura-Saito divergence
KKT	Karush-Kuhn-Tucker condition
KL	Kullback-Leibler divergence
KNN	K-Nearest-Neighbor
LDA	Latent Dirichlet Allocation

LLE	Locally Linear Embedding
LSD	Left Stochastic Matrix Decomposition
MDL	Minimum Description Length criterion
MDS	Multidimensional Scaling
MKL	Multiple Kernel Learning
ML	Maximum Likelihood
MML	Minimum Message Length criterion
NMD	Nonnegative Matrix Decomposition
NMF	Nonnegative Matrix Factorization
NMFR	Nonnegative Matrix Factorization using Graph Random Walk
NMI	Normalized Mutual Information
NP-hard	Non-deterministic Polynomial-time hard
NSC	Nonnegative Spectral Clustering
ONMF	Orthogonal tri-factor Nonnegative Matrix Factorization
PCA	Principle Component Analysis
PLSI	Probabilistic Latent Semantic Analysis
PMF	Positive Matrix Factorization
PNMF	Projective Nonnegative Matrix Factorization
QNMF	Quadratic Nonnegative Matrix Factorization
SOFM	Self-Organizing Feature Maps
SOM	Self-Organizing Maps
SVD	Singular Value Decomposition
SVMs	Support Vector Machines
WNMF	Weighted Nonnegative Matrix Factorization

1. Introduction

1.1 Motivation and scope

Matrix decomposition is a fundamental theme in algebra with both scientific and engineering significance. For example, a large data matrix can be approximately factorized into several low-rank matrices. Two popular factorization methods are Singular Value Decomposition (SVD) [65] and Principle Component Analysis (PCA) [91] that are extensively used for data analysis. Many real-world data are nonnegative and the corresponding hidden components convey physical meanings only when the nonnegative condition holds. The factorizing matrices in SVD or PCA can have negative entries, which makes it hard or impossible to obtain physical interpretations from the factorizing results.

Nonnegative Matrix Factorization (NMF) [109] imposes the nonnegativity constraint on some of the factorizing matrices. When all involved matrices are constrained to be nonnegative, NMF allows only additive but not subtractive combinations during the factorization. Such nature can result in parts-based representation of the data, which can discover the hidden components that have specific structures and physical meanings.

Originally, the NMF approximation is factorized into two nonnegative matrices based on either Euclidean distance or I-divergence. Actually there are many other divergence measurements, and the factorization can take many different forms. Besides the nonnegativity constraint, NMF has been extended by incorporating other constraints and regularizations, such as stochasticity and orthogonality, on the factorizing matrices in order to enhance the capability to find true structures in data. Many numerical algorithms have been developed to optimize NMF objectives, among which multiplicative update algorithms are commonly utilized. The mul-

tiplicative algorithms can automatically maintain the nonnegativity constraint and require no user-specified parameters.

NMF has found a variety of applications in, for example, image processing [109, 83], text mining [109, 186], sound or music analysis [161, 54], bioinformatics [24], etc., among which NMF is mainly used for analyzing multivariate data, i.e. working on features. Recently, NMF has been extended to handle the graph input or similarity matrix between data points, i.e. grouping samples [117]. Actually, NMF has a close relationship with the classical k-means clustering [44]. In practice, NMF is easy to implement. These merits of NMF thus give the motivation of this thesis to study NMF methods with their application in data clustering tasks.

This thesis presents advances in NMF with the application in cluster analysis. Cluster analysis such as k-means is not a linear problem, and thus using standard NMF for clustering is an approximation. Therefore, the thesis reviews a class of higher order NMF methods, called Quadratic Nonnegative Matrix Factorization (QNMF) [191], where some factorizing matrices occur twice in the approximation. In addition to k-means clustering, QNMF can be applied to various other learning problems, such as graph matching and bi-clustering. Two important special cases of QNMF are Projective NMF (PNMF) [194] and NMF based on graph Random Walk (NMFR) [187], both of which can yield a highly orthogonal factorizing matrix desired for cluster analysis.

For some other problems, especially probabilistic models, even QNMF is not enough, but more complex matrix factorizations must be used. The thesis reviews a structural decomposition technique based on Data-Cluster-Data (DCD) random walk. NMF and QNMF are restricted to the scope of matrix factorization, whereas DCD goes beyond matrix factorization since the decomposition of the approximating matrix includes operations other than matrix product. In particular, DCD directly learns the cluster assigning probabilities as the only decomposing matrix. Besides, DCD adopts the Kullback-Leibler divergence as the approximation error measure that takes into account sparse similarities between samples, which is advantageous for clustering large-scale manifold data.

It is known that multiplicative update rules may suffer from slow convergence [62]. To improve the optimization speed, we present an adaptive multiplicative update algorithm, where a constant exponent in the update rule is replaced with a varied one. The adaptive algorithm can increase the convergence speed of QNMF for various applications while maintain

the monotonic decrease of their objective functions.

In addition to a fast convergence, initialization often has a great impact on the clustering accuracy. This thesis presents a co-initialization strategy, where a set of diverse clustering methods provide initializations for each other to improve the clustering results. We have empirically shown that the clustering performance can be greatly improved by using more comprehensive co-initialization strategies.

The presented methods have been tested on a variety of real-world datasets, including facial images, textual documents, and protein data, etc. Specifically, the thesis will present an experimental study on clustering emotional data, where the presented approaches can achieve improved clustering performance over other existing clustering methods.

1.2 Contributions of the thesis

This thesis presents advances in nonnegative matrix decomposition with application to cluster analysis. Its major contributions are:

- A new class of NMF methods called Quadratic Nonnegative Matrix Factorization (QNMF) is reviewed, where some factorizing matrices occur twice in the approximation. The properties of QNMF are discussed.

- An adaptive multiplicative update scheme is proposed for QNMF, where the constant exponent in the update rules is replaced by an variable one to accelerate the convergence speed of QNMF while its monotonic objective decrease is still maintained.

- A novel nonnegative low-rank approximation clustering method is reviewed, where the approximation is formed by only cluster assigning probabilities based on Data-Cluster-Data (DCD) random walk. DCD goes beyond matrix factorization and belongs to the class of probabilistic clustering methods.

- A co-initialization strategy is proposed, where a set of base clustering methods provide initializations for each other to improve clustering results. A hierarchy of initializations is presented, where a higher level can better facilitate methods that require careful initialization such as the DCD.

- A gaze-and speech-enhanced Contented-Based Image Retrieval System is proposed, and an emerging research area called Emotional Semantic Image Retrieval is introduced, where low-level generic image features are proposed for describing people's high-level affective feelings evoked by viewing images within affective image classification and retrieval tasks.

- An experimental study is performed, where the proposed co-initialization approach is applied to the cluster analysis of a widely-used emotional image dataset. Our approach has been compared with several other existing clustering methods.

1.3 Author's contribution in the publications

The author's contributions in the publications are described as follows:

Publication I studies the replacement of the I-divergence with the original Kullback-Leibler (KL-) divergence for NMF, and presents a projective gradient algorithm for NMF. The author participated in running the experiments and in writing the paper.

In Publication II, the author proposed a pairwise clustering algorithm by generalizing Probabilistic Latent Semantic Analysis (PLSI) to $t-$expo -nential family based on a criterion called $t-$divergence. The proposed method can improve clustering performance in purity for certain datasets. The author formulated the learning objective, developed the Majorization-Minimization algorithm, ran all the experiments, and wrote a major part of the paper.

Publication III provides an online Projective Nonnegative Matrix Factorization (PNMF) algorithm for handling large-scale datasets. The empirical studies on synthetic and real-world datasets indicate that the online algorithm runs much faster than the existing batch version. The author participated in designing the experiments and in writing the paper.

The original PNMF optimization algorithm can not guarantee the theoretical convergence during the iterative learning process. In Publication IV, the author proposed an adaptive multiplicative update algorithm for PNMF which not only ensures the theoretical convergence but also significantly accelerates its convergence speed. An adaptive exponent scheme is adopted by the author to replace the old unitary one. The author also de-

rived two new multiplicative update rules for PNMF based on the squared Euclidean distance and the I-divergence, performed all the numerical experiments, and wrote a major part of the paper.

In Publication V, the author extended the method proposed in Publication IV by generalizing the adaptive exponent scheme to Quadratic Nonnegative Matrix Factorization (QNMF). The author claimed that the proposed method is general and thus can be applied to QNMF with a variety of factorization forms and with the most commonly used approximation error measures. In addition to PNMF, the author has applied the adaptive scheme to two other special cases of QNMF, i.e. bi-clustering and estimation of hidden Markov chains. The extensive experimental results show that the new method is effective in these applications on both synthetic and real-world datasets. The author was responsible for both the theoretical and empirical contributions, as well as writing the article.

Publication VI discusses a comprehensive initialization strategy for improving cluster analysis. The author proposed a co-initialization method, where a set of different clustering methods provide initializations for each other to boost their clustering performance. The author also presented an initialization hierarchy, from simple to comprehensive. The extensive empirical results show that a higher-level initialization often gives better clustering results, and the proposed method is especially effective for advanced clustering methods such as the Data-Cluster-Data (DCD) decomposition technique. The author was responsible for conducting all the empirical analysis and writing a major part of the article.

Publication VII presents a novel gaze-and speech-enhanced Content-Based Image Retrieval (CBIR) system. The author recruited 18 users to evaluate the system in an image tagging task. Both the qualitative and quantitative results show that using gaze and speech as relevance feedback can improve the accuracy and the speed of finding wrongly-tagged images. The author was responsible for setting up the CBIR system, collecting each user's gaze and speech data, and recording the user's experience feedback after each evaluation. The author also conducted the numerical analysis and wrote a major part of the paper.

In Publication VIII, the author studied human's emotions evoked by viewing abstract art images within a classification framework. The author proposed to utilize generic low-level color, shape, and textual features for describing people's high-level emotions. The empirical results show that people's emotions can be predicted to certain extent using rather low-

level image features. The author created the abstract image dataset, conducted the online user survey participated by 20 test subjects, performed the empirical analysis, and wrote the whole paper.

Publication IX extends Publication VIII by conducting affective classification and retrieval of abstract art images using the Multiple Kernel Learning (MKL) framework. The experimental results on two abstract image datasets demonstrate the advantage of the MKL framework for image affect detection in terms of feature selection, classification performance, and interpretation. The author performed all the numerical experiments and wrote the whole paper.

1.4 Organization of the thesis

This thesis consists of an introductory part and nine publications. The structure of the thesis is organized as follows:

After this introduction in Chapter 1, Chapter 2 gives a brief review on the basic Nonnegative Matrix Factorization (NMF) method, algorithms, constraints, and its applications. Chapter 3 summarizes well-known clustering approaches. In Chapter 4, we first review existing clustering methods with NMF. We then review advances in NMF including Quadratic Nonnegative Matrix Factorization (QNMF) and a structural matrix decomposition technique based on Data-Cluster-Data (DCD) random walk. A novel adaptive multiplicative update algorithm is presented for increasing the convergence speed for QNMF. Chapter 5 presents a co-initialization approach for improving the performance of clustering methods such as the DCD. In Chapter 6, an experimental study is presented, in which we apply the proposed approach to clustering emotional data. Finally in Chapter 7, we conclude the thesis and discuss potential future directions.

2. Nonnegative Matrix Factorization

Nonnegative Matrix Factorization (NMF) is a popular matrix decomposition method with various applications in e.g. machine learning, data mining, pattern recognition, and signal processing. The nonnegativity constraints have been shown to result in parts-based representation of the data, and such additive property can lead to the discovery of data's hidden structures that have meaningful interpretations. In this chapter, we review the related work of NMF, including its algorithms, constraints, and applications.

2.1 Introduction

In linear algebra, a Matrix Factorization (MF) is a decomposition of a matrix into a product of matrices. Let the input data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ contain n data vectors of dimensionality m, $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_r)$, and $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_n)$. To factorize matrix $\mathbf{X}$ into the product of matrices $\mathbf{W}$ and $\mathbf{H}$, one can write:

$$\mathbf{X} = \mathbf{WH}. \tag{2.1}$$

In conventional MF, both the input matrix $\mathbf{X}$ and the factorized matrices $\mathbf{W}$ and $\mathbf{H}$ can contain either positive or negative entries.

The idea of Nonnegative Matrix Factorization (NMF) originated from the work by Paatero and Tapper in 1994 [142], in which they introduced a factor analysis method called Positive Matrix Factorization (PMF). Given an observed positive data matrix $\mathbf{X}$, PMF solves the following weighted factorization problem with nonnegativity constraints:

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} ||\mathbf{A} \odot (\mathbf{X} - \mathbf{WH})||_F, \tag{2.2}$$

where $||\cdot||_F$ denotes Frobenius norm, $\odot$ denotes Hadamard (element-wise) product, $\mathbf{A}$ is the weighting matrix, and $\mathbf{W}$, $\mathbf{H}$ are factor matrices that are

constrained to be nonnegative. The authors in [142] proposed an alternating least squares (ALS) algorithm by minimizing Eq. 2.2 with respect to one matrix while keeping the other constant.

Lee and Seung's milestone work [109] has made NMF attract more research attentions and gain more applications in various fields. Given a nonnegative input data matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$, NMF finds two nonnegative matrix $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ such that

$$\mathbf{X} \approx \mathbf{WH}. \tag{2.3}$$

The rank r is often chosen so that $r < \min(m, n)$, An appropriate selection of the value r is critical in practice, but its choice is usually problem dependent.

Let us write $\mathbf{x}_i \approx \mathbf{W}\mathbf{h}_i = \sum_{k=1}^{r} \mathbf{w}_k \cdot h_{ki}$. One can see that NMF approximates each nonnegative input data vector in $\mathbf{X}$ by an additive linear combination of r nonnegative basis columns in $\mathbf{W}$, with nonnegative coefficients in the corresponding column in $\mathbf{H}$. Therefore the matrix factor $\mathbf{W}$ is usually regarded as the basis matrix, the factor $\mathbf{H}$ as the coefficient matrix, and the product term $\mathbf{WH}$ is called the compressed version of the $\mathbf{X}$ or the approximating matrix of $\mathbf{X}$. As illustrated in [109], the additive nature of NMF can often generate parts-based data representation that conveys physical meanings .

2.2 NMF algorithms

Many numerical algorithms have been developed to solve the NMF problem [36]. Generally, the algorithms can be divided into three major classes: multiplicative update algorithms, projected gradient algorithms, and alternating nonnegative least squares algorithms.

2.2.1 Multiplicative update algorithms

The multiplicative update algorithm originates from the work by Lee and Seung [110], where they proposed to solve the NMF in Eq. 2.3 by minimizing two criteria: (1) the squared Euclidean (EU) distance $D_{EU}(\mathbf{X}||\mathbf{WH}) = \sum_{ij} \left(X_{ij} - (WH)_{ij}\right)^2$ and (2) the generalized Kullback-Leibler (KL) divergence $D_{KL}(\mathbf{X}||\mathbf{WH}) = \sum_{ij} \left(X_{ij} \log \frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij}\right)$. For example, the update rules based on squared EU distance are given by

$$W_{ia} \leftarrow W_{ia} \frac{\left(XH^T\right)_{ia}}{\left(WHH^T\right)_{ia}}, \quad H_{a\mu} \leftarrow H_{a\mu} \frac{\left(W^T X\right)_{a\mu}}{\left(W^T WH\right)_{a\mu}}. \tag{2.4}$$

Lee and Seung utilized the gradient descent to obtain the above multiplicative update algorithm by choosing a smart step size. For the implementation purpose (e.g, using Matlab), a small constant in each update rule is added to the denominator to avoid division by zero.

NMF algorithms that use multiplicative updates are characterized in Algorithm 1. ∇ denotes the gradient of NMF objective with respect to $\mathbf{W}$ or $\mathbf{H}$; ∇^+ and ∇^- are the positive and (unsigned) negative parts of ∇ respectively, i.e. $\nabla = \nabla^+ - \nabla^-$; $\oslash$ denotes the element-wise division.

Algorithm 1 Basic Multiplicative Update Algorithm for NMF

Initialize $\mathbf{W} \geq 0$ as a $m \times r$ random matrix

Initialize $\mathbf{H} \geq 0$ as a $r \times n$ random matrix

repeat

$\mathbf{W} = \mathbf{W} \odot \left(\nabla^-_{\mathbf{W}} \oslash \nabla^+_{\mathbf{W}}\right)$

$\mathbf{H} = \mathbf{H} \odot \left(\nabla^-_{\mathbf{H}} \oslash \nabla^+_{\mathbf{H}}\right)$

until converged or the maximum number of iterations is reached

Recently, Yang and Oja [189] presented a unified principle for deriving NMF multiplicative update algorithms with theoretical monotonicity guarantee for a large number of objectives under various divergence measures. Although Lee and Seung [110] claimed their algorithm converges to a local minimum, many researchers (e.g. [66, 119]) later revealed that their algorithm can only keep the non-increasing property of the objective. Rigorous proof of convergence to stationary points still lacks in general.

Another issue of conventional multiplicative update algorithms is that their convergence speed is relatively slow, compared with alternatives such as the projected gradient algorithms and the alternating nonnegative least squares algorithms discussed below. In [120], Lin demonstrated that one of his proposed methods converges faster than the algorithm by Lee and Seung [110]. Gonzalez and Zhang [66] accelerated Lee and Seung's algorithm by adding extra step-length parameters to rows of $\mathbf{W}$ and columns of $\mathbf{H}$. In Publication IX and Publication V, an adaptive exponential scheme was adopted instead of the constant one for speeding up the convergence of multiplicative updates.

2.2.2 Projected gradient algorithms

The second class of NMF algorithms has update rules of the form given in Algorithm 2. The step size parameters $\epsilon_{\mathbf{W}}$ and $\epsilon_{\mathbf{H}}$ are often problem dependent, and are selected alone the negative gradient direction. In [83],

Algorithm 2 Basic Projected Gradient Algorithm for NMF

Initialize $\mathbf{W} \geq 0$ as a $m \times r$ random matrix

Initialize $\mathbf{H} \geq 0$ as a $r \times n$ random matrix

Choose parameters $\epsilon_{\mathbf{W}}$ and $\epsilon_{\mathbf{H}}$

repeat

$\mathbf{W} = \mathbf{W} - \epsilon_{\mathbf{W}} \odot \nabla_{\mathbf{W}}$

Set all negative elements of $\mathbf{W}$ to be 0

$\mathbf{H} = \mathbf{H} - \epsilon_{\mathbf{H}} \odot \nabla_{\mathbf{H}}$

Set all negative elements of $\mathbf{H}$ to be 0

until converged or the maximum number of iterations is reached

Hoyer set the initial step size values to be 1 and multiplied them by one-half at each subsequent iteration. However, this additive setup can not prevent the entries of the updated matrices $\mathbf{W}$ and $\mathbf{H}$ from getting negative values. A common practice used by many projected gradient algorithms (see e.g. Hoyer [83], Chu et al. [34], and Pauca et al. [145] etc.) is a simple projection step, where, after each update rule, the updated matrices are projected to the nonnegative orthant by setting all negative values to be zero.

