Mach Learn (2015) 99:137–163
DOI 10.1007/s10994-014-5470-z

A unified statistical approach to non-negative matrix factorization and probabilistic latent semantic indexing

Karthik Devarajan · Guoli Wang · Nader Ebrahimi

Received: 3 October 2012 / Accepted: 4 October 2014 / Published online: 15 November 2014

Abstract Non-negative matrix factorization (NMF) is a powerful machine learning method for decomposing a high-dimensional nonnegative matrix V into the product of two nonnegative matrices, W and H, such that $V \sim WH$. It has been shown to have a parts-based, sparse representation of the data. NMF has been successfully applied in a variety of areas such as natural language processing, neuroscience, information retrieval, image processing, speech recognition and computational biology for the analysis and interpretation of large-scale data. There has also been simultaneous development of a related statistical latent class modeling approach, namely, probabilistic latent semantic indexing (PLSI), for analyzing and interpreting co-occurrence count data arising in natural language processing. In this paper, we present a generalized statistical approach to NMF and PLSI based on Renyi's divergence between two non-negative matrices, stemming from the Poisson likelihood. Our approach unifies various competing models and provides a unique theoretical framework for these methods. We propose a unified algorithm for NMF and provide a rigorous proof of monotonicity of multiplicative updates for W and H. In addition, we generalize the relationship between NMF and PLSI within this framework. We demonstrate the applicability and utility of our approach as well as its superior performance relative to existing methods using real-life and simulated document clustering data.

Keywords Nonnegative matrix factorization · Probabilistic latent semantic indexing · Renyi's divergence · λ-log-likelihood · EM algorithm · Biomedical informatics

Editor: Kristian Kersting.

K. Devarajan (✉)
Department of Biostatistics and Bioinformatics, Fox Chase Cancer Center, Temple University Health System, Philadelphia, PA 19111, USA
e-mail: karthik.devarajan@fccc.edu

G. Wang
3M Health Information Systems, Bethesda, MD 20814, USA

N. Ebrahimi
Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA

1 Introduction

Nonnegative matrix factorization (NMF) was introduced as an unsupervised parts-based learning paradigm in which a nonnegative matrix V is decomposed into the product of two nonnegative matrices, W and H, such that $V \sim WH$, by a multiplicative updates algorithm (Lee and Seung 2001). In the past decade, NMF has been widely used in a variety of areas including natural language processing such as text mining and document clustering (Pauca et al. 2004; Shahnaz and Berry 2004; Shahnaz et al. 2006; Chagoyen et al. 2006), neuroscience (Cheung and Tresch 2005; Devarajan and Cheung 2012), information retrieval (Tsuge et al. 2001; Xu et al. 2003), image processing and facial pattern recognition (Li et al. 2001; Buciu and Pitas 2004), sparse coding (Hoyer 2002, 2003; Liu et al. 2003), speech recognition (Behnke 2003; Cho et al. 2003), video summarization (Cooper and Foote 2002), and Internet research (Lu et al. 2003; Mao and Saul 2004). More recently, this approach has found its way into the domain of computational biology, particularly in the analysis and interpretation of high-throughput biological data (Devarajan and Ebrahimi 2005, 2008; Okun and Priisalu 2006; Devarajan 2006, 2008, 2011a, b; Qi et al. 2009; Zhang et al. 2011; Gaujoux and Seoighe 2012). For a complete review of the applications of NMF, the interested reader is referred to Devarajan (2008) and references therein. These developments in NMF have been paralleled by the formulation of a statistical latent class modeling approach called probabilistic latent semantic indexing (PLSI). PLSI is a model-based extension of latent semantic indexing and is used for analyzing and interpreting co-occurrence count data arising in text mining and document clustering applications (Hoffman 2001).

Lee and Seung (2001) outlined algorithms for NMF based on the Poisson and Gaussian likelihoods. They applied it to text mining and facial pattern recognition. Since its introduction, several variations and extensions of their algorithm have been proposed in the literature. For instance, Hoyer (2004), Shahnaz et al. (2006), Pascual-Montano et al. (2006) and Berry et al. (2007) extended NMF to include sparseness constraints. Wang et al. (2006) developed LS-NMF that incorporated variability in the data. Cichocki et al. (2006, 2008, 2009, 2011) extensively developed a series of generalized algorithms for NMF based on α- and β-divergences. In addition, Dhillon and Sra (2005) and Kompass (2007) have proposed generalized divergence measures for NMF. Cheung and Tresch (2005) and Devarajan and Cheung (2012) extended the NMF algorithm to include members of the exponential family of distributions while Devarajan and Ebrahimi (2005, 2008), Devarajan (2006, 2008, 2011a, b) formulated a generalized approach to NMF based on the Poisson likelihood that included various well-known distance measures as special cases. Ding et al. (2008) showed the relationship between NMF and PLSI while Ding et al. (2010) proposed a Bayesian non-parametric approach to NMF. Lin (2007), Cichocki et al. (2007), Cichocki and Phan (2009), Cichocki et al. (2009), Wang and Li (2010), Févotte and Idier (2011), Gillis and Glineur (2010, 2012) and Zhou et al. (2012) have developed efficient algorithms for various divergence measures used in NMF. The work of Cichocki et al. (2009) is a detailed reference on this subject. In previous work (Devarajan and Ebrahimi 2005, 2008), we applied NMF based on Renyi's divergence between two non-negative matrices to gene expression data from cancer microarray studies. Renyi's divergence is indexed by a parameter γ and represents a continuum of divergence measures based on the choice of this parameter (Renyi 1970). In this paper, we propose a unique theoretical framework for NMF and PLSI based on Renyi's divergence between two non-negative matrices, related to the Poisson likelihood. This model-based approach includes several well-known divergence measures as special cases and is also related to some recently proposed divergence measures, thus unifying various competing models into a single statistical framework. We describe a generalized algorithm for NMF based on Renyi's divergence