The convergence of projected gradient algorithm depends on the choices of step size parameters $\mathbf{W}$ and $\mathbf{H}$. The extra step of nonnegativity projection makes the analysis even more difficult. In general, projected gradient methods can not guarantee the monotonic decrease of objective function.

2.2.3 Alternating nonnegative least squares algorithms

Alternating Nonnegative Least Squares (ANLS) algorithms are the third class of NMF algorithms. This class has a general update form described in Algorithm 3, where a least squares step is followed by another least squares step in an alternating manner. The advantage of ANLS algorithms lies in the fact that, although the NMF problem of Eq. 2.3 is not convex in both $\mathbf{W}$ and $\mathbf{H}$, it is convex in either $\mathbf{W}$ or $\mathbf{H}$. Therefore, one can fix one matrix and solve for the other matrix using a simple least squares method.

The alternating least squares algorithm was first used by Paatero and Tapper [142] for minimizing the PMF problem of Eq. 2.2. To ensure the nonnegativity, a truncation step is added after each least squares step to project all negative elements of factorizing matrices to be 0. This simple approach facilitates sparsity, and gives more flexibility than multiplica-

Algorithm 3 Basic Alternating Nonnegative Least Squares (ANLS) Algorithm for NMF

Initialize $\mathbf{W} \geq 0$ as a $m \times r$ random matrix
repeat
  Solve for $\mathbf{H}$ using equation $\mathbf{W}^T\mathbf{W}\mathbf{H} = \mathbf{W}^T\mathbf{X}$
  Set all negative elements of $\mathbf{H}$ to be 0
  Solve for $\mathbf{W}$ using equation $\mathbf{H}\mathbf{H}^T\mathbf{W}^T = \mathbf{H}\mathbf{X}^T$
  Set all negative elements of $\mathbf{W}$ to be 0
until converged or the maximum number of iterations is reached

tive update algorithms because it allows the iterative process to escape from a poor path.

For NMF based on Euclidean distance, the ANLS algorithms in general have nice optimization properties and converge faster than the multiplicative update approach [120]. However, the underlying distribution of least square measurement is Gaussian, which is not suitable for handling data with other types of distributions.

2.3 NMF with additional constraints or regularizations

For many real-world problems, the input $\mathbf{X}$ to be analyzed often has noise or other data degradations inside. To alleviate this issue, researchers have proposed combining various auxiliary constraints with NMF objectives to enforce the agreement between the factorization results and the expected physical interpretations. The forms of constraints are application dependent, which can be characterized by extending the NMF objective of Eq. 2.3 as follows:

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} ||\mathbf{X} - \mathbf{W}\mathbf{H})||_F^2 + \alpha J_1(\mathbf{W}) + \beta J_2(\mathbf{H}), \tag{2.5}$$

where $J_1(\mathbf{W})$ and $J_2(\mathbf{H})$ are penalty terms regarded as auxiliary constraints; α and β are small regularization parameters that balance the trade-off between the approximation error and the added constraints.

Smoothness constraints are often used to counteract the noise in data. For example, Pauca et al. [145] presented an NMF algorithm by incorporating additional constraints as

$$J_1(\mathbf{W}) = ||\mathbf{W}||_F^2 \text{ and } J_2(\mathbf{H}) = ||\mathbf{H}||_F^2 \tag{2.6}$$

in order to penalize $\mathbf{W}$ and $\mathbf{H}$ solutions of large Frobenius norm and thus

enforce the smoothness in both matrices.

Sparsity constraints are often used in situations where only a few features are enough to represent data vectors and/or an emphasis on the extraction of local rather than global features. Cichochi et al. [35] achieved a sparse representation of data by setting

$$J_1(\mathbf{W}) = \sum_{ij} W_{ij} \text{ and } J_2(\mathbf{H}) = \sum_{jk} H_{jk}. \tag{2.7}$$

Hoyer [83] proposed a sparseness criterion by leveraging the relationship between the $\mathbf{L}1$ and $\mathbf{L}2$ norm:

$$sparseness(\mathbf{x}) = \frac{\sqrt{n} - (\sum |\mathbf{x}_i|)/\sqrt{\sum \mathbf{x}_i^2}}{\sqrt{n} - 1}, \tag{2.8}$$

where $\mathbf{x}$ denotes a given vector with dimension n. For instance, the sparseness criterion imposed on a $m \times k$ matrix $\mathbf{W}$ can be formulated as the following penalty term:

$$J_1(\mathbf{W}) = (\alpha||vec(\mathbf{W})||_2 - ||vec(\mathbf{W})||_1)^2, \tag{2.9}$$

where $\alpha = \sqrt{mk} - (\sqrt{mk} - 1)\alpha$ and $vec(\cdot)$ is an operator that transforms a matrix into a vector by stacking its columns. The sparseness in $\mathbf{W}$ is specified by setting α to a value between 0 and 1.

In addition to the smoothness and sparsity constraints, there are situations when certain prior knowledge about the application is known beforehand. In such cases, the prior information can be transformed to be auxiliary constraints for helping to better achieve the desired results. A good example is the semi-supervised NMF, which imposes the label information as constraints (see [113] for a recent survey). More examples using the different constraint forms discussed above will be given in the following NMF applications' section.

2.4 Initialization

Most NMF objectives are not convex and are thus sensitive to the initialization of factor matrices $\mathbf{W}$ and $\mathbf{H}$ (see e.g. [1, 201, 101]). A good initialization strategy can significantly relieve the convergence problem of NMF algorithms. Here the term *good* refers to the initialization strategy that leads to a rapid error reduction of an NMF objective with a fast algorithmic convergence speed [18].

Among the literature, random initialization has been commonly used, for instance in Lee and Seung's work [109], where one needs to run an

NMF algorithm several times with different initial matrices and picks the best solution. However, choosing the initial matrices randomly often gives a poor solution. Wild et al. [182] proposed a structured initialization approach for NMF, where they initialized the factor matrices using the cluster centroids of the Spherical k-means clustering [42]. Boutsidis et al. [18] described an initialization method based on two SVD processes to speed up the approximation error reduction. Other SVD-based initialization approaches include [108, 201, 101]. In Publication VI, we proposed a comprehensive initialization strategy to relieve the problem, where several algorithms provide initializations until convergence (see a detailed discussion in Chapter 5). These alternatives all demonstrated better performance than the random initialization approach.

2.5 Selected applications of NMF

2.5.1 Image processing

In the work by Lee and Seung [109], they demonstrated with a facial image dataset that NMF can be used to obtain a set of bases that correspond to the intuitive notion of facial parts such as eyes, nose, and mouth. Since then, many NMF algorithms have been developed for image processing tasks. It is argued that the nonnegativity of linear factorization is particularly suitable for analyzing image libraries that consist of images showing a composite object in many articulations and poses [109, 49]. The nonnegativity constraints lead to a parts-based representation as they allow only additive, not subtractive, combinations of the original data. More specifically, since $||\mathbf{X} - \mathbf{WH}||_F^2 = \sum_i (\mathbf{x}_i - \mathbf{W}\mathbf{h}_i)^2 = \sum_i (\mathbf{x}_i - \sum_k \mathbf{w}_k h_{ki})^2$, each column $\mathbf{x}_i$ of the nonnegative input $\mathbf{X}$ now represents a m dimensional (column-wise aligned) image vector; each column $\mathbf{w}_k \in \mathbb{R}^m$ is a basis element that corresponds a facial part (eyes or nose etc.); each column $\mathbf{h}_i \in \mathbb{R}^r$ is considered as a coefficient array representing the i-th image in the basis elements. Therefore, the NMF model can be interpreted that each image represented as a vector $\mathbf{x}_i$ is obtained by superimposing a collection of standard parts $\mathbf{w}_k$ with respective weighting coefficients $\mathbf{h}_i$. The resulting parts (pixel values) and the coefficients are all nonnegative. Figure 2.1 shows an illustrative example.

Since the additive parts learned by original NMF are not necessarily

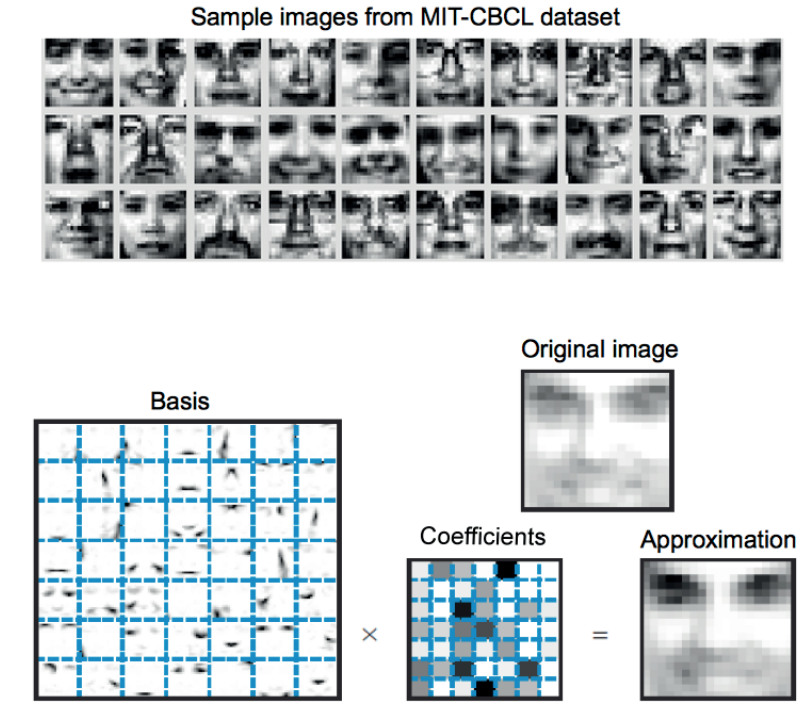


Figure 2.1. The above figure is taken from Lee and Seung's paper [109]. The training samples are from MIT-CBCL [2] image dataset that contains $2,429$ facial images, each consisting of 19×19 pixels. The basis vectors correspond to facial parts such as eyes, mouths, and noses, etc. NMF seeks to approximate an original input image by a linear combination of the (learned) basis weighted by their corresponding (learned) coefficients.

localized, several NMF methods have been proposed to enhance the localization. For example, Li et al. [116, 53] combined NMF with localization constraints for learning spatially localized representation of facial images. Guillamet et al. [70] presented a weighted NMF (WNMF) algorithm to reduce basis redundancies with a factorization form $\mathbf{XQ} \approx \mathbf{WHQ}$, where the weighting matrix $\mathbf{Q}$ is learned from the training data (newspaper images). In [71], Guillamet and co-authors introduced a probabilistic framework to compare PCA, NMF, and WNMF in the context of image patch classification. Hoyer [83] incorporated sparseness constraints on matrices $\mathbf{W}$ and $\mathbf{H}$ for improving the parts-based decompositions of facial images and natural image patches. Wang et al. [181] proposed a subspace method called Fisher NMF that encodes within-class and between-class scattering information for face recognition. Yuan and Oja [194] presented a projective NMF method with a factorization form $\mathbf{X} \approx \mathbf{W}\mathbf{W}^T\mathbf{X}$ to induce a stronger sparseness on basis $\mathbf{W}$ for facial image compression and

feature extraction. NMF has also been used for image hashing [132, 166] and image watermarking [140, 124]. More works of NMF related to image processing can be found, e.g., in [195, 143, 163, 77, 26, 67, 121].

2.5.2 Text mining

Besides the applications to images, Lee and Seung [109] utilized NMF for the semantic analysis of textual documents (articles in the encyclopedia) as well. For this application, $\mathbf{X}$ contains a corpus of documents, where X_{ij} is the number of times the i-th word in the vocabulary appears in the j-th document; each column $\mathbf{w}_k$ corresponds to a semantic feature or topic; each column $\mathbf{h}_i$ is a projection array for approximating the i-th document in $\mathbf{X}$. In each semantic feature, NMF groups together semantically related words. Each document is represented by additively combining several of these features.

NMF gives a natural choice for document clustering, since the coefficient matrix $\mathbf{H}$ can be directly used to determine the cluster membership of each document, i.e., assigning document $\mathbf{x}_i$ to cluster k if $k = \arg\max_k h_{ki}$, for $k = 1, \ldots, r$. For example, Xu et al. [186] showed with a document clustering task that NMF surpasses the latent semantic indexing and spectral clustering methods in terms of reliability and clustering accuracy. Pauca and co-authors [146, 157] conducted text mining by using a hybrid NMF method, where $\mathbf{W}$ is updated with gradient descent approach whereas $\mathbf{H}$ is updated with constrained least squares. Berry et al. [11] applied NMF for email surveillance. Ding et al. [48] proposed a 3-factor NMF having the form $\mathbf{X} \approx \mathbf{FSG}^T$ with orthogonality constraints on both $\mathbf{F}$ and $\mathbf{G}$. They showed that the 3-factor model can cluster words and documents simultaneously. A semi-supervised NMF model was proposed in [113]. Online NMF methods can be found in [176, 68]. Other related works using NMF for text mining include [117, 175, 47, 114, 174].

2.5.3 Music analysis

Another successful application area of NMF is sound source separation. In real-world audio signals, multiple sound sources are often mixed together within a single channel, and one needs to separate the mixed sounds for a better analysis and manipulation of audio data. Usually the separation is done by using prior knowledge of the sources, which results in highly complex systems. In contrast, NMF does the separation by find-

ing a nonnegative decomposition of the input signal $\mathbf{x}_t \approx \sum_{k=1}^{r} \mathbf{w}_k \cdot h_{kt}$, without using any other source-specific prior information than the nonnegativity. Each source is modeled as a sum of one or more components. The term *component* refers to a basis $\mathbf{w}_k$ and its time-varying gain h_{kt} for $t = 1, \ldots, T$. The input $\mathbf{x}_t$ denotes a magnitude or power spectrum vector in frame t, with T being the number of frames. In the case of music signals, each component usually represents a musically meaningful entity or parts of it (e.g., the sounds produced by a drum or piano).

Smaragdis and Brown [161] applied NMF for analyzing polyphonic music passages by following the basic NMF algorithms presented in [110]. Later, Smaragdis [160] presented a deconvolution version of NMF for extracting multiple sound sources produced by drums. Virtanen [170] combined NMF with temporal continuity and sparseness criteria for separating sound sources in single-channel music signals. Févotte et al. [54] applied NMF with Itakura-Saito (IS) divergence for piano music analysis, where they experimentally show that the scale invariance property of IS divergence is advantageous in the estimation of low-power components, such as note attacks. The authors in [12] presented the constrained NMF within a Bayesian Framework for polyphonic piano music transcription. Other works that use NMF for sound source separation include [80, 155, 171, 141, 81].

2.5.4 Computational biology

NMF has recently been utilized for the analysis and interpretation of biological data. NMF as an unsupervised method is particular useful when there is no prior knowledge of the expected gene expression patterns for a given set of genes. In this context, $\mathbf{X}$ denotes the gene expression data matrix that consists of observations on m genes from n samples; each column $\mathbf{w}_k$ defines a meta-gene; each column $\mathbf{h}_i$ represents the meta-gene expression pattern of the corresponding sample in $\mathbf{X}$.

A large amount of works focus on using NMF in the area of molecular pattern discovery, especially for gene and protein expression microarray studies. For instance, Kim and Tidor [99] applied NMF to cluster genes and predict functional cellular relationships in yeast with gene expression data. Brunet et al. [24] utilized NMF to detect cancer-related microarray data. Sparse versions of NMF were presented by Gao and Church [60], Kim and Park [98], for cancer-class discovery and gene expression data analysis. Pascual-Montano et al. [144] and Carmona-Saez et al. [27] con-

ducted two-way clustering of gene expression profiles using non-smooth NMF [143]. NMF has also used for applications such as protein fold recognition [137], cross-platform and cross-species characterization [165], and gene functional characterization [147, 169]. Other works of using NMF for computational biology include [177, 93, 31, 100, 61, 148, 178].

2.5.5 Other applications

In addition to the above applications, NMF has been used in many other areas, including Electroencephalogram (EEG) signal classification [32, 112, 111], financial data analysis [50, 150], remote sensing and object identification [145, 130, 90], as well as color and vision research [25].

3. Clustering

Clustering, or cluster analysis, plays an essential role in machine learning and data mining. Clustering aims to group data points according to certain similarity criterion, without knowing the data labeling information. This chapter gives an introduction on cluster analysis and provides a brief review on the well-known clustering methods. In the next chapter, we will go into details on using NMF for cluster analysis.

3.1 Introduction

Clustering is a combinatorial problems whose aim is to find the cluster assignment of data that optimizes certain objective. The aim of clustering is to group a set of objects in such a way that the objects in the same cluster are more similar to each other than to the objects in other clusters, according to a particular objective. Clustering belongs to the *unsupervised* learning scope that involves unlabeled data only, which makes it a more difficult and challenging problem than classification because no labeled data or ground truth can be used for training. Cluster analysis is prevalent in many scientific fields with a variety of applications. For example, image segmentation, an important research area of computer vision, can be formulated as a clustering problem (e.g. Shi and Malik [159]). Documents can be grouped by topics for efficient information retrieval (e.g. Hofmann [82]). Clustering techniques are also used for volatility analysis in financial markets (e.g. Lux and Marchesi [126]) and genome analysis in bioinformatics (e.g. Ben-Dor et al. [10]).

Many clustering algorithms have been published in the literature. These algorithms can be divided into two groups: *hierarchical* and *partitioning*. Hierarchical clustering algorithms recursively find nested clusters according to certain hierarchy, whereas partitioning clustering algorithms locate

all the clusters simultaneously without imposing any hierarchical structure. In the following section, we give a brief overview on some of the major clustering approaches.

3.2 Major clustering approaches

3.2.1 Hierarchical clustering

Hierarchical clustering organizes data into a cluster hierarchy or a binary tree known as *dendrogram*, according to the proximity matrix. The roof node of the dendrogram represents the whole dataset and each leaf node is regarded as a data object. Hierarchical clustering methods can be further divided into *agglomerative* (bottom-up) mode and *divisive* (top-down) mode. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more of the most similar clusters to form a cluster hierarchy, whereas a divisive clustering starts with a single cluster containing all data points and recursively splits the most appropriate cluster into smaller clusters. The process continues until the stopping criterion (e.g. the requested number of clusters) is achieved. To split or merge clusters, a linkage metric needs to be defined for measuring the distance between two clusters. The most popular metrics are single-linkage, complete linkage, and average linkage (see surveys in [135] and [40]), all of which can be derived from the Lance-Williams updating formula [106].

Typical agglomerative clustering algorithms include CURE (Clustering Using Representatives) by Guha et al. [69], CHAMELEON by Karypis et al. [94], and BIRCH (Balanced Iterative Reduction and Clustering using Hierarchies) by Zhang et al. [199]. CURE represents a cluster by a fixed set of points scattered around it, which makes it possible to handle clusters of arbitrary shapes. CHAMELEON uses the connectivity graph G sparsified by a K-nearest neighbor model: only the edges of K most similar points to any given point are preserved, the rest are dropped. BIRCH is designed for clustering very large databases, where it represents data by its statistics summaries instead of using original data features. For divisive clustering, a good example is the Principle Direction Divisive Partitioning) algorithm presented by Boley [16], in which the author applied SVD to hierarchical divisive clustering of document collections.

Hierarchical clustering facilitates data exploration on different levels of

granularity. However, most of the hierarchical algorithms do not reconsider clusters once constructed, which means that they are not capable of correcting previous improper cluster assignments.

3.2.2 Partitioning relocation clustering

Compared with hierarchical clustering, partitioning clustering finds all the clusters simultaneously and imposes no hierarchical structure. Perhaps the most well-known and the simplest partitioning algorithm is k-means presented by Lloyd [123]. The k-mean algorithm finds a partition by minimizing the distance between the empirical mean of a cluster and the points inside the cluster. It consists of two-step major iterations: (1) re-assign all the points to their nearest centroids, and (2) re-compute centroids of newly assembled clusters. The iterations continues until a stopping criterion (e.g. no changes on cluster assignments) is achieved. Despite of the popularity, the k-means algorithm can not choose the best number of clusters K by itself. Also, k-means only facilitates the detection of spherical shaped clusters due to its usage of Euclidean distance in most cases, which makes it rather sensitive to initializations and outliers.

Many extensions of k-means have been developed, such as ISODATA by Ball and Hall [7] and FORGY by Forgy [57]. Other representative examples include *fuzzy c-means* by Dunn [51], *kernel k-means* by Scholkopf et al. [156], and *k-medoid* by Kaufman and Rousseeuw [95]. Fuzzy c-means assigns each data point a membership of multiple clusters, i.e., making a “soft" assignment rather than a “hard" assignment of k-means. Kernel k-means leverages the power of kernels to detect arbitrary shaped clusters. In k-medoid methods, a cluster is represented by one of its points instead of the geometric centroid, which makes it robust against outliers.

Unlike traditional hierarchical clustering methods, in which clusters are revisited after being constructed, relocation clustering methods can gradually improve the intermediate results to achieve high quality clusters.

3.2.3 Generative models

In clustering based on generative models, each data object is assumed to be independently generated from a *mixture* model of several probability distributions (see McLachlan and Basford [129]). If the distributions are known, finding the clusters of a given dataset is equivalent to estimating the parameters of several underlying models. Multivariate Gaussian

mixture model is often utilized due to its analytical tractability, where the Maximum Likelihood (ML) estimation is commonly used for estimating parameters. However, the closed-form solution of maximizing the likelihood does not exist for mixture models. Dempster et al. [41] proposed the Expectation-Maximization (EM) algorithm, which is the most popular approach for approximating the ML estimates. EM iterates two steps: the E-step computes the expectation of the complete data log-likelihood, and the M-step selects a parameter that maximizes the log-likelihood. The process continues until the log-likelihood converges. An important observation by Celeux and Govaert [29] has shown that the classification EM algorithm under a spherical Gaussian mixture assumption is equivalent to the classical k-means algorithm. Therefore, EM is sensitive to initial parameter selection and is slow to converge. Detailed descriptions regarding EM algorithm and extensions can be found in [128].

Hofmann [82] proposed a clustering method named Probabilistic Latent Semantic Indexing (PLSI) based on a statistical latent class model. PLSI was further developed into a more comprehensive model called Latent Dirichlet Allocation (LDA) by Blei et al. [15], where a Bayesian treatment is applied to improve the mixture models for data clustering. In Publication II, we presented a pairwise clustering algorithm by generalizing PLSI to $t-$exponential family based on a criterion called $t-$divergence.

Clustering based on generative models provides an easy and straightforward interpretation on cluster system.

3.2.4 Graph-based partitioning

Graph-based partitioning has become an active research field in recent years, thanks to the concepts and properties of graph theories. Let $G = (V, E)$ denote a weighted graph, where the nodes V represent the data points and the edges E reflect the proximities between each pair of data points. Early works (e.g. [184]) concentrate on finding the *minimum cut* of a graph, that is, to partition the nodes V into two subsets A and B such that the cut size, defined as the sum of the weights assigned to the edges connecting between nodes in A and B, is minimized. However, methods based on the minimum cut often result in un-balanced data clusters.

Hagen and Kahng [72] proposed the ratio cut algorithm for balancing the cluster size, i.e., the number of data points in a cluster. Shi and Malik [159] presented the normalized cut algorithm, which measures both the total dissimilarity between the different groups and the total similarity

within the groups. A multi-class version was later proposed by Yu and Shi [193]. Ng et al. [136] presented another variant by deriving data representation from the normalized eigenvectors of a kernel matrix, whereas Belkin and Niyogi [8] constructed data representation based on the eigenvectors of a graph Laplacian.

Graph-based partitioning based on (generalized) eigen decomposition of a matrix is simple to implement and can be solved efficiently by standard linear algebra methods. The weighted or similarity graph can be easily computed no matter if the data is numerical or categorical. An open problem for graph partitioning methods is: how to compute the similarity graph or how to select the similarity function (see related discussions by Luxburg [172]).

3.2.5 Large-scale and high-dimensional data clustering

Many techniques have been developed for clustering large-scale databases. These techniques can be divided into: incremental clustering, data squashing, and data sampling. Incremental clustering handles one data point at a time and then discards it, which is in contrast to most clustering algorithms that require multiple passes over data points before identifying the cluster centroids. Typical examples include the hierarchical clustering algorithm COBWEB proposed by Fisher [55], and the scaling clustering algorithms proposed by Bradley et al. [22]. Data squashing techniques scan data to compute certain data summaries (sufficient statistics) that are then used instead of the original data for clustering. An algorithm with high impact is BIRCH [199]. Data sampling methods subsample a large dataset selectively, and perform clustering over the smaller set. The resulting information is later transferred to the larger dataset. A typical example is CURE [69].

The term *curse of dimensionality* [9] is often used to describe the problems caused by high dimensional spaces. It is theoretically proved that the distance between the nearest points becomes indistinguishable from the distance to the majority of points when the dimensionality is high enough [13]. Therefore dimensionality reduction is an important procedure in cluster analysis to keep the proximity measures valid. Principal Component Analysis (PCA) [91] is a popular approach for reducing the dimension of multivariate data. Other typical approaches include projection pursuit [84], multidimensional scaling (MDS) [17], locally linear embedding (LLE) [152, 8], and semi-supervised dimensionality reduction [196].

A survey on dimension reduction techniques is given in [56].

3.2.6 Other clustering techniques

Neural networks-based clustering finds clusters based on Aritificial Neural Networks (ANN). The most typical method is Self-Organizing Maps (SOM) or Self-Organizing Feature Maps (SOFM), proposed by Kohonen [102]. SOM projects high-dimensional input data points onto a low-dimensional (usualy 2-dimensional) lattice structure while preserving the topological properties of the input space, which makes SOM become one of the few clustering methods that can be used as a helpful tool for visualizing high-dimensional data. One can see Kohonen's book [103] for more details on SOM and various variants. In addition to SOM, other ANN methods, such as Adaptive Resonance Theory (ART) presented by Carpenter et al. [28], have received much research attention as well. Further details related to ANN can be found in the tutorial by Jain et al. [89].

Nature-inspired clustering, or evolutionary techniques, is characterized by Genetic Algorithms (GA) (see Goldberg [64] for details). A set of evolutionary operators, usually the *selection*, *crossover*, and *mutation*, are iteratively applied to the population until the objective, called *fitness* function, satisfies the stopping condition. Hall et al. [73] proposed a Genetically Guided Algorithm (GGA) that can be considered as a general scheme for center-based (hard or fuzzy) clustering problems. Evolutionary techniques rely on user-defined parameters and have high computational complexity, which limits their applications in large-scale datasets. Some researches combine genetic techniques with other clustering methods for better clustering performance. For instance, Krishna and Murty [104] proposed a hybrid method called genetic k-means algorithm (GKA) that can converge to the best known optimum corresponding to the given data.

3.3 Evaluation measures

Effective evaluation measures are crucial for quantifying and comparing the performance of clustering methods. Generally there are three different categories of evaluation criteria: *internal*, *relative*, and *external* [88]. Internal criteria examine the resulting clusters directly from the original input data. Relative criteria compare several clustering structures, which

can be produced by different algorithms, and decide which one may best characterize the data to certain extent. External criteria has been commonly used; they measure the clustering performance by using the known information (often referred to as ground truth). Two widely used external criteria are

- *Purity*, (e.g. [48, 190]), defined as

$$\text{purity} = \frac{1}{n}\sum_{k=1}^{r}\max_{1\le l\le q} n_k^l,$$

  where n_k^l is the number of vertices in the partition k that belong to the ground-truth class l. Purity is easy to understand, as it can be interpreted in a similar way as the classification accuracy in supervised learning. However, purity has a drawback in that it tends to emphasize the large clusters.

- *Normalized Mutual Information* [164], defined as

$$\text{NMI} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{i,j}\log\left(\frac{n_{i,j}n}{n_i m_j}\right)}{\sqrt{\sum_{i=1}^{K} n_i \log\left(\frac{n_i}{n}\right)\sum_{j=1}^{K'} m_j \log\left(\frac{m_j}{n}\right)}},$$

  where K and K' respectively denote the number of clusters and classes; $n_{i,j}$ is the number of data points agreed by cluster i and class j; n_i and m_j denote the number of data points in cluster i and class j respectively; and n is the total number of data points in the dataset. NMI examines the quality of clusters from an information-theoretic perspective. Compared with purity, NMI tends to be less affected by the cluster sizes due to the normalization step given by its denominator, but it is not that intuitive as purity for people to interpret.

For a given clustering structure of the data, both purity and NMI give a value between 0 and 1, where a larger value in general indicates a better clustering performance. Other typical external criteria include, for instance, *Rand Index* [149] and *Adjusted Rand Index* [85].

4. Nonnegative matrix decomposition for clustering

NMF can often produce parts-based data representation. If we transpose the data matrix, the grouping can appear over samples, instead of features. This is a desired property for cluster analysis. In this chapter, we review two recent nonnegative matrix decomposition methods. One is a new class of NMF methods called Quadratic Nonnegative Matrix Factorization (QNMF). The other one is a matrix decomposition method based on Data-Cluster-Data (DCD) random walk.

4.1 Early NMF methods for clustering

Assume there are n samples to be grouped into r disjoint clusters, the cluster assignment can be represented by a binary indicator matrix $\mathbf{W} \in \{0,1\}^{n \times r}$, where $\mathbf{W}_{ik} = 1$ if the i-th sample is assigned to the k-th cluster and 0 otherwise.

Many existing clustering methods employ a non-convex objective function over the cluster indicator matrix $\mathbf{W}$, and directly optimizing $\mathbf{W}$ is difficult as the solution space is *discrete*, which usually leads to an NP-hard problem. Therefore, a relaxation to “soft" clustering is often required to obtain computationally efficient solutions.

NMF relaxes clustering problems by nonnegative low-rank approximation, in which the cluster indicator matrix $\mathbf{W}$ can be operated within a *continuous* space to ease the optimization process.

4.1.1 Related work

Originally, NMF was applied to analyzing vectorial data, i.e. extracting features. Later, NMF was extended to take graph or pairwise similarities as input in order to group samples (see [117]). An early example can be found in [186], where NMF was applied to clustering textual documents.

Actually, the clustering property of NMF was not well discussed in Lee and Seung's works [109, 110]. Ding et al. have shown that the basic NMF is equivalent to the classical k-means clustering under certain constraints [46], and that some NMF extensions with least square error measure are equivalent to kernel k-means and spectral clustering [44].

In the following, we briefly review some recent NMF-based clustering methods. In the subsequent sections, we will present our two methods based on nonnegative matrix decomposition.

- Orthogonal tri-factor NMF (ONMF), proposed by Ding et al. [48]. Given the nonnegative input matrix $\mathbf{X} \in \mathbb{R}_+^{m \times n}$, ONMF has a factorization form of $\mathbf{X} \approx \mathbf{WSH}$, where $\mathbf{W} \in \mathbb{R}_+^{m \times r}$, $\mathbf{S} \in \mathbb{R}_+^{r \times l}$, $\mathbf{H} \in \mathbb{R}_+^{l \times n}$, and solves the following optimization problem:

$$\underset{\mathbf{W} \geq 0, \mathbf{S} \geq 0, \mathbf{H} \geq 0}{\text{minimize}} ||\mathbf{X} - \mathbf{WSH}||_F^2, \text{ subject to } \mathbf{W}^T\mathbf{W} = \mathbf{I}, \ \mathbf{H}\mathbf{H}^T = \mathbf{I}. \quad (4.1)$$

  The tri-factorization form of ONMF makes it capable of grouping rows and columns of the input data matrix simultaneously. An important special case is that the input $\mathbf{X}$ contains a matrix of pairwise similarities, i.e., $\mathbf{X} = \mathbf{X}^T = \mathbf{A}$. In this case, $\mathbf{W} = \mathbf{H}^T \in \mathbb{R}_+^{n \times r}$, $\mathbf{S} \in \mathbb{R}_+^{r \times r}$, and the optimization problem becomes:

$$\underset{\mathbf{W} \geq 0, \mathbf{S} \geq 0}{\text{minimize}} ||\mathbf{A} - \mathbf{WSW}^T||_F^2, \text{ subject to } \mathbf{W}^T\mathbf{W} = \mathbf{I}. \quad (4.2)$$

- Nonnegative Spectral Clustering (NSC), proposed by Ding et al. [45]. NSC solves the normalized cut [159] by using a multiplicative update algorithm. Let $\mathbf{W} \in \mathbb{R}_+^{n \times r}$ be the cluster indicator matrix. NSC solves the following optimization problem:

$$\underset{\mathbf{W}^T\mathbf{DW} = \mathbf{I}, \ \mathbf{W} \geq 0}{\text{minimize}} -trace(\mathbf{W}^T\mathbf{AW}), \quad (4.3)$$

  where $\mathbf{D} = diag(d_1, \ldots, d_n)$ with $d_i = \sum_{j=1}^n A_{i,j}$. Unlike the normalized cut, where the solutions (eigenvectors) contain mixed signs, the nonnegative constraint on $\mathbf{W}$ makes the cluster assignment easy to interpret.

- Projective NMF (PNMF), proposed by Yuan and Oja [194]. PNMF tries to find a nonnegative projection matrix $\mathbf{P} \in \mathbb{R}_+^{m \times m}$ of rank r such that $\mathbf{X} \approx \mathbf{PX} = \mathbf{WW}^T\mathbf{X}$, where $\mathbf{W} \in \mathbb{R}_+^{m \times r}$. This equals to solve the following optimization problem:

$$\underset{\mathbf{W} \geq 0}{\text{minimize}} ||\mathbf{X} - \mathbf{WW}^T\mathbf{X}||. \quad (4.4)$$

Compared with the basic NMF [109] where $\mathbf{X} \approx \mathbf{WH}$, PNMF replaces the factorizing matrix $\mathbf{H}$ with $\mathbf{W}^T\mathbf{X}$, which brings a sparser factorized matrix desired for cluster analysis. The kernelized version of PNMF is discussed by Yang and Oja [188], where the term $\mathbf{X}^T\mathbf{X}$ is replaced by the similarity matrix $\mathbf{A}$.

- Semi-NMF and Convex-NMF, proposed by Ding et al. [46]. Both Semi-NMF and Convex-NMF allow the input data matrix $\mathbf{X}$ to have mixed signs, which extends the applicability of NMF methods. Semi-NMF directly connects to the K-means clustering with the factorization form $\mathbf{X}_\pm \approx \mathbf{W}_\pm\mathbf{H}$, where $\mathbf{X}_\pm$ is the data matrix, $\mathbf{W}_\pm$ ($\in \mathbb{R}^{m\times r}$) contains the cluster centroids, and $\mathbf{H}$ ($\in \mathbb{R}_+^{r\times n}$) contains the cluster membership indicators. Only $\mathbf{H}$ is constrained to be nonnegative. Convex-NMF further restricts the columns of $\mathbf{W}$ to be convex combinations of data points (columns) in $\mathbf{X}$ and considers the factorization form $\mathbf{X}_\pm \approx \mathbf{X}_\pm\mathbf{WH}$, where $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$. The term $\mathbf{X}_\pm\mathbf{W}$ can be interpreted as weighted cluster centroids. Note that both $\mathbf{W}$ and $\mathbf{H}$ are constrained to be nonnegative. The authors also discussed kernel Convex-NMF, where a special case is PNMF [194].

- Left Stochastic Matrix Decomposition (LSD), proposed by Arora et al. [3]. LSD is a probabilistic clustering method. Given a similarity matrix $\mathbf{A}$, it estimates a scaling factor c^* and a cluster probability matrix $\mathbf{W}^*$ to solve the following optimization problem:

$$\underset{c\in\mathbb{R}_+}{\text{minimize}} \left\{ \underset{\mathbf{W}\geq 0}{\text{minimize}} \, ||c\mathbf{A} - \mathbf{W}\mathbf{W}^T||_F^2, \text{ subject to } \sum_{k=1}^{r} W_{ik} = 1 \right\}, \quad (4.5)$$

  where $\mathbf{W} \in \mathbb{R}_+^{n\times r}$. Note that minimizing the scaling factor c^* is given in a closed form and does not depend on a particular solution $\mathbf{W}^*$, which means that only $\mathbf{W}$ needs to be updated. The authors also developed a rotation-based algorithm to compute the objective of Eq. 4.5.

In [133], Mørup et al. tackled the matrix factorization problem with Archetypal Analysis (AA). One advantage of AA is that its solution or factorization result is unique. More recent works concerning their work on AA can be found in [134, 167].

4.2 Quadratic nonnegative matrix factorization

Given a nonnegative input matrix $\mathbf{X} \in \mathbb{R}_+^{m \times n}$, conventional NMF and its variants find a number of matrices $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(Q)}$, some of which are constrained to be nonnegative, so that the distance between $\mathbf{X}$ and its approximation matrix $\widehat{\mathbf{X}} = \prod_{q=1}^{Q} \mathbf{W}^{(q)}$ can be minimized. Most existing NMF methods are *linear* in that each factorizing matrix $\mathbf{W}^{(q)}$ appears only once in the approximation. In [191], Yang and Oja introduced a higher-order class of NMF methods called Quadratic Nonnegative Matrix Factorization (QNMF), where some of the factorizing matrices appear twice in the approximation, or formally, $\mathbf{W}^{(s)} = \mathbf{W}^{(t)^T}$ for a series of non-overlapping pairs (s, t) with $1 \leq s < t \leq Q$.

There are many important real-world problems that employ quadratic factorization forms. One example is graph matching, when presented as a matrix factorization problem [45]: given two adjacency matrices $\mathbf{A}$ and $\mathbf{B}$, graph matching aims to find a permutation matrix $\mathbf{W}$ so that $\mathbf{A} = \mathbf{W}\mathbf{B}\mathbf{W}^T$. If such matrix $\mathbf{W}$ exists, then matrices $\mathbf{A}$ and $\mathbf{B}$ are considered to be isomorphic. This is actually an NMF problem when minimizing the distance between $\mathbf{A}$ and $\mathbf{W}\mathbf{B}\mathbf{W}^T$ with respect to $\mathbf{W}$, under certain constraints. Note that the approximation here is *quadratic* in $\mathbf{W}$ since the matrix $\mathbf{W}$ occurs twice. Another example is clustering. Assume an input data matrix $\mathbf{X}$ with n columns to be grouped into r disjoint clusters, the classical k-means objective function can be written as $\mathcal{J}_1 = \text{trace}(\mathbf{X}^T\mathbf{X}) - \text{trace}(\mathbf{U}^T\mathbf{X}^T\mathbf{X}\mathbf{U})$ [43], where $\mathbf{U} \in \mathbb{R}^{n \times m}$ is the binary cluster indicator matrix. It has been shown in [188] that minimizing the objective of Projective NMF (PNMF) [194] $\mathcal{J}_2 = ||\mathbf{X}^T - \mathbf{W}\mathbf{W}^T\mathbf{X}^T||_F^2$ achieves the same solution except for the binary constraint. In this example, the factorization is also *quadratic* in $\mathbf{W}$.

4.2.1 Factorization form

Following [191], we can write the general approximating factorization form of QNMF as:

$$\widehat{\mathbf{X}} = \mathbf{A}\mathbf{W}\mathbf{B}\mathbf{W}^T\mathbf{C}, \tag{4.6}$$

where we consider only one doubly occurring matrix $\mathbf{W}$ first. Note that matrix $\mathbf{A}$ or $\mathbf{B}$ or $\mathbf{C}$ can be the products of any number of linearly appearing matrices. Here we focus on optimizing the matrix $\mathbf{W}$, since the optimization of other linearly occurring matrices can be done by solving each matrix independently using standard NMF methods such as [110].

The factorization form of Eq. 4.6 is general such that it accommodates several QNMF objectives proposed earlier:

- when $\mathbf{A} = \mathbf{B} = \mathbf{I}$ and $\mathbf{C} = \mathbf{X}$, the factorization becomes PNMF, i.e., $\widehat{\mathbf{X}} = \mathbf{W}\mathbf{W}^T\mathbf{X}$, which is also called *Clustering-NMF* as a constrained case of Convex-NMF [46].

- when $\mathbf{X}$ is a square matrix and $\mathbf{A} = \mathbf{B} = \mathbf{C} = \mathbf{I}$, the factorization reduces to the Symmetric Nonnegative Matrix Factorization (SNMF) $\widehat{\mathbf{X}} = \mathbf{W}\mathbf{W}^T$ as a special case of 3-factor NMF [48].

- when $\mathbf{X}$ and $\mathbf{B}$ are the same-size square matrices and $\mathbf{A} = \mathbf{C} = \mathbf{I}$, the factorization has the form $\widehat{\mathbf{X}} = \mathbf{W}\mathbf{B}\mathbf{W}^T$. With the orthogonal constraint $\mathbf{W}\mathbf{W}^T = \mathbf{I}$, this corresponds to learning a permutation matrix for solving, for example, the graph matching problem [45]. With the constraint that each column of $\mathbf{W}$ sums to be 1, the learned $\mathbf{W}$ serves for estimating parameters of hidden Markov chains (see Section 5.3 of [191]).

In addition to one doubly occurring matrix $\mathbf{W}$, there can be cases where two or more doubly appearing matrices exist as well. For instance, when $\mathbf{A} = \mathbf{C}^T = \mathbf{U}$, the factorization 4.6 becomes $\widehat{\mathbf{X}} = \mathbf{U}\mathbf{W}\mathbf{B}\mathbf{W}^T\mathbf{U}^T$; when $\mathbf{A} = \mathbf{B} = \mathbf{I}$ and $\mathbf{C} = \mathbf{X}\mathbf{U}\mathbf{U}^T$, it gives $\widehat{\mathbf{X}} = \mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{U}\mathbf{U}^T$, and the solution of the latter QNMF problem can be used for solving the bi-clustering or co-clustering problem, i.e., to group the rows and columns of data matrix $\mathbf{X}$ simultaneously. In such cases we can utilize an alternative optimization approach over each doubly appearing matrix.

It is worth mentioning that the quadratic NMF problems can not be considered as special cases of linear NMF. In linear NMF, the factorizing matrices are all different and each of them can be optimized while keeping the others unchanged. However, in quadratic NMF, the optimization is more complex in that there are at least two matrices changing at the same time, leading to higher-order objectives. For example, the objective of linear NMF [110] with Euclidean distance $||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2$ is *quadratic* with respect to $\mathbf{W}$ and $\mathbf{H}$, whereas the objective of PNMF $||\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}||_F^2$ is *quartic* with respect to $\mathbf{W}$. Mathematically, minimizing a fourth-order objective as such is more difficult than minimizing a quadratic function.

4.2.2 Multiplicative update algorithms

Multiplicative update rules have been widely adopted for optimizing NMF objectives because they are easy to implement and to use. Multiplicative algorithms can automatically maintain the nonnegativity constraints and require no user-specified parameters during the iterative updating stage. Similar to linear NMF, there exists a multiplicative update algorithm for a wide variety of quadratic NMF that theoretically guarantees convergence or the monotonic decrease of the objective function in each iteration, if the QNMF objective can be written as a generalized polynomial form [189]. In the rest, we recapitulate the essence of multiplicative algorithms for QNMF. We write $\widetilde{\mathbf{X}} = \mathbf{A}\widetilde{\mathbf{W}}\mathbf{B}\widetilde{\mathbf{W}}^T\mathbf{C}$ for denoting the approximation that contains the variable $\widetilde{\mathbf{X}}$, and $\widehat{\mathbf{X}} = \mathbf{AWBW}^T\mathbf{C}$ for the current estimate.

Following the approach by Lee and Seung [110], the convergence proof of the QNMF objective is carried out by minimizing a certain auxiliary upper-bounding function. First we define the objective of QNMF as:

$$\mathcal{J}(\mathbf{W}) \stackrel{def}{=\!=} D(\mathbf{X}||\mathbf{AWBW}^T\mathbf{C}), \tag{4.7}$$

where $D()$ is a divergence measure given in the Appendix. $G(\mathbf{W}, \mathbf{U})$ is defined as an auxiliary function if it satisfies:

$$G(\mathbf{W}, \mathbf{U}) \geq \mathcal{J}(\mathbf{W}), \text{ and } G(\mathbf{W}, \mathbf{W}) = \mathcal{J}(\mathbf{W}). \tag{4.8}$$

Let us define

$$\mathbf{W}^{new} = \arg\min_{\widetilde{\mathbf{W}}} G(\widetilde{\mathbf{W}}, \mathbf{W}). \tag{4.9}$$

By construction, we have

$$\mathcal{J}(\mathbf{W}) = G(\mathbf{W}, \mathbf{W}) \geq G(\mathbf{W}^{new}, \mathbf{W}) \geq G(\mathbf{W}^{new}, \mathbf{W}^{new}) = \mathcal{J}(\mathbf{W}^{new}), \tag{4.10}$$

where the first inequality results from minimization and the second from the upper bound. By iteratively applying the update rule of Eq. 4.9, one can obtain a monotonically decreasing sequence of $\mathcal{J}$ to guarantee the convergence of the objective function. Further, by setting $\frac{\partial G}{\partial \widetilde{\mathbf{W}}} = 0$, one can have a closed-form iterative update rule that usually takes the form

$$W_{ik}^{new} = W_{ik} \left(\frac{\nabla_{ik}^-}{\nabla_{ik}^+} \right)^{\eta}, \tag{4.11}$$

where the terms ∇^+ and ∇^- denote respectively the sums of positive and unsigned negative parts of $\nabla = \partial \mathcal{J}(\widetilde{\mathbf{X}})/\partial \widetilde{\mathbf{X}}|_{\widetilde{\mathbf{X}}=\mathbf{W}}$ (i.e. $\nabla = \nabla^+ - \nabla^-$). The exponent η is determined by the specific NMF objective and the

Majorization-Minimization optimization procedure, and it guarantees a monotonic decrease of approximation errors (see more details in [189]).

For QNMF, the update rule can be unified by the following equation [189, 191]:

$$W_{ik}^{new} = W_{ik} \left[\frac{\left(A^T Q C^T W B^T + C Q^T A W B\right)_{ik}}{\left(A^T P C^T W B^T + C P^T A W B\right)_{ik}} \cdot \theta \right]^{\eta}, \tag{4.12}$$

where $\mathbf{P}, \mathbf{Q}, \theta$, and η are specified in Table 4.1. For example, the update rule for the QNMF problem $\mathbf{X} \approx \mathbf{WBW}^T$ based on the (squared) Euclidean distance (i.e. $\beta \to 1$) takes the form

$$W_{ik}^{new} = W_{ik} \left[\frac{\left(XWB^T + X^T WB\right)_{ik}}{\left(WBW^T WB^T + WB^T W^T WB\right)_{ik}} \right]^{1/4}. \tag{4.13}$$

Table 4.1. Notations in the multiplicative update rules of QNMF examples, where $\widehat{\mathbf{X}} = \mathbf{AWBW}^T\mathbf{C}$.

Divergence	$\mathbf{P}_{ij}$	$\mathbf{Q}_{ij}$	θ	η
α-Divergence	1	$X_{ij}^{\alpha}\widehat{X}_{ij}^{-\alpha}$	1	$1/(2\alpha)$ for $\alpha > 1$
				1/2 for $0 < \alpha < 1$
				$1/(2\alpha - 2)$ for $\alpha < 0$
β-Divergence	$\widehat{X}_{ij}^{\beta}$	$X_{ij}\widehat{X}_{ij}^{\beta-1}$	1	$1/(2 + 2\beta)$ for $\beta > 0$
				$1/(2 - 2\beta)$ for $\beta < 0$
γ-Divergence	$\widehat{X}_{ij}^{\gamma}$	$X_{ij}\widehat{X}_{ij}^{\gamma-1}$	$\frac{\sum_{ab}\widehat{X}_{ab}^{\gamma+1}}{\sum_{ab}\widehat{X}_{ab}^{\gamma}}$	$1/(2 + 2\gamma)$ for $\gamma > 0$
				$1/(2 - 2\gamma)$ for $\gamma < 0$
Rényi divergence	1	$X_{ij}^{r}\widehat{X}_{ij}^{-r}$	$\frac{\sum_{ab}\widehat{X}_{ab}}{\sum_{ab}X_{ab}^{r}\widehat{X}_{ab}^{1-r}}$	$1/(2\alpha)$ for $r > 1$
				1/2 for $0 < r < 1$

4.2.3 Adaptive multiplicative updates for QNMF

The original QNMF multiplicative update rules in Eq. 4.12 have a fixed form, which means the exponent η does not change during all iterations. Despite the simplicity, the constant exponent corresponds to overly conservative learning steps and thus often leads to mediocre convergence speed [62, 154].

In Publication V, we proposed an adaptive multiplicative update scheme for QNMF algorithms to overcome this drawback: we replace the constant

Algorithm 4 Multiplicative Updates with Adaptive Exponent for QNMF

Usage: $\mathbf{W} \leftarrow \text{FastQNMF}(\mathbf{X}, \eta, \mu)$.

Initialize $\mathbf{W}$; $\rho \leftarrow \eta$.

repeat

$$U_{ik} \leftarrow W_{ik} \left[\frac{\left(A^T Q C^T W B^T + C Q^T A W B\right)_{ik}}{\left(A^T P C^T W B^T + C P^T A W B\right)_{ik}} \cdot \theta \right]^{\rho}$$

if $D(\mathbf{X}||\mathbf{AUBU}^T\mathbf{C}) < D(\mathbf{X}||\mathbf{AWBW}^T\mathbf{C})$ **then**

$\mathbf{W} \leftarrow \mathbf{U}$

$\rho \leftarrow \rho + \mu$

else

$\rho \leftarrow \eta$

end if

until convergent conditions are satisfied

exponent in multiplicative update rules by a variable one, which accelerates the optimization while still maintaining the monotonic decrease of QNMF objective function. In particular, the proposed approach increases the exponent steadily if the new objective is smaller than the old one and otherwise shrinks back to the safe choice, η. We have empirically used $\eta = 0.1$ in all related experiments in this work. Algorithm 4 gives the pseudo-code for implementing the adaptive QNMF multiplicative update algorithm, and its monotonicity proof is straightforward by following the theorem proof in [191]:

Proposition 1. $D(\mathbf{X}||\mathbf{AWBW}^T\mathbf{C})$ *monotonically decreases after each of the iterations in Algorithm 4.*

Proof. If $D(\mathbf{X}||\mathbf{AUBU}^T\mathbf{C}) < D(\mathbf{X}||\mathbf{AWBW}^T\mathbf{C})$, then after $\mathbf{W}$ is replaced by $\mathbf{U}$ and the exponent ρ is replaced by $\rho + \mu$, $D(\mathbf{X}||\mathbf{AWBW}^T\mathbf{C})$ monotonically decreases; otherwise, $\mathbf{W}$ remains unchanged and the exponent ρ returns to the initial value η. Thus $D(\mathbf{X}||\mathbf{AUBU}^T\mathbf{C}) < D(\mathbf{X}||\mathbf{AWBW}^T\mathbf{C})$, which is guaranteed by Theorem 1 in [191]. □

A similar adaptive scheme was presented in our earlier work of Publication IV for accelerating the convergence speed of Projective NMF (PNMF) objective. PNMF is a special case of QNMF, and Publication V has generalized the adaptive multiplicative update algorithm for a wide variety of QNMF applications.

Here we show the performance of the adaptive multiplicative update algorithm using several real-world datasets. The statistics of the datasets are summarized in Table 4.2. These datasets were obtained from the UCI

Table 4.2. Datasets used in the PNMF experiments.

Datasets	Dimensions	#Samples
wine	13	178
mfeat	292	2000
orl	10304	400
feret	1024	2409
swimmer	1024	256
cisi	1460	5609
cran	1398	4612
med	1033	5831

repository[1], the University of Florida Sparse Matrix Collection[2], and the LSI text corpora[3], as well as other publicly available websites.

Figure 4.1 shows the objective function evolution curves using the original and the adaptive PNMF multiplicative update algorithms for the eight datasets. It is clear to see that the dashed lines are below the solid ones in respective plots, which indicates that the adaptive update algorithm is significantly faster than the original one.

Table 4.3 gives the mean and standard deviation of the convergence time of PNMF using the compared algorithms. The convergence time is calculated at the earliest iteration where the objective D is sufficiently close to the minimum D^*, i.e. $|D-D^*|/D^* < 0.001$. Each algorithm on each dataset has been repeated 100 times with different random seeds for initialization. These quantitative results confirm that the adaptive algorithm is significantly faster: it is 3 to 5 times faster than the original one.

The adaptive algorithm can be applied for other QNMF applications, for example, bi-clustering (also called co-clustering or two-way clustering). Bi-clustering aims to simultaneously group rows and columns of given input matrix, and can be formulated as: $\mathbf{X} \approx \mathbf{L}\mathbf{L}^T\mathbf{X}\mathbf{R}\mathbf{R}^T$ [191]. This is a two-sided QNMF problem, which can be solved by alternatively optimizing $\mathbf{X} \approx \mathbf{L}\mathbf{L}^T\mathbf{Y}^{(R)}$ with $\mathbf{Y}^{(R)} = \mathbf{X}\mathbf{R}\mathbf{R}^T$ fixed and $\mathbf{X} \approx \mathbf{Y}^{(L)}\mathbf{R}\mathbf{R}^T$ with $\mathbf{Y}^{(L)} = \mathbf{L}\mathbf{L}^T\mathbf{X}$ fixed. The bi-cluster indices of rows and columns are given by taking the maximum of each row in $\mathbf{L}$ and $\mathbf{R}$.

The above *Bi-clustering QNMF* (Bi-QNMF) was originally implemented by interleaving multiplicative updates between $\mathbf{L}$ and $\mathbf{R}$ using constant

[1] http://archive.ics.uci.edu/ml/

[2] http://www.cise.ufl.edu/research/sparse/matrices/index.html

[3] http://www.cs.utk.edu/~lsi/corpa.html

Table 4.3. The mean and standard deviation of the convergence time (seconds) of PNMF using the compared algorithms.

(a) PNMF based on Euclidean distance

dataset	original	adaptive
wine	0.22±0.11	0.06±0.03
mfeat	68.57±1.75	19.10±0.70
orl	117.26±1.74	29.89±1.48
feret	107.58±24.43	19.97±5.60

(b) PNMF based on I-divergence

dataset	original	adaptive
swimmer	613.04±20.63	193.47±5.43
cisi	863.89±69.23	193.23±18.70
cran	809.61±62.64	189.41±18.50
med	566.99±64.44	132.67±13.86

exponents. Figure 4.2 shows the comparison results between the previous implementation and the adaptive algorithm using variable exponents, on both synthetic and real-world data. The synthetic data is a 200×200 blockwise nonnegative matrix, where each block has dimensions $20, 30, 60$ or 90 and the matrix entries in a block are randomly drawn from the same Poisson distribution whose mean is chosen from $1, 2, 4$, or 7. The real-world data matrix contains a subset of the whole *webkb* textual dataset[4], with two classes of 1433 documents and 933 terms. The ij-th entry of the matrix is the number of the j-th term that appears in the i-th document.

We can see that the dashed curves are below the solid ones for both datasets, which indicates that the adaptive algorithm brings efficiency improvement. The advantage is further quantified in Table 4.4, where we ran each algorithm 10 times and recorded their mean and standard deviation of the convergence times.

It is worth mentioning that, in addition to QNMF algorithms, the proposed adaptive exponent technique is readily extended to other fixed-point algorithms that use multiplicative updates.

4.2.4 QNMF with additional constraints

Similar to NMF objectives, there are situations where one needs to append certain constrains to QNMF objectives for achieving the desired out-

[4] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

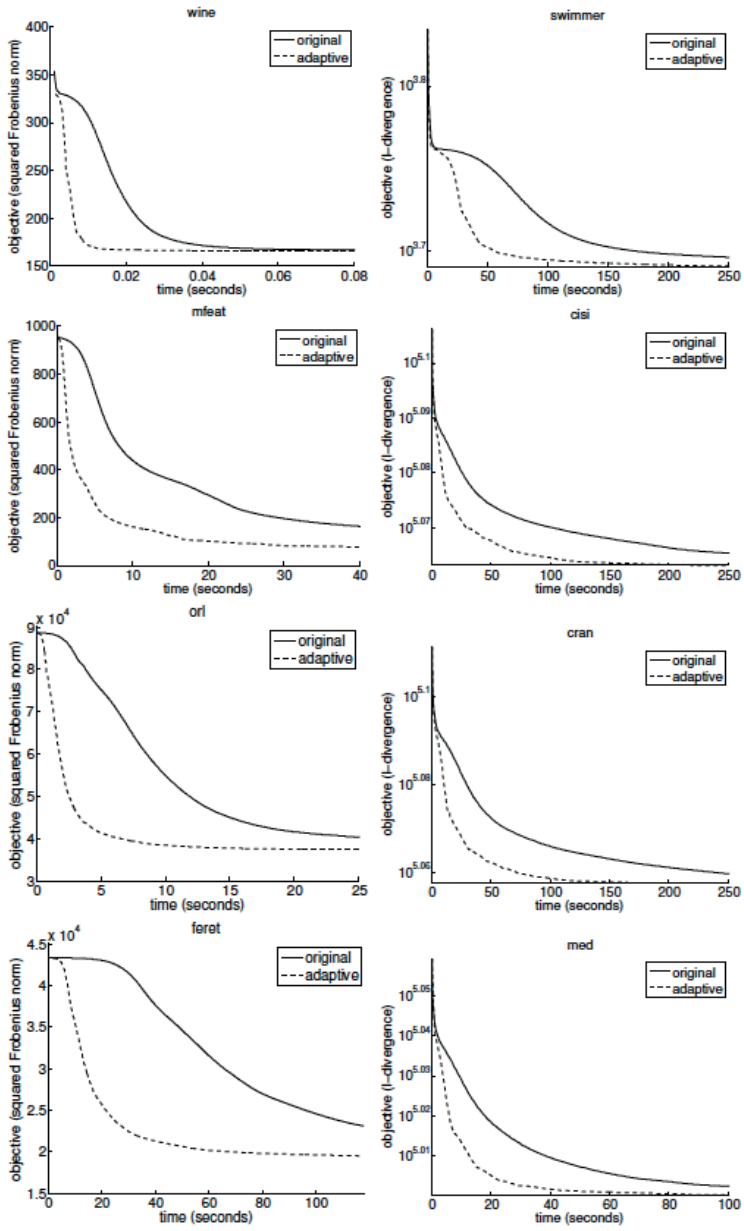


Figure 4.1. Evolutions of objectives using the original and adaptive PNMF multiplicative update algorithms based on (left) squared Euclidean distance and (right) I-divergence.

Table 4.4. The mean and standard deviation of the converged time (seconds) of Bi-QNMF using the compared algorithms.

data	original	adaptive
`synthetic`	17.96 ± 0.26	5.63 ± 0.10
`webkb`	139.35 ± 76.81	25.93 ± 13.61

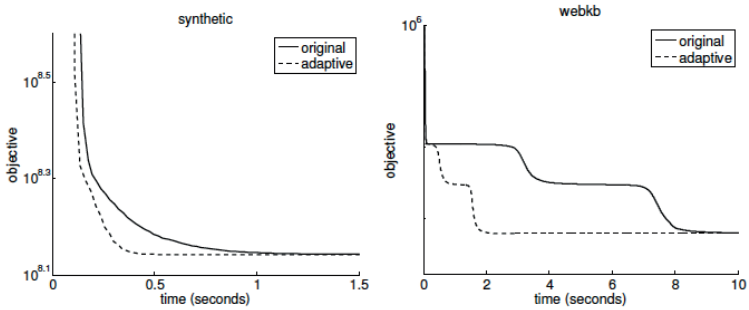


Figure 4.2. Evolutions of objectives using the original and adaptive Bi-QNMF multiplicative update algorithms based on (left) synthetic data and (right) webkb data.

puts. For solving the constrained QNMF problems, a relaxation technique has been adopted by the authors in [191], summarized as below:

Step 1 The (soft) constraints are attached to the QNMF objective as regularization terms;

Step 2 A multiplicative update rule is constructed for the augmented objective, where the Lagrangian multipliers are solved by using the K.K.T. conditions;

Step 3 Inserting back the multipliers one thus obtains new update rules with the (soft) constraints incorporated.

The above algorithm is called *iterative Lagrangian solution* of the constrained QNMF problem. It minimizes the QNMF approximation error while forcing the factorizing matrices to approach the manifold specified by the constraints. A similar idea was also employed by Ding et al. [48]. Below we give solutions or update rules of QNMF with two widely-used constraints.

Stochastic constraints for nonnegative matrices are often utilized to represent probabilities, which ensures the summation of all or partial matrix elements to be one. The left-stochastic constraint ensures a column-wise unitary sum (i.e. $\sum_i W_{ik} = 1$); the right-stochastic constraint ensures a row-wise unitary sum; the matrix-wise stochastic constraint ensures a matrix-wise unitary sum. A general principle that incorporates the stochastic constraints to an existing convergent QNMF algorithm is given in [191].

Orthogonal constraints for nonnegative matrices are frequently utilized for approximating cluster indicator matrices, because a nonnegative orthogonal matrix has only one non-zero entry in each row. Usually a strict

orthogonality is not required since the discrete optimization problem is often NP-hard, and relaxation is needed so that there is only one large nonzero entry in each row of the matrix $\mathbf{W}$, leaving the other entries close to zero. A general principle that incorporates the orthogonal constraint to a theoretically convergent QNMF algorithm is given in [191].

4.2.5 NMF using graph random walk

In cluster analysis, the cluster assignment can be learned from pairwise similarities between data points. Let $\mathbf{S} \in \mathbb{R}_+^{n \times n}$ denote a pairwise similarity matrix encoding n data samples. Since clustered data tend to have higher similarities within clusters but lower similarities between clusters, the matrix $\mathbf{S}$ should have a nearly diagonal appearance if its rows and columns are sorted by clusters. This structure has motivated approximative low-rank factorization of $\mathbf{S}$ by the binary cluster indicator matrix $\mathbf{U} \in \{0,1\}^{n \times r}$ for r clusters: $\mathbf{S} \approx \mathbf{U}\mathbf{U}^T$, where $U_{ik} = 1$ if the i-th sample belongs to the k-th cluster and 0 otherwise, as shown by the example in Figure 4.3.

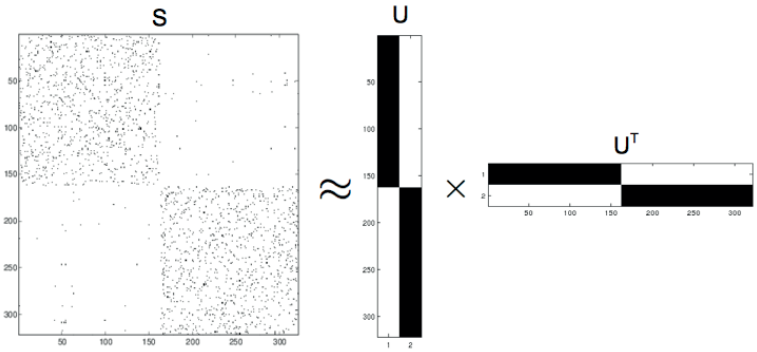


Figure 4.3. An illustrative example showing the approximation $\mathbf{S} \approx \mathbf{U}\mathbf{U}^T$, where $\mathbf{U} \in \{0,1\}^{321 \times 2}$ is the binary cluster indicator matrix, and $\mathbf{S} \in \mathbb{R}_+^{321 \times 321}$ is the symmetrized 5-NN similarity matrix constructed by taking the subset of digits "2" (162 samples) and "3" (159 samples) from the *SEMEION* handwritten digit dataset. The matrix entries are visualized as image pixels, with darker pixels representing higher similarities.

However, as stated in Chapter 4, directly optimizing over $\mathbf{U}$ often leads to an NP-hard problem due to the discrete solution space, and the continuous relaxation is thus needed to ease the problem. A popular relaxation technique is to combine the nonnegativity and orthogonality constraints, that is, replacing $\mathbf{U}$ with $\mathbf{W}$ where $W_{ik} \geq 0$ and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. After that, each row of $\mathbf{W}$ has only one non-zero entry since two nonnegative orthogonal vectors do not overlap. Therefore the factorization takes the form

$\mathbf{S} \approx \mathbf{W}\mathbf{W}^T$, with the constraints of $W_{ik} \geq 0$ and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. This problem is called Symmetric NMF (SNMF) [48], which is also a special case of the QNMF problem with auxiliary constraints.

A commonly used approximation error or divergence is the Least Square Error (LSE) or squared Euclidean (EU) distance (Frobenius norm) [110, 78]. By using LSE, the above SNMF problem can be solved by minimizing the following objective function:

$$\mathcal{J} = ||\mathbf{S} - \mathbf{W}\mathbf{W}^T||_F^2 = \sum_{ij} \left[S_{ij} - (WW^T)_{ij}\right]^2 \tag{4.14}$$

$$\text{subject to } W_{ik} \geq 0 \text{ and } \mathbf{W}^T\mathbf{W} = \mathbf{I}. \tag{4.15}$$

The EU distance measure is the most popular approximation criterion. However, NMF methods using LSE often give mediocre clustering results.

To see the reason, one may consider the example as illustrated in Figure 4.4. To minimize $||\mathbf{S} - \widehat{\mathbf{S}}||_F^2$ for a given similarity matrix $\mathbf{S}$, the approximating matrix $\widehat{\mathbf{S}}$ should be diagonal blockwise for clustering, as shown in Figure 4.4 (b), and the ideal input $\mathbf{S}$ should be similar to Figure 4.4 (c). However, the similarity matrices of real-world data often look like Figure 4.4 (a), where the non-zero entries are much sparser than the ideal case in Figure 4.4 (c). This is because the raw data features are usually weak such that a simple metric like EU distance is only valid in a small neighborhood, but tends to be unreliable for non-neighboring data points with long distances. Therefore the similarities between those non-neighboring samples are usually set to be zero, and the resulting similarity matrix is often sparse as in Figure 4.4 (a).

Since the underlying distribution of LSE is Gaussian, which is good at handling dense matrices, it is a mismatch to approximate a sparse similarity matrix by a dense diagonal blockwise matrix using LSE. Because the squared Euclidean distance is a symmetric metric, the learning objective can be dominated by the approximation to the majority of zero entries, which may hinder from finding correct cluster assignments. Yet little research effort has been made to overcome the above mismatch.

To reduce the sparsity gap between input and output matrices, Yang et al. [187] proposed to approximate a smoothed version of $\mathbf{S}$ using graph random walk, which implements multi-step similarities that considers farther relationships between data samples. Assume the normalized similarity matrix $\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$. By using graph random walk, the similarities between data nodes are given by $(\alpha\mathbf{Q})^j$, where $\alpha \in (0, 1)$ is the decay parameter that

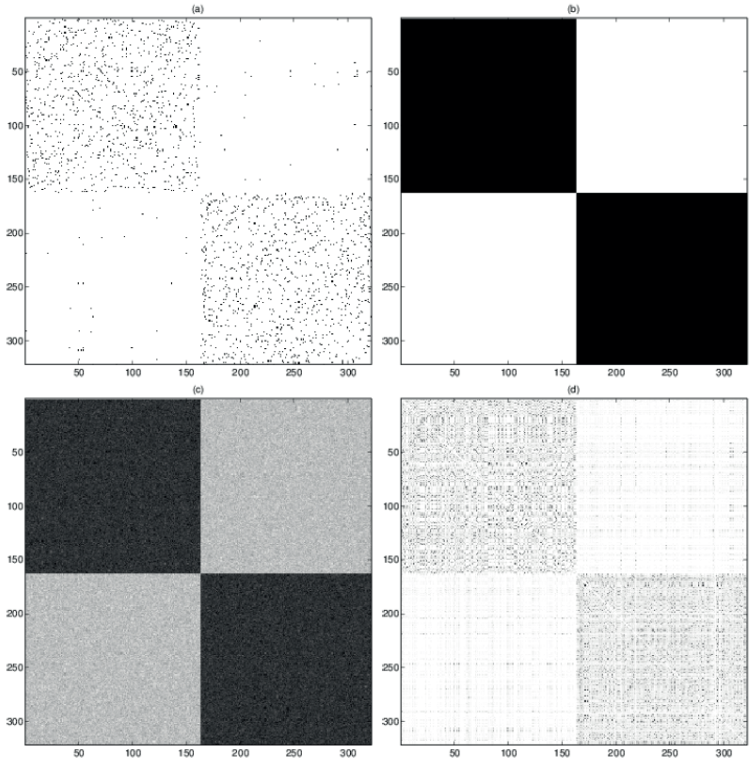


Figure 4.4. An example from [187] showing the LSE-based NMF clustering: (a) the symmetrized 5-NN similarity matrix constructed by taking the subset of digits "2" (162 samples) and "3" (159 samples) from the *SEMEION* handwritten digit dataset, (b) the correct clusters to be found, (c) the ideally assumed data that suits the least square error, (d) the smoothed input by using graph random walk. The matrix entries are visualized as image pixels, with darker pixels representing higher similarities. In [187], Yang et al. proposed to find correct clusters using (d) $\approx$ (b) instead of (a) $\approx$ (b) by NMF with LSE, because (d) is "closer" to (c) than (a).

controls the random walk extent. Summing over all possible numbers of steps gives $\sum_{j=1}^{\infty}(\alpha\mathbf{Q})^j = (\mathbf{I}-\alpha\mathbf{Q})^{-1}$. The authors thus proposed to replace $\mathbf{S}$ with the following smoothed similarity matrix:

$$\mathbf{A} = c^{-1}\left(\mathbf{I} - \alpha\mathbf{Q}\right)^{-1}, \tag{4.16}$$

where $c = \sum_{ij}[(I - \alpha Q)^{-1}]_{ij}$ is a normalizing factor. The parameter α constrols the smoothness: a larger α tends to produce a smoother $\mathbf{A}$, whereas a smaller one makes $\mathbf{A}$ shrink on its diagonal. Figure 4.4 (d) shows a smoothed approximated matrix $\mathbf{A}$, where one can see that the sparsity gap to the approximating matrix $\widehat{\mathbf{S}}$ has been reduced. Therefore, the opti-

mization problem of Eq. 4.14 (with constraints) becomes:

$$\underset{\mathbf{W}\geq 0}{\text{minimize}}\, \mathcal{J}(\mathbf{W}) = -trace(\mathbf{W}^T\mathbf{A}\mathbf{W}) + \lambda \sum_i \left(\sum_k W_{ik}^2 \right)^2 \tag{4.17}$$

$$\text{subject to } \mathbf{W}^T\mathbf{W} = \mathbf{I}, \tag{4.18}$$

where λ is the tradeoff parameter. The extra penalty term has two functions [187]: (1) it emphasizes off-diagonal correlations in the trace, and (2) it tends to equalize the norms of $\mathbf{W}$ rows.

The above constrained optimization problem can be solved by using the following multiplicative update rule [187]:

$$W_{ik}^{new} = W_{ik} \left[\frac{\left(AW + 2\lambda WW^TVW\right)_{ik}}{\left(2\lambda VW + WW^TAW\right)_{ik}} \right]^{1/4}, \tag{4.19}$$

where $\mathbf{V}$ is a diagonal matrix with $V_{ii} = \sum_l W_{il}^2$. Note that the update rule of Eq 4.19 only needs the product of $(\mathbf{I} - \alpha\mathbf{Q})^{-1}$ with a low-rank matrix instead of $\mathbf{A}$, which avoids expensive computation and storage of a large smoothed similarity matrix.

4.3 Clustering by low-rank doubly stochastic matrix decomposition

Clustering methods based on NMF or QNMF are restricted to the scope of matrix factorization. Yang and Oja [190] proposed a nonnegative low-rank approximation method to improve the clustering. The proposed method is based on Data-Cluster-Data random walk and thus named DCD. DCD goes beyond matrix factorization because the decomposition of the approximating matrix includes operations other than matrix product.

4.3.1 Learning objective

Given n data samples to be grouped into r disjoint clusters. Let i, j, and v be indices for data points, and k and l for clusters. Assume $\mathbf{A} \in \mathbb{R}_+^{n\times n}$ as the similarity matrix between samples, and $P(k|i)$ ($i = 1, \ldots, n$ and $k = 1, \ldots, r$) as the probability of assigning the ith sample to the kth cluster.

DCD seeks an approximation to $\mathbf{A}$ by another matrix $\widehat{\mathbf{A}}$ whose elements correspond to the probabilities of two-step random walks between data points through clusters 1 to r. By using the Bayes formula and the uniform prior $P(i) = 1/n$, the random walk probabilities are given by

$$\widehat{A}_{ij} = P(i|j) = \sum_k P(i|k)P(k|j) = \sum_k \frac{P(k|i)P(k|j)}{\sum_v P(k|v)}. \tag{4.20}$$

Let us write $W_{ik} = P(k|i)$ for convenience, and thus

$$\widehat{A}_{ij} = \sum_k \frac{W_{ik} W_{jk}}{\sum_v W_{vk}}. \tag{4.21}$$

By using the Kullback-Leibler (KL-) divergence, the DCD approximation is formulated as the following optimization problem:

$$\underset{\mathbf{W} \geq 0}{\text{minimize}}\, D_{KL}(\mathbf{A}||\widehat{\mathbf{A}}) = \sum_{ij} \left(A_{ij} \log \frac{A_{ij}}{\widehat{A}_{ij}} - A_{ij} + \widehat{A}_{ij} \right), \tag{4.22}$$

$$\text{subject to} \sum_k W_{ik} = 1, i = 1, \ldots, n. \tag{4.23}$$

4.3.2 Optimization

The optimization is solved by a Majorization-Minimization algorithm [86, 189, 191, 202] that iteratively applies a multiplicative update rule:

$$W_{ik}^{new} = W_{ik} \left(\frac{\nabla_{ik}^{-} a_i + 1}{\nabla_{ik}^{+} a_i + b_i} \right), \tag{4.24}$$

where the terms ∇^+ and ∇^- denote respectively the sums of positive and unsigned negative parts of the gradient, $a_i = \sum_l \frac{W_{il}}{\nabla_{il}^{+}}$, and $b_i = \sum_l W_{il} \frac{\nabla_{il}^{-}}{\nabla_{il}^{+}}$.

Given a good initial decomposing matrix $\mathbf{W}$, DCD can achieve better cluster purity than many other existing clustering approaches, especially for large-scale datasets where the data samples lie in a curved manifold such that only similarities in a small neighborhood are reliable. The advantages of DCD owe to its objective in three aspects: 1) the approximation error measure by Kullback-Leibler divergence takes into account sparse similarities; 2) the decomposition form ensures relatively balanced clusters and equal contribution of each data sample; 3) the probabilities from samples to clusters form the only decomposing matrix to be learned, and directly give the answer for probabilistic clustering.

5. Improving cluster analysis using co-initialization

Many clustering methods, including NMF-based approaches, adopt objective functions that are not convex. These objectives are usually solved by employing iterative algorithms that start from an initial value (or matrix). A proper initialization is thus critical for finding clusters with high qualities. In this chapter, we introduce a novel initialization strategy to improve clustering performance through combining a set of diverse clustering methods. We also present an initialization hierarchy, from simple to comprehensive, and empirically demonstrate that a higher level of initialization often achieves better clustering results, especially for methods that require a careful initialization such as the DCD approximation.

5.1 Motivation

For many clustering methods, their objective functions are non-convex and their optimization generally involves iterative algorithms that start from an initial guess. A proper initialization plays a key role in achieving good clustering results. Random initialization has widely been used by researchers due to its simplicity. However, random guesses often yield poor results and the iterative clustering algorithm has to be run many times with different starting points in order to get better solutions.

Many advanced initialization techniques have been proposed to improve the efficiency, for example, specific choices of the initial cluster centers of the classical k-means method (e.g. [21, 118, 97, 52]), or singular value decomposition for clustering based on nonnegative matrix factorization [201, 101]. However, there still lacks an initialization principle that is commonly applicable for a wide range of iterative clustering methods. Especially, there is little research on whether one clustering method could benefit from initializations by the results of other clustering methods.

5.2 Clustering by co-initialization

In Publication VI, we proposed a *co-initialization* strategy, where a set of base clustering methods provide initializations for each other to improve clustering performance. The proposed approach is based on two observations as follows:

1. Many clustering methods that use iterative optimization algorithms are sensitive to initializations, and random starting guesses often lead to poor local optima.

2. On the other hand, the iterative algorithms often converge to a much better result given a starting point that is sufficiently close to the optimal result or the ground truth.

These two observations inspired us to systematically study the behavior of an ensemble of clustering methods through *co-initializations*, i.e., providing starting guesses for each other. We presented a hierarchy of initializations towards this direction, where a higher level represents a more extensive strategy.

In the following, we call the clustering method used for initialization the *base method*, in contrast to the *main method* used for the actual consequent cluster analysis. The proposed initialization hierarchy is summarized into five levels as below:

Level 0 Random initialization: using random starting points. Typically the starting point is drawn from a uniform distribution. To find a better local optimum, one may repeat the optimization algorithm several times with different starting assignments (e.g. with different random seeds). Although the random initialization is easy to implement, such a heuristic approach often leads to clustering results which are far from a satisfactory partition.

Level 1 Simple initialization: initializing by a fast and computationally simple method such as k-means or NCUT. We call this strategy *simple initialization* because here the base method is simpler than the main clustering method. This strategy has been widely used in NMF-based clustering methods (e.g. [48, 45, 188]).

Level 2 Family initialization: using base methods from a same param-

eterized family as the main method for initialization. That is, both the base and the main methods use the same form of objective and metric but only differ by a few parameters. For example, in the above DCD method [190], varying α in the Dirichlet prior can provide different base methods [190]; the main method ($\alpha = 1$) and the base methods ($\alpha \neq 1$) belong to the same parametric family.

Level 3 Heterogeneous initialization: using any base methods to provide initialization for the main method. We call this strategy *heterogeneous initialization* because we remove the constraint of the same parameterized family and generalize the above family initialization such that any clustering methods can be used as base methods. Similar to the strategies for combining classifiers, it is reasonable to have base methods as diverse as possible for better exploration. Algorithm 5 gives the pseudocodes for heterogeneous initialization.

Level 4 Heterogeneous co-initialization: running in multiple iterations, where in each iteration all participating methods provide initialization for each other. Here we make no difference from base and main methods. The participating methods can provide initializations to each other, and such cooperative learning can run for more than one iteration. That is, when one algorithm finds a better local optimum, the resulting cluster assignment can again serve as the starting guess for the other clustering methods. The loop will converge when none of the involved methods can find a better local optimum. Note that the convergence is guaranteed if the involved objective functions are all bounded. A special case of this strategy was used for combining NMF and Probabilistic Latent Semantic Indexing [47]. Here we generalize this idea to any participating clustering methods. Algorithm 6 gives the pseudocodes for heterogeneous co-initialization.

Ensemble clustering is another way to combine a set of clustering methods, where several base clustering results from different clustering methods are combined into a single categorical output through a combining function called *consensus function* (e.g. [63, 59, 87]). However, the performance of ensemble clustering methods will not bring extraordinary improvement over the base clustering methods if the base methods fall into poor local optima during the optimization (see the numerical examples in the following Section 5.3).

Algorithm 5 Cluster analysis using heterogeneous initialization. We denote $\mathbf{W} \leftarrow \mathcal{M}(\mathcal{D}, \mathbf{U})$ a run of clustering method $\mathcal{M}$ on data $\mathcal{D}$, with starting guess matrix $\mathbf{U}$ and output cluster indicator matrix $\mathbf{W}$.

Input: data $\mathcal{D}$, base clustering methods $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_T$, and main clustering method $\mathcal{M}$

Initialize $\{\mathbf{U}_t\}_{t=1}^T$ by e.g. random or simple initialization

for $t = 1$ to T **do**

  $\mathbf{V} \leftarrow \mathcal{B}_t(\mathcal{D}, \mathbf{U}_t)$

  $\mathbf{W}_t \leftarrow \mathcal{M}(\mathcal{D}, \mathbf{V})$

end for

Output: $\mathbf{W} \leftarrow \arg\min_{\mathbf{W}_t} \{\mathcal{J}_M(\mathbf{W}_t)\}_{t=1}^T$.

Algorithm 6 Cluster analysis using heterogeneous co-initialization.

Input: data $\mathcal{D}$ and clustering methods $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_T$

$\mathcal{J}_t \leftarrow \infty, t = 1, \ldots, T$.

Initialize $\{\mathbf{W}_t\}_{t=1}^T$ by e.g. random or simple initialization

repeat

  bContinue←False

  for $i = 1$ to T **do**

    for $j = 1$ to T **do**

      if $i \neq j$ **then**

        $\mathbf{U}_i \leftarrow \mathcal{M}_i(\mathcal{D}, \mathbf{W}_j)$

      end if

    end for

    $\mathcal{J} \leftarrow \min_{\mathbf{U}_j} \{\mathcal{J}_{M_j}(\mathcal{D}, \mathbf{U}_j)\}_{j=1}^T$

    $\mathbf{V} \leftarrow \arg\min_{\mathbf{U}_j} \{\mathcal{J}_{M_j}(\mathcal{D}, \mathbf{U}_j)\}_{j=1}^T$

    if $\mathcal{J} < \mathcal{J}_i$ **then**

      $\mathcal{J}_i \leftarrow \mathcal{J}$

      $\mathbf{W}_i \leftarrow \mathbf{V}$

      bContinue←True

    end if

  end for

until bContinue=False or maximum iteration is reached

Output: $\{\mathbf{W}_t\}_{t=1}^T$.

Table 5.1. Statistics of the datasets.

DATASET	# SAMPLES	# CLASSES
ORL	400	40
COIL20	1440	20
CITESEER	3312	6
USPS	9298	10
PROTEIN	17766	3
20NEWS	19938	20
LET-REC	20000	26
MNIST	70000	10

5.3 Empirical results

We have performed clustering experiments on real-world datasets covering various domains such as facial images, textual documents, handwritten digit / letter images, and protein etc. For each dataset, we constructed its similarity matrix using KNN graph ($K = 10$), which is then binarized and symmetrized as the nonnegative input. There are 19 different datasets used in our Publication VI. Here we show a subset of them for the illustration purpose (see Table 5.1 for their statistics). The clustering performance is evaluated by two-widely used criteria (described in Section 3.3), i.e. Purity and Normalized Mutual Information (NMI).

We have tested various clustering methods with different initializations in the hierarchy described previously. We focus on the following four levels: *random initialization*, *simple initialization*, *heterogeneous initialization*, and *heterogeneous co-initialization* in these experiments, while treating *family initialization* as a special case of *heterogeneous initialization*. For *heterogeneous co-initialization*, the number of co-initialization iterations was set to 5, as in practice we found that there is no significant improvement after five rounds. These levels of initializations have been applied to six clustering methods, i.e. PNMF [194, 188], NSC [45], ONMF [48], PLSI [82], LSD [3], and DCD [190]. For comparison, we also include the results of two other methods based on graph cut, i.e. Normalized Cut (NCUT) [159] and 1-Spectral Ratio Cheeger Cut (1-SPEC) [79].

The extensive empirical results are given in Table 5.2, shown in cells with quadruples. We can see from the results that more comprehensive initialization strategies often lead to better clusterings, where the four numbers in most cells monotonically increase from left to right. Some

clustering methods such as Normalized Cut [159] are not sensitive to initializations but tend to return less accurate clustering, whereas some methods can find more accurate clusters but require careful initialization (see discussions in e.g. [190, 187]). DCD is a typical clustering method of the latter kind, since the geometry of the KL-divergence cost function is more complex than the commonly-used cost functions based on the Euclidean distance.

We have also compared our co-initialization method with three ensemble clustering methods, i.e. the BEST algorithm [63], the co-association algorithm (CO) [59], and the link-based algorithm (CTS) [87]. For fair comparison, the set of base methods (i.e. same objective and same optimization algorithm) is the same for all compared approaches: the 11 bases are from NCUT, 1-SPEC, PNMF, NSC, ONMF, LSD, PLSI, DCD1, DCD1.2, DCD2, and DCD5 respectively. Here we chose the result by DCD for the comparison with the ensemble methods, as we find that this method benefits the most from co-initializations. The results are given in Table 5.3. We can see that DCD wins most clustering tasks and the superiority of DCD using co-initializations is especially distinct for large datasets that lie a curved manifold.

Table 5.2. Clustering performance of various clustering methods with different initializations. Performances are measured by (top) Purity and (bottom) NMI. Rows are ordered by dataset sizes. In cells with quadruples, the four numbers from left to right are results using *random*, *simple*, and *heterogeneous initialization* and *heterogeneous co-initialization*.

DATASET	KM	NCUT	1-SPEC	PNMF	NSC	ONMF	LSD	PLSI	DCD
ORL	0.70	0.81	0.81	0.81 0.81 0.81 0.81	0.81 0.81 0.81 0.81	0.53 0.78 0.80 0.80	0.82 0.82 0.82 0.82	0.65 0.81 0.83 0.83	0.67 0.81 0.83 0.83
COIL20	0.63	0.71	0.67	0.67 0.71 0.71 0.71	0.73 0.71 0.72 0.72	0.63 0.71 0.72 0.72	0.71 0.68 0.68 0.68	0.58 0.75 0.69 0.70	0.62 0.75 0.69 0.70
CITESEER	0.61	0.30	0.31	0.28 0.29 0.29 0.28	0.26 0.28 0.25 0.25	0.31 0.32 0.28 0.28	0.38 0.43 0.45 0.47	0.35 0.44 0.44 0.48	0.37 0.44 0.44 0.48
USPS	0.74	0.74	0.74	0.67 0.80 0.75 0.68	0.72 0.74 0.74 0.74	0.62 0.80 0.75 0.68	0.80 0.79 0.84 0.85	0.48 0.73 0.80 0.85	0.51 0.75 0.80 0.85
PROTEIN	0.46	0.46	0.46	0.46 0.46 0.46 0.46	0.46 0.46 0.46 0.46	0.46 0.46 0.46 0.46	0.46 0.46 0.48 0.50	0.46 0.51 0.51 0.50	0.47 0.46 0.51 0.50
20NEWS	0.07	0.43	0.36	0.39 0.39 0.39 0.38	0.39 0.43 0.40 0.40	0.27 0.38 0.38 0.38	0.41 0.48 0.48 0.49	0.22 0.44 0.49 0.49	0.23 0.45 0.49 0.50
LET-REC	0.29	0.21	0.15	0.36 0.37 0.34 0.35	0.17 0.21 0.21 0.21	0.29 0.35 0.35 0.34	0.31 0.31 0.37 0.37	0.16 0.26 0.32 0.38	0.17 0.25 0.32 0.38
MNIST	0.60	0.77	0.88	0.57 0.87 0.73 0.57	0.74 0.79 0.79 0.79	0.57 0.75 0.65 0.57	0.93 0.75 0.97 0.97	0.46 0.79 0.97 0.98	0.55 0.81 0.97 0.98

DATASET	KM	NCUT	1-SPEC	PNMF	NSC	ONMF	LSD	PLSI	DCD
ORL	0.85	0.90	0.92	0.89 0.90 0.89 0.89	0.89 0.90 0.90 0.90	0.76 0.88 0.89 0.89	0.90 0.90 0.90 0.90	0.84 0.90 0.91 0.91	0.83 0.90 0.91 0.91
COIL20	0.77	0.79	0.77	0.75 0.79 0.79 0.79	0.81 0.79 0.80 0.80	0.74 0.79 0.79 0.79	0.79 0.78 0.77 0.77	0.71 0.80 0.80 0.80	0.74 0.80 0.80 0.80
CITESEER	0.34	0.10	0.12	0.07 0.07 0.07 0.07	0.07 0.08 0.07 0.07	0.10 0.12 0.07 0.07	0.13 0.18 0.20 0.20	0.10 0.17 0.19 0.21	0.11 0.17 0.18 0.21
USPS	0.62	0.77	0.80	0.66 0.75 0.71 0.66	0.71 0.78 0.78 0.78	0.62 0.75 0.71 0.66	0.75 0.77 0.81 0.82	0.40 0.75 0.77 0.81	0.46 0.76 0.77 0.81
PROTEIN	0.00	0.01	0.01	0.02 0.01 0.01 0.02	0.01 0.01 0.01 0.01	0.00 0.01 0.01 0.00	0.01 0.00 0.02 0.04	0.01 0.04 0.04 0.04	0.02 0.01 0.04 0.04
20NEWS	0.05	0.54	0.52	0.36 0.36 0.36 0.34	0.48 0.54 0.52 0.52	0.24 0.34 0.34 0.34	0.36 0.43 0.44 0.44	0.14 0.44 0.44 0.45	0.14 0.45 0.44 0.45
LET-REC	0.35	0.38	0.26	0.43 0.43 0.42 0.43	0.21 0.38 0.37 0.37	0.35 0.43 0.43 0.42	0.39 0.41 0.45 0.45	0.17 0.37 0.42 0.46	0.18 0.36 0.42 0.46
MNIST	0.51	0.81	0.89	0.59 0.82 0.72 0.59	0.73 0.84 0.84 0.84	0.58 0.75 0.64 0.59	0.87 0.76 0.93 0.93	0.34 0.81 0.92 0.94	0.48 0.80 0.92 0.93

Table 5.3. Clustering performance comparison of DCD using *heterogeneous co-initialization* with three ensemble clustering methods. Rows are ordered by dataset sizes. Boldface numbers indicate the best. The 11 bases are from NCUT, 1-SPEC, PNMF, NSC, ONMF, LSD, PLSI, DCD1, DCD1.2, DCD2, and DCD5 respectively.

	Purity				NMI			
DATASET	BEST	CO	CTS	DCD	BEST	CO	CTS	DCD
ORL	0.81	0.81	0.80	**0.83**	0.90	0.90	0.90	**0.91**
COIL20	**0.73**	0.69	0.72	0.70	**0.80**	0.77	0.79	**0.80**
CITESEER	0.43	0.34	0.35	**0.48**	0.18	0.12	0.15	**0.21**
USPS	0.75	0.65	0.73	**0.85**	0.76	0.69	0.78	**0.81**
PROTEIN	0.46	0.46	0.46	**0.50**	0.01	0.01	0.01	**0.04**
20NEWS	0.45	0.28	0.40	**0.50**	0.45	0.38	**0.47**	0.45
LET-REC	0.26	0.23	0.24	**0.38**	0.37	0.35	0.39	**0.46**
MNIST	0.96	0.57	0.76	**0.98**	0.92	0.68	0.84	**0.93**

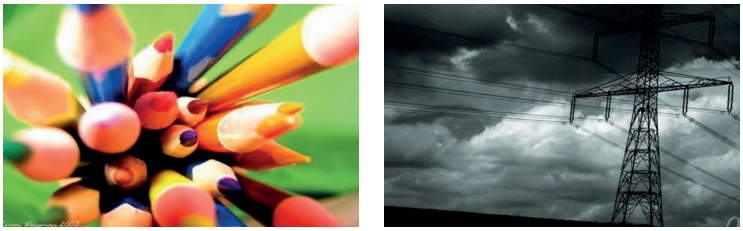
6. Cluster analysis on emotional images

This chapter presents an experimental study on emotional image clustering. Emotional semantic image analysis as a new research area has attracted an increasing attention in recent years. Most researchers attack this problem within a supervised learning context. Here we study image emotions from an unsupervised learning context and perform cluster analysis on a widely-used emotional image dataset.

6.1 Introduction

Content-Based Image Retrieval (CBIR) is the study that helps organize and index digital images by their visual content or similarity (see detail surveys in [122, 39]). A major challenge of the current approaches is handling the *semantic gap* between low-level visual features that the computer is relying upon and high-level semantic concepts that humans naturally associate with images [162]. To narrow down the semantic gap, relevance feedback techniques, including both explicit feedback and implicit feedback (also called enriched feedback [197]), have been utilized for decades to improve the CBIR performance [6, 96, 76].

In Publication VII, we presented a gaze-and-speech-enhanced CBIR system, where we analyzed the use of implicit relevance feedback from the user's gaze tracking patterns for boosting up the CBIR performance. A client-side information collector, implemented as a Firefox browser extension, can unobtrusively record a user's feedback forms on a displayed image, including eye movements, pointer tracks and clicks, keyboard strokes, and speech input, which are then transmitted asynchronously to the remote CBIR server for assisting in retrieving relevant images (operational details can be found in [198]). The information collector was integrated with an existing CBIR server named PicSOM [105], and the effectiveness



(a) Amusement (b) Fear

Figure 6.1. Example images from a photo sharing site (ArtPhoto [127]) with the ground truth labels of Amusement and Fear. Though emotions are highly subjective human factors, still they have stability and generality across different people and cultures [139]. Intuitively, an “Amusement" picture usually makes people feel pleasant or induces high valence, whereas a “Fear" picture may induce low valence but high arousal to the viewer.

of the system has been verified by real users in image-tagging tasks.

Images contain emotional information that can trigger people’s affective feelings (see Figure 6.1 as an example). Recently, studies related to affective image classification and retrieval have attracted an increasing research effort. Unlike most modern CBIR systems that were designed for recognizing objects and scenes such as plants, animals, outdoor places etc., an Emotional Semantic Image Retrieval (ESIR) system (see [179] for a recent survey) aims at incorporating the user’s affective states or emotions to bridge the so called “affective gap" [74], by enabling queries like “beautiful flowers", “cute dogs", “exciting games", etc. The major challenges in this area are (1) modeling image affects or emotions [138, 153, 131], and (2) designing features and classifiers for better classification performance (see e.g. [14, 139, 180, 127, 125, 37, 115, 183, 38, 192]).

In Publication VIII, we adopted generic low-level color, shape, and texture features to describe people’s high-level affective states evoked by viewing abstract art images. Our empirical results show that image emotions can be well recognized even using rather low-level image descriptors (see Table 6.1). In Publication IX, we proposed to utilize Multiple Kernel Learning (MKL) for classifying and retrieving abstract art images with low-level features. MKL can utilize various image features simultaneously, such that it jointly learns the feature weights and the corresponding classifier for selecting automatically the most suitable feature or a combination of them [5, 4]. Our experimental results demonstrate the advantage of MKL framework for affective image recognition, in terms of feature selection, classification performance, and interpretation.

Table 6.1. The low-level features used in Publication VIII. These features were originally used in [158] for representing emotional images.

Group of Features	Type	Dimension
First four moments	Image Statistics	48
Haralick features	Texture	28
Multiscale histograms	Texture	24
Tamura features	Texture	6
Radon transform features	Texture	12
Chebyshev statistic features	Polynomial	400
Chebyshev-Fourier features	Polynomial	32
Zernike features	Shape & Edge	72
Edge statistics features	Shape & Edge	28
Object statistics	Shape & Edge	34
Gabor filters	Shape & Edge	7

6.2 Cluster analysis on affective images

6.2.1 Motivation

The prediction of image emotions can be conceived as a multiclass classification problem that has been well attempted by *supervised learning* methods. However, a major challenge in affective image classification is that there are few emotional image databases available to the research community, as obtaining controlled experimental data is expensive in time and cost [92]. For example, the ground truth label or the most dominant emotion of an image is supposed to be agreed by a sufficient number of human evaluators whose ages, backgrounds, and cultures should be as diverse as possible.

In the absence of ground truth labels, *unsupervised learning* methods can learn patterns among data in a more impromptu fashion so that people may gain a quick insight on the data at hand. Moreover, it has been found that unsupervised methods often achieve successful intermediate results for classification tasks [23, 33]. For instance, Bressan et al. [23] described a procedure to compute a similarity matrix between painters, which was then used to infer connections between painter styles and genres. The authors in [33] proposed an automatic photo recommendation system, where they carried out a preprocessing step on photos with k-

means clustering and graph-based partitioning methods. In addition to images or pictures, clustering methods are often coupled with classification approaches for detecting emotions in videos or movies [185, 200].

6.2.2 Affective image clustering

In this section we perform an experimental study on affective image clustering. To our knowledge, there are still no similar studies that explore NMF-based methods for the cluster analysis on affective images.

Dataset. We use the `International Affective Picture System (IAPS)` [107] dataset in our clustering experiment. The `IAPS` data set is a widely-used stimulus set in emotion-related studies. It contains altogether 1182 color images that cover contents across a large variety of semantic categories, including snakes, insects, animals, landscapes, babies, guns, and accidents, among others. Each image is evaluated by subjects (males & females) on three continuously varying scales from 1 to 9 for Valence, Arousal, and Dominance. Figure 6.2 shows some of the `IAPS` images. A subset of 394 `IAPS` images have been grouped into 8 discrete emotional categories based on a psychophysical study [131], which are Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, and Sad. The emotion (ground truth label) for each image was selected as the most dominant emotion that had received the majority of votes from people. Low-level features as described in Table 6.2 are extracted from each image. The dataset is symmetrized and then binarized to be a KNN ($K = 15$) similarity matrix as the nonnegative data input.

Compared methods. There are 8 clustering methods compared in the experiment:

- The classical k-means [123] (KM) clustering,
- Projective NMF (PNMF) [194],
- Symmetric Tri-Factor Orthogonal NMF (ONMF) [48],
- NMF using Graph Random Walk (NMFR) [187],
- 1-Spectral Ratio Cheeger Cut (1-SPEC) [79],
- Data-Cluster-Data random walks (DCD) [190],

Table 6.2. The set of low-level image features used in the experiment. The features are extracted both globally and locally using the PicSOM system [105].

Index	Feature	Type	Zoning	Dims.
F1	Scalable Color	Color	Global	256
F2	Dominant Color	Color	Global	6
F3	Color Layout	Color	8×8	12
F4	5Zone-Color	Color	5	15
F5	5Zone-Colm	Color	5	45
F6	Edge Histogram	Shape	4×4	80
F7	Edge Fourier	Shape	Global	128
F8	5Zone-Edgehist	Shape	5	20
F9	5Zone-Edgecoocc	Shape	5	80
F10	5Zone-Texture	Texture	5	40

- DCD using heterogeneous initialization (DCD-heter-init), i.e. Algorithm 1 in Publication VI,

- DCD using heterogeneous co-initialization (DCD-heter-co-init), i.e. Algorithm 2 in Publication VI.

We have implemented PNMF, ONMF, NMFR, and DCD using multiplicative updates and ran each of these programs for $10,000$ iterations to ensure their convergence. We adopted the 1-SPEC software by Hein and Bühler[1] with its default setting. For k-means, we directly utilized the Matlab function *kmeans*. For DCD-heter-init, the involved methods are PNMF, NSC [45], ONMF, LSD [3], PLSI [82], DCD with 4 different Dirichlet priors ($\alpha = 1, 1.2, 2, 5$), NMFR, and 1-SPEC. For DCD-heter-co-init, the involved methods are the same as DCD-heter-init excluding the 1-SPEC method. The number of co-initialization iterations was set to 5, as in practice we found that there is no significant improvement after five rounds.

Clustering results. Figure 6.3 shows the clustering result of the compared methods. It can be seen that our new nonnegative matrix approximation methods DCD and NMFR are more advantageous over several other compared approaches on emotional image clustering, especially with larger numbers of clusters. Moreover, as the number of clusters increases, DCD methods using co-initialization have generally better perfor-

[1] http://www.ml.uni-saarland.de/code/oneSpectralClustering/oneSpectralClustering.html

mance than the original DCD and others. This makes sense in that emotional image datasets, such as IAPS, are often very “noisy", which means that the data points lie in a rather curved manifold and need to be handled by more advanced clustering techniques than traditional methods, such as the k-means.

Note that the clustering results above are rather preliminary at this stage; yet it would rather serve as a starting point for using NMF-type methods into the area of image affect analysis.

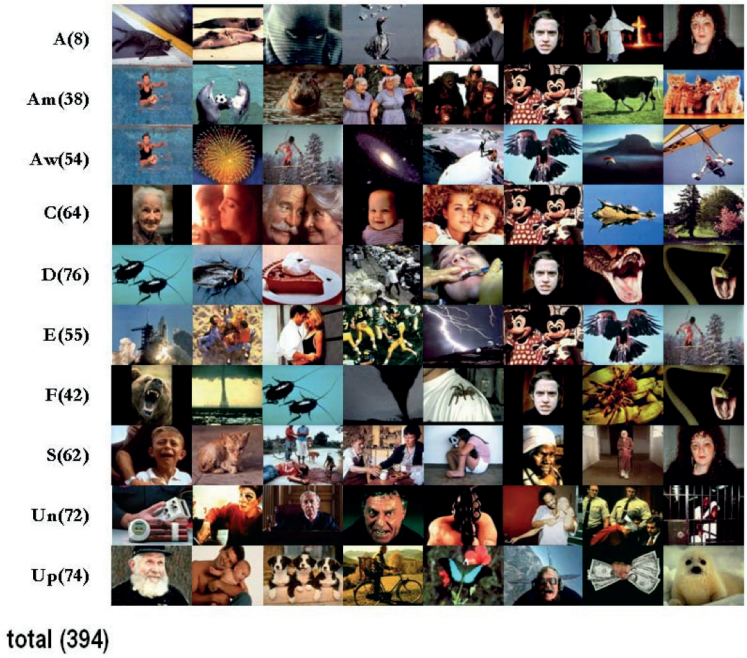


Figure 6.2. Example images from IAPS dataset (note: these example images are taken from the work [192]. A - anger, Am - amusement, Aw - awe, C - contentment, D - disgust, E - excitement, F - fear, S - sadness, Un - undifferentiated negative, Up - undifferentiated positive.

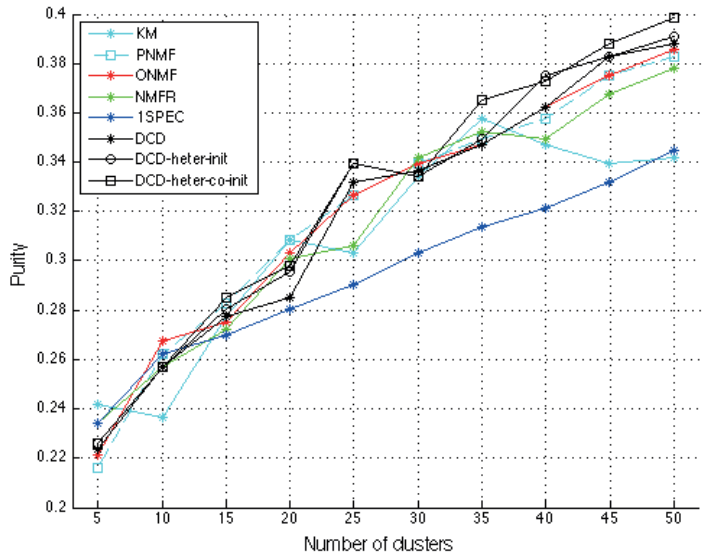


Figure 6.3. Clustering purities on the IAPS image dataset using 8 compared clustering methods with different numbers of clusters.

7. Conclusion

Since Lee and Seung's *Nature* article published in 1999, Nonnegative Matrix Factorization (NMF) has been an active research field with applications in various areas. NMF approximates a nonnegative data matrix by a product of several low-rank factorizing matrices under nonnegative constraints. The additive nature of NMF can often result in parts-based representation of the data, and this property is especially desired for feature extraction and data clustering.

This thesis presents advances in Nonnegative Matrix Decomposition (NMD) with application in cluster analysis. It reviews a higher-order class of NMF methods called Quadratic Nonnegative Matrix Factorization (QNMF), where some factorizing matrices appear twice in the approximation. Further, the thesis reviews a matrix decomposition method based on Data-Cluster-Data (DCD) random walk. DCD goes beyond matrix factorizations since it involves operations other than matrix multiplications.

Despite the advantage of QNMF, its convergence speed is mediocre. This thesis has presented an adaptive multiplicative update algorithm for QNMF, where the constant exponent in the update rules is replaced by a variable one to accelerate the convergence speed while the monotonic decrease of QNMF objective is still ensured. In addition, a co-initialization strategy has been presented, where a set of base clustering methods provide initializations for each other to improve the clustering performance. The co-initialization approach is especially suitable for those methods that require careful initializations such as the DCD.

The proposed methods have been tested on a variety of real-world datasets, such as textual documents, facial images, and protein data, etc. In particular, the thesis has presented an experimental study on clustering emotional data, where DCD and the presented co-initialization strategy have been applied for the cluster analysis on a benchmark emotional im-

age dataset. Although the experimental study is rather preliminary, the clustering results have shown that NMD-type methods are suitable for analyzing the structure of affective images. Besides, the comparison result with other existing clustering methods has also demonstrated the advantage of DCD and the co-initialization approach in terms of clustering purity.

The research on NMD is still going on. For QNMF, in addition to matrix products, nonlinear operators could be considered in the approximation, such as a nonlinear activation function that interleaves a factorizing matrix and its transpose. This type of approximation can be extended to the field of nonnegative neural networks and connected to the deep learning principle. For DCD, one might consider different priors or regularization techniques for smoothing the objective function space in order to obtain better initialization matrices. Besides, one should notice that using more informative features or a better similarity measure to construct the input similarity matrix can significantly improve the clustering results.

The presented adaptive algorithm facilitates the applications of QNMF methods. More comprehensive adaptation schemes, such as the more efficient line search, could be applied for further increasing the convergence speed of QNMF objectives. Moreover, the presented adaptive exponent technique is readily extended to other fixed-point algorithms that use multiplicative updates. As for the co-initialization strategy, the participating methods are currently chosen heuristically in the presented work. A more rigorous and computable diversity measure between clustering methods could be helpful for more efficient co-initializations.

The presented experimental study shows the suitability of NMD-type decomposition methods in affective image analysis, where DCD using co-initialization outperforms other compared methods in terms of clustering purity. The study here is rather preliminary at this stage. More large-scale affective image datasets are worth to be tested in the future. Besides, due to the high cost in obtaining the affective image labels, semi-supervised clustering could be a promising alternative for further improving the clustering performance, especially when only partial labels are available. In addition, one might investigate the feature-extraction aspect of NMD in affective image classification tasks. Because of the varying subjectivity in humans and the limit of available affective databases, the development in affective image analysis relies on the joint efforts from, for instance, artificial intelligence, computer vision, and psychology.

It is worth mentioning that the number of data clusters is assumed to be known beforehand in our thesis. Usually people pre-specify the number of clusters, K, based on their expertise. However, in many real-world applications, the value of K is unknown and needs to be estimated from the data in question. One way is to run a clustering algorithm with different values of K and choose the best value of K according to a predefined criterion. The criteria frequently used include: Minimum Description Length (MDL) criterion [151, 75], Minimum Message Length (MML) criterion [173], Bayes Information Criterion (BIC) [58], and Akaike's Information Criterion (AIC) [19, 20]. Other approaches include, for example, the gap statistics [168] and the entropy criterion [30]. Despite the various objective criteria, the most appropriate number of clusters largely relies on the parameter tuning and is application dependent.

A. Appendix

A.1 Divergence

Divergences are commonly used to measure the difference between two probability distributions P and Q. A divergence measure, denoted by $D(P||Q)$, satisfies the condition:

$$D(P||Q) \geq 0, \tag{A.1}$$

where $D(P||Q) = 0$ if and only if $P = Q$. There are various kinds of divergences. Typical measures include the Euclidean distance (Frobenius norm) and the non-normalized Kullback–Leibler divergence (I-divergence). Note that divergences are non-symmetric measurements because $D(P||Q)$ is not necessarily equal to $D(Q||P)$. For NMF problems, P corresponds to the input data matrix $\mathbf{X}$ and Q to the approximation matrix $\widehat{\mathbf{X}}$.

Table A.1 summarizes the divergence measures used in the thesis. For α-divergence, when setting $\alpha \rightarrow 1$, its limiting value leads to the I-divergence:

$$D_I(\mathbf{X}||\widehat{\mathbf{X}}) = \sum_{ij} \left(X_{ij} \ln \frac{X_{ij}}{\widehat{X}_{ij}} - X_{ij} + \widehat{X}_{ij} \right); \tag{A.2}$$

when setting $\alpha \rightarrow 0$, its limiting value gives the Dual I-divergence:

$$D_I(\widehat{\mathbf{X}}||\mathbf{X}) = \sum_{ij} \left(\widehat{X}_{ij} \ln \frac{\widehat{X}_{ij}}{X_{ij}} - \widehat{X}_{ij} + X_{ij} \right). \tag{A.3}$$

For β-divergence, when setting $\beta = 0$, it gives the I-divergence as in Eq. A.2; when setting $\beta = 1$, it gives the Euclidean distance:

$$D_{EU}(\mathbf{X}||\widehat{\mathbf{X}}) = \sum_{ij} \left(X_{ij} - \widehat{X}_{ij} \right)^2; \tag{A.4}$$

Table A.1. Divergence measure used in the thesis. $\mathbf{X}$ denotes the input data matrix and $\widehat{\mathbf{X}}$ denotes the approximation matrix.

Divergence measure	Definition
α-divergence	$D_\alpha(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \frac{1}{\alpha(1-\alpha)} \sum_{ij} \left(\alpha X_{ij} + (1-\alpha)\, \widehat{X}_{ij} - X_{ij}^{\alpha} \widehat{X}_{ij}^{1-\alpha} \right)$
β-divergence	$D_\beta(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij} \left(X_{ij} \frac{X_{ij}^{\beta} - \widehat{X}_{ij}^{\beta}}{\beta} - \frac{X_{ij}^{\beta+1} - \widehat{X}_{ij}^{\beta+1}}{\beta+1} \right)$
γ-divergence	$D_\gamma(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij} \frac{1}{\gamma(1+\gamma)} \Bigg(\ln \left(\sum_{ij} X_{ij}^{1+\gamma} \right) + \gamma \ln \left(\sum_{ij} \widehat{X}_{ij}^{1+\gamma} \right) - (1+\gamma) \ln \left(\sum_{ij} X_{ij} \widehat{X}_{ij}^{\gamma} \right) \Bigg)$
Rényi divergence	$D_\rho(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \frac{1}{\rho-1} \ln \left[\sum_{ij} \left(\frac{X_{ij}}{\sum_{ab} X_{ab}} \right)^{\rho} \left(\frac{\widehat{X}_{ij}}{\sum_{ab} \widehat{X}_{ab}} \right)^{1-\rho} \right]$, where $\rho > 0$

when setting $\beta \to -1$, its limiting value leads to the Itakura-Saito divergence:

$$D_{IS}(\mathbf{X}||\widehat{\mathbf{X}}) = \sum_{ij} \left(-\ln \frac{X_{ij}}{\widehat{X}_{ij}} + \frac{X_{ij}}{\widehat{X}_{ij}} - 1 \right). \tag{A.5}$$

For γ-divergence, when setting $\gamma \to 0$, it gives the Kullback-Leibler (KL) divergence:

$$D_{KL}(\mathbf{X}||\widehat{\mathbf{X}}) = \sum_{ij} \left(X_{ij} \ln \frac{X_{ij}}{\widehat{X}_{ij} / \sum_{ab} \widehat{X}_{ab}} \right), \tag{A.6}$$

where $\sum_{ij} X_{ij} = 1$. For Rényi divergence, when setting $\rho \to 1$, its limiting value gives the KL-divergence (i.e. Eq. A.6) as well.

Bibliography

[1] R. Albright, J. Cox, D. Duling, A. Langville, and C. Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, NCSU Technical Report Math 81706. http://meyer. math. ncsu. edu/Meyer/Abstracts/Publications.html, 2006.

[2] M. Alvira and R. Rifkin. An empirical comparison of SNoW and SVMs for face detection. A.I. memo 2001-004, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2001.

[3] R. Arora, M. Gupta, A. Kapila, and M. Fazel. Clustering by left-stochastic matrix factorization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 761–768, 2011.

[4] F.R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research (JMLR))*, 9:1179–1225, 2008.

[5] F.R. Bach, G.R.G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1–8. ACM, 2004.

[6] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. ACM press New York, 1999.

[7] G. Ball and D. Hall. ISODATA, a novel method of data analysis and pattern classification. Technical report, Technical Report NTIS, Stanford Research Institute, 1965.

[8] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–591, 2001.

[9] R. Bellmann. *Adaptive control processes: A guided tour*. Princeton University Press, 1961.

[10] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297, 1999.

[11] M.W. Berry and M. Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, 2005.

[12] 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, 2010.

[13] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *International Conference on Database Theory*, pages 217–235. Springer, 1999.

[14] N. Bianchi-Berthouze. K-DIME: An affective image filtering system. *IEEE Multimedia*, 10(3):103–106, 2003.

[15] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

[16] D. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344, 1998.

[17] I. Borg. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[18] C. Boutsidis and E. Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[19] H. Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[20] H. Bozdogan. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44(1):62–91, 2000.

[21] P.S. Bradley and U.M. Fayyad. Refining initial points for k-means clustering. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 91–99, 1998.

[22] P.S. Bradley, U.M. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 9–15, 1998.

[23] M. Bressan, C. Cifarelli, and F. Perronnin. An analysis of the relationship between painters based on their work. In *Proceedings of International Conference on Image Processing (ICIP)*, pages 113–116. IEEE, 2008.

[24] J. Brunet, P. Tamayo, T.R. Golub, and J.P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[25] G. Buchsbaum and O. Bloch. Color categories revealed by non-negative matrix factorization of Munsell color spectra. *Vision Research*, 42(5):559–563, 2002.

[26] D. Cai, X. He, J. Han, and T.S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(8):1548–1560, 2011.

[27] P. Carmona-Saez, R. Pascual-Marqui, F. Tirado, J. Carazo, and A. Pascual-Montano. Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*, 7(1):78, 2006.

[28] G.A. Carpenter, S. Grossberg, and D.B. Rosen. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4(6):759–771, 1991.

[29] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[30] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

[31] M. Chagoyen, P. Carmona-Saez, H. Shatkay, J.M. Carazo, and A. Pascual-Montano. Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics*, 7(1):41, 2006.

[32] Z. Chen, A. Cichocki, and T.M. Rutkowski. Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer disease. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages V–V. IEEE, 2006.

[33] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *Proceedings of ACM International Conference on Multimedia*, pages 291–300. ACM, 2010.

[34] M. Chu, F. Diele, R. Plemmons, and S. Ragni. Optimality, computation, and interpretation of nonnegative matrix factorizations. *SIAM Journal on Matrix Analysis*, pages 1–18, 2004.

[35] A. Cichocki, R. Zdunek, and S. Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 621–624. IEEE, 2006.

[36] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[37] C. Colombo, A. Del Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*, 6(3):38–53, 1999.

[38] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 288–301, 2006.

[39] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):1–60, 2008.

[40] W.H. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24, 1984.

[41] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[42] I.S. Dhillon and D.S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, 2001.

[43] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1–8. ACM, 2004.

[44] C. Ding, X. He, and H.D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of SIAM International Conference on Data Mining (SDM)*, pages 606–610, 2005.

[45] C. Ding, T. Li, and M.I. Jordan. Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding. In *Proceedings of International Conference on Data Mining (ICDM)*, pages 183–192. IEEE, 2008.

[46] C. Ding, T. Li, and M.I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(1):45–55, 2010.

[47] C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927, 2008.

[48] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 126–135. ACM, 2006.

[49] D.L. Donoho and V.C. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–8. MIT Press, 2004.

[50] K. Drakakis, S. Rickard, R. de Fréin, and A. Cichocki. Analysis of financial data using non-negative matrix factorization. *International Mathematical Forum*, 3(38):1853–1870, 2008.

[51] J.C. Dunn. A fuzzy relative of the ISODATA process and its use in detection compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[52] M. Erisoglu, N. Calis, and S. Sakallioglu. A new algorithm for initial cluster centers in k-means algorithm. *Pattern Recognition Letters*, 32(14):1701–1705, 2011.

[53] T. Feng, S.Z. Li, H. Shum, and H. Zhang. Local non-negative matrix factorization as a visual representation. In *Proceedings of International Conference on Development and Learning*, pages 178–183. IEEE, 2002.

[54] C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[55] D.H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987.

[56] I.K. Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Laboratory, 2002.

[57] E.W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[58] C. Fraley and A.E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[59] A.L.N. Fred and A.K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.

[60] Y. Gao and G. Church. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21(21):3970–3975, 2005.

[61] R. Gaujoux and C. Seoighe. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infection, Genetics and Evolution*, 12(5):913–921, 2012.

[62] N. Gillis. The why and how of nonnegative matrix factorization. Technical report, Machine Learning and Pattern Recognition Series, arXiv:1401.5226, 2014.

[63] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. In *Proceedings of International Conference on Data Engineering (ICDE)*, pages 341–352. IEEE, 2005.

[64] D.E. Goldberg and J.H. Holland. Genetic algorithms and machine learning. *Machine Learning*, 3(2):95–99, 1988.

[65] G.H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970.

[66] E.F. Gonzalez and Y. Zhang. Accelerating the Lee-Seung algorithm for nonnegative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, 2005.

[67] N. Guan, D. Tao, Z. Luo, and B. Yuan. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Transactions on Image Processing*, 20(7):2030–2048, 2011.

[68] N. Guan, D. Tao, Z. Luo, and B. Yuan. Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 23(7):1087–1099, 2012.

[69] S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 73–84. ACM, 1998.

[70] D. Guillamet, M. Bressan, and J. Vitria. A weighted non-negative matrix factorization for local representations. In *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–942–I–947. IEEE, 2001.

[71] D. Guillamet, J. Vitria, and B. Schiele. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognition Letters*, 24(14):2447–2454, 2003.

[72] L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11(9):1074–1085, 1992.

[73] L.O. Hall, I.B. Ozyurt, and J.C. Bezdek. Clustering with a genetically optimized approach. *IEEE Transactions on Evolutionary Computation*, 3(2):103–112, 1999.

[74] A. Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006.

[75] M.H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[76] D.R. Hardoon, J. Shawe-Taylor, A. Ajanki, K. Puolamäki, and S. Kaski. Information retrieval by inferring implicit queries from eye movements. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 179–186, 2007.

[77] T. Hazan, S. Polak, and A. Shashua. Sparse image coding using a 3D non-negative tensor factorization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 50–57. IEEE, 2005.

[78] Z. He, S. Xie, R. Zdunek, G. Zhou, and A. Cichocki. Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering. *IEEE Transactions on Neural Networks (TNN)*, 22(12):2117–2131, 2011.

[79] M. Hein and T. Bühler. An inverse power method for nonlinear eigenproblems with applications in 1-Spectral clustering and sparse PCA. In *Advances in Neural Information Processing Systems (NIPS)*, pages 847–855, 2010.

[80] M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *European Signal Processing Conference*, pages 1–4, 2005.

[81] M.D. Hoffman, D.M. Blei, and P.R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 439–446, 2010.

[82] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57. ACM, 1999.

[83] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research (JMLR)*, 5:1457–1469, 2004.

[84] P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[85] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[86] D.R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[87] N. Iam-On, T. Boongoen, and S. Garrett. Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. In *Proceedings of International Conference on Discovery Science (DS)*, pages 222–233. Springer, 2008.

[88] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[89] A.K. Jain, J. Mao, and K.M. Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[90] S. Jia and Y. Qian. Constrained nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):161–173, 2009.

[91] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[92] D. Joshi, R. Datta, E. Fedorovskaya, Q. Luong, J.Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.

[93] I. Jung, J. Lee, S. Lee, and D. Kim. Application of nonnegative matrix factorization to improve profile-profile alignment features for fold recognition and remote homolog detection. *BMC Bioinformatics*, 9(1):298, 2008.

[94] G. Karypis, E. Han, and V. Kumar. CHAMELEON: Hierrchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[95] L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[96] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *ACM SIGIR Forum*, 37(2):18–28, 2003.

[97] S.S. Khan and A. Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25(11):1293–1302, 2004.

[98] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[99] P.M. Kim and B. Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research*, 13(7):1706–1718, 2003.

[100] S. Kim, Y.N. Rao, D. Erdogmus, J.C. Sanchez, M.A.L. Nicolelis, and J.C. Principe. Determining patterns in neural activity for reaching movements using nonnegative matrix factorization. *EURASIP Journal on Applied Signal Processing*, 2005:3113–3121, 2005.

[101] Y. Kim and S. Choi. A method of initialization for nonnegative matrix factorization. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 537–540, 2007.

[102] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[103] T. Kohonen. *Self-organizing maps*. Springer, 2001.

[104] K. Krishna and M. Narasimha Murty. Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(3):433–439, 1999.

[105] J. Laaksonen, M. Koskela, and E. Oja. PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. *IEEE Transactions on Neural Networks (TNN)*, 13(4):841–853, 2002.

[106] G.N. Lance and W.T. Williams. A general theory of classificatory sorting strategies 1. Hierarchical systems. *The Computer Journal*, 9(4):373–380, 1967.

[107] P.J. Lang, M.M. Bradley, and B.N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, University of Florida, Gainesville, FL, 2008.

[108] A.N. Langville, C.D. Meyer, R. Albright, J. Cox, and D. Duling. Initializations for the nonnegative matrix factorization. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 23–26. Citeseer, 2006.

[109] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[110] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562, 2001.

[111] H. Lee and S. Choi. Group nonnegative matrix factorization for EEG classification. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 320–327, 2009.

[112] H. Lee, A. Cichocki, and S. Choi. Nonnegative matrix factorization for motor imagery EEG classification. In *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, pages 250–259. Springer, 2006.

[113] H. Lee, J. Yoo, and S. Choi. Semi-supervised nonnegative matrix factorization. *Signal Processing Letters*, 17(1):4–7, 2010.

[114] J. Lee, S. Park, C. Ahn, and D. Kim. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1):20–34, 2009.

[115] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009.

[116] S.Z. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–207–I–212. IEEE, 2001.

[117] T. Li and C. Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *Proceedings of International Conference on Data Mining (ICDM)*, pages 362–371. IEEE, 2006.

[118] A. Likas, N. Vlassis, and J. J Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003.

[119] C. Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks (TNN)*, 18(6):1589–1596, 2007.

[120] C. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[121] H. Liu, Z. Wu, X. Li, D. Cai, and T.S. Huang. Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1299–1311, 2012.

[122] Y. Liu, D. Zhang, G. Lu, and W. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[123] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[124] W. Lu, W. Sun, and H. Lu. Robust watermarking based on DWT and nonnegative matrix factorization. *Computers & Electrical Engineering*, 35(1):183–188, 2009.

[125] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M.G. Newman, and J.Z. Wang. On shape and the computability of emotions. In *Proceedings of International Conference on Multimedia*, pages 229–238. ACM, 2012.

[126] T. Lux and M. Marchesi. Volatility clustering in financial markets: A microsimulation of interacting agents. *International Journal of Theoretical and Applied Finance*, 3(04):675–702, 2000.

[127] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of International Conference on Multimedia*, pages 83–92, 2010.

[128] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[129] G.J. McLachlan and K.E. Basford. *Mixture models. Infrence and applications to clustering*. New York: Marcel Dekker, 1988.

[130] L. Miao and H. Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777, 2007.

[131] J.A. Mikels, B.L. Fredrickson, G.R. Larkin, C.M. Lindberg, S.J. Maglio, and P.A. Reuter-Lorenz. Emotional category data on images from the International Affective Picture System. *Behavior Research Methods*, 37(4):626–630, 2005.

[132] V. Monga and M.K. Mihçak. Robust and secure image hashing via non-negative matrix factorizations. *IEEE Transactions on Information Forensics and Security*, 2(3):376–390, 2007.

[133] M. Mørup and L.K. Hansen. Archetypal analysis for machine learning. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 172–177. IEEE, 2010.

[134] M. Mørup and L.K. Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, 2012.

[135] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[136] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, 2002.

[137] O. Okun and H. Priisalu. Fast nonnegative matrix factorization and its application for protein fold recognition. *EURASIP Journal on Advances in Signal Processing*, 2006(1–8), 2006.

[138] C.E. Osgood, G.J. Suci, and P. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, 1957.

[139] L. Ou, M.R. Luo, A. Woodcock, and A. Wright. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research & Application*, 29(3):232–240, 2004.

[140] M. Ouhsain and A.B. Hamza. Image watermarking scheme using nonnegative matrix factorization and wavelet transform. *Expert Systems with Applications*, 36(2):2123–2129, 2009.

[141] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[142] 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, 1994.

[143] A. Pascual-Montano, J.M. Carazo, K. Kochi, D. Lehmann, and R.D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(3):403–415, 2006.

[144] A. Pascual-Montano, P. Carmona-Sáez, R.D. Pascual-Marqui, F. Tirado, and J.M. Carazo. Two-way clustering of gene expression profiles by sparse matrix factorization. In *Proceedings of IEEE Computational Systems Bioinformatics Conference Workshops*, pages 103–104. IEEE, 2005.

[145] V.P. Pauca, J. Piper, and R.J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications*, 416(1):29–47, 2006.

[146] V.P. Pauca, F. Shahnaz, M.W. Berry, and R.J. Plemmons. Text mining using nonnegative matrix factorizations. In *Proceedings of SIAM International Conference on Data Mining*, pages 452–456, 2004.

[147] P. Pehkonen, G. Wong, and P. Törönen. Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, 6(1):162, 2005.

[148] Q. Qi, Y. Zhao, M. Li, and R. Simon. Non-negative matrix factorization of gene expression profiles: a plug-in for BRB-ArrayTools. *Bioinformatics*, 25(4):545–547, 2009.

[149] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[150] B. Ribeiro, C. Silva, A. Vieira, and J. Neves. Extracting discriminative features using non-negative matrix factorization in financial distress data. In *Proceedings of Internatonal Conference on Adaptive and Natural Computing Algorithms*, pages 537–547. Springer, 2009.

[151] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[152] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[153] J.A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[154] R. Salakhutdinov and S. Roweis. Adaptive overrelaxed bound optimization methods. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 664–671, 2003.

[155] M.N. Schmidt and M. Mørup. Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation*, pages 700–707. Springer, 2006.

[156] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[157] F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[158] L. Shamir, T. Macura, N. Orlov, D.M. Eckley, and I.G. Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 7:1–17, 2010.

[159] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):888–905, 2000.

[160] P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Blind Signal Separation*, pages 494–499. Springer, 2004.

[161] 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*, pages 177–180. IEEE, 2003.

[162] A.WM. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[163] M.W. Spratling. Learning image components for object recognition. *Journal of Machine Learning Research (JMLR))*, 7:793–815, 2006.

[164] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research (JMLR)*, 3:583–617, 2003.

[165] P. Tamayo, D. Scanfeld, B.L. Ebert, M.A. Gillette, C.W.M. Roberts, and J.P. Mesirov. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proceedings of the National Academy of Sciences*, 104(14):5959–5964, 2007.

[166] Z. Tang, S. Wang, X. Zhang, W. Wei, and S. Su. Robust image hashing for tamper detection using non-negative matrix factorization. *Journal of Ubiquitous Convergence and Technology*, 2(1):18–26, 2008.

[167] J.C. Thøgersen, M. Mørup, S. Damkiær, S. Molin, and L. Jelsbak. Archetypal analysis of diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC bioinformatics*, 14(1):279, 2013.

[168] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[169] E. Tjioe, M. Berry, and R. Homayouni. Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization). *BMC Bioinformatics*, 11(Suppl 6):S14, 2010.

[170] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

[171] T. Virtanen, A.T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1825–1828. IEEE, 2008.

[172] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[173] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(3):240–265, 1987.

[174] D. Wang, T. Li, and C. Ding. Weighted feature subset non-negative matrix factorization and its applications to document understanding. In *Proceedings of SIAM International Conference on Data Mining*, pages 541–550. IEEE, 2010.

[175] D. Wang, T. Li, S. Zhu, and C. Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314. ACM, 2008.

[176] F. Wang, C. Tan, A.C. König, and P. Li. Efficient document clustering via online nonnegative matrix factorizations. In *Proceedings of SIAM International Conference on Data Mining*, pages 908–919. SIAM, 2011.

[177] G. Wang, A.V. Kossenkov, and M.F. Ochs. Ls-nmf: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, 7(1):175, 2006.

[178] J. Wang, X. Wang, and X. Gao. Non-negative matrix factorization by maximizing correntropy for cancer clustering. *BMC Bioinformatics*, 14(1):107, 2013.

[179] W. Wang and Q. He. A survey on emotional semantic image retrieval. In *Proceedings of International Conference on Image Processing (ICIP)*, pages 117–120, 2008.

[180] W. Wang, Y. Yu, and S. Jiang. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, pages 3534–3539, 2006.

[181] Y. Wang and Y. Jia. Fisher non-negative matrix factorization for learning local features. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, pages 27–30. Citeseer, 2004.

[182] S. Wild, J. Curry, and A. Dougherty. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37(11):2217–2232, 2004.

[183] Q. Wu, C. Zhou, and C. Wang. Content-based affective image classification and retrieval using support vector machines. In *Affective Computing and Intelligent Interaction*, pages 239–247. Springer, 2005.

[184] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.

[185] M. Xu, J.S. Jin, S. Luo, and L. Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of ACM International Conference on Multimedia*, pages 677–680. ACM, 2008.

[186] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 267–273. ACM, 2003.

[187] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja. Clustering by nonnegative matrix factorization using graph random walk. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1088–1096, 2012.

[188] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(5):734–749, 2010.

[189] Z. Yang and E. Oja. Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Transactions on Neural Networks (TNN)*, 22(12):1878–1891, 2011.

[190] Z. Yang and E. Oja. Clustering by low-rank doubly stochastic matrix decomposition. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 831–838, 2012.

[191] Z. Yang and E. Oja. Quadratic nonnegative matrix factorization. *Pattern Recognition*, 45(4):1500–1510, 2012.

[192] V. Yanulevskaya, J.C. Van Gemert, K. Roth, A.K. Herbold, N. Sebe, and J.M. Geusebroek. Emotional valence categorization using holistic image features. In *Proceedings of International Conference on Image Processing (ICIP)*, pages 101–104. IEEE, 2008.

[193] S.X. Yu and J. Shi. Multiclass spectral clustering. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 313–319. IEEE, 2003.

[194] Z. Yuan and E. Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proceedings of Scandinavian conference on Image Analysis (SCIA)*, pages 333–342. Springer-Verlag, 2005.

[195] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks (TNN)*, 17(3):683–695, 2006.

[196] D. Zhang, Z. Zhou, and S. Chen. Semi-supervised dimensionality reduction. In *Proceedings of SIAM International Conference on Data Mining (SDM)*, pages 629–634, 2007.

[197] H. Zhang, M. Koskela, and J. Laaksonen. Report on forms of enriched relevance feedback. Technical report, Technical Report TKK-ICS-R10, Helsinki University of Technology, Department of Information and Computer Science (November 2008), 2008.

[198] H. Zhang, M. Sjöberg, J. Laaksonen, and E. Oja. A multimodal information collector for content-based image retrieval system. In *Proceedings of International Conference on Neural Information Processing (ICONIP)*, pages 737–746. Springer, 2011.

[199] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 103–114. ACM, 1996.

[200] S. Zhao, H. Yao, X. Sun, X. Jiang, and P. Xu. Flexible presentation of videos based on affective content analysis. In *Advances in Multimedia Modeling*, pages 368–379. Springer, 2013.

[201] Z. Zheng, J. Yang, and Y. Zhu. Initialization enhancer for non-negative matrix factorization. *Engineering Applications of Artificial Intelligence*, 20(1):101–110, 2007.

[202] Z. Zhu, Z. Yang, and E. Oja. Multiplicative updates for learning with stochastic matrices. In *Proceedings of Scandinavian conference on Image Analysis (SCIA)*, pages 143–152. Springer, 2013.

Aalto-DD 127/2014

ISBN 978-952-60-5828-3
ISBN 978-952-60-5829-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**