and provide a rigorous proof of its convergence using the Expectation-Maximization (EM) algorithm (Dempster et al. 1977). Furthermore, we generalize the equivalence of NMF and PLSI using our framework and show that the currently known relationship between these methods is embedded within this framework as a special case. Throughout this paper, NMF refers to that based on the Poisson likelihood unless specified otherwise.

We demonstrate the utility and applicability of our generalized approach using several real-life and simulated data sets from text mining and document clustering. We use consensus clustering to quantitatively evaluate the homogeneity and accuracy of clustering for different choices of the parameter γ using a variety of metrics. Our methods are implemented in high-performance computing clusters using message-passing interface. The extension of our methods to other problems of interest is straightforward.

This paper is organized as follows. Section 2 gives an overview of the fundamental concepts and provides a brief discussion of Renyi's divergence and related divergence measures. In Sect. 3, we explore the applicability of these measures in the context of NMF, propose our unified NMF algorithm and provide update rules based on Renyi's divergence. In addition, we generalize the equivalence of NMF and PLSI within the unified framework provided by Renyi's divergence. In Sect. 4, we describe the quantitative evaluation of clustering based on our approach and in Sect. 5, we illustrate our methods in detail by applying it to a variety of real-life and simulated document clustering data sets. The last section provides a discussion and concluding remarks. Detailed proofs of the theoretical results presented in Sect. 3 are relegated to the Appendix.

2 A generalized divergence measure

Consider the problem of discriminating between two probability models F and G for a random prospect X that ranges over the space S. Let f and g be the probability density (mass) functions corresponding to F and G, respectively. Given an observation $X = x$, the logarithm of the likelihood ratio $\log\left[\frac{f(x)}{g(x)}\right]$ quantifies the information in $X = x$ in favor of F against G. Suppose that x is not given and there is not specific information on the whereabouts of x, other than $x \epsilon S$, then the mean observation per x from F for the discrimination information between F and G is

$$K(f : g) = \int \left(\log \frac{f(x)}{g(x)}\right) dF(x), \tag{2.1}$$

given that F is absolutely continuous with respect to G. The discrimination information function (2.1) is a measure for comparing two distributions, and is referred to as the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951; Kullback 1959). See Ebrahimi and Soofi (2004) for properties of this measure.

Renyi's divergence, which is referred to as the information divergence of order γ between two distributions F and G, is defined by

$$R_\gamma(f : g) = \frac{1}{\gamma - 1} \log \int \left(\frac{f(x)}{g(x)}\right)^{\gamma - 1} dF(x), \tag{2.2}$$

where $\gamma \neq 1$ (Renyi 1970). Various well-known distance measures, including KL divergence, arise from Renyi's divergence as special cases. An important feature of Renyi's divergence is that it is invariant under any nonsingular transformation, $Y = T(X)$, on the original data. That is, for any γ, $R_\gamma(f_X : g_X) = R_\gamma(f_Y : g_Y)$.

For two Poisson random variables with parameters m_1 and m_2, i.e., $f(x) = \frac{e^{-m_1} m_1^x}{x!}$ and $g(x) = \frac{e^{-m_2} m_2^x}{x!}$, one can easily show that

$$R_\gamma(f : g) = \frac{1}{\gamma - 1}\left(-\gamma m_1 - (1-\gamma)m_2 + m_1^\gamma m_2^{1-\gamma}\right). \tag{2.3}$$

It is well-known that Renyi's divergence reduces to KL divergence for the limiting case $\gamma \to 1$,

$$K(f : g) = m_1 \log\left(\frac{m_1}{m_2}\right) - m_1 + m_2. \tag{2.4}$$

In the special case that $\gamma = \frac{1}{2}$, $R_{\frac{1}{2}}(f : g) = R_{\frac{1}{2}}(g : f) = \left(\sqrt{m}_1 - \sqrt{m}_2\right)^2$. This is the well-known Bhattacharya distance and is a symmetric measure (Freeman and Tukey 1950). For this case, it is also the logarithm of the squared Matusita or Hellinger distance (Matusita 1954). If $\gamma = 2$, $R_2(f : g) = \frac{(m_1 - m_2)^2}{m_2}$, which is the Pearson Chi-squared estimator. When $\gamma = -1$, we obtain the modified Chi-squared estimator due to Neyman (1949). And for $\gamma = \frac{5}{3}$, we obtain the Cressie-Read distance estimator (Cressie et al. 2003).

Our motivation for a generalized approach to NMF using the Poisson likelihood is based on the power-divergence family of statistics (Agresti 1990). It is given by

$$\phi_\lambda(m_1, m_2) = \frac{2}{\lambda(\lambda+1)} m_1 \left[\left(\frac{m_1}{m_2}\right)^\lambda - 1\right] \tag{2.5}$$

for $\lambda \neq -1$ and $\lambda \neq 0$. This family of measures and its variants have been extensively studied in the statistical literature in the context of discrete multivariate data analysis (see Cressie et al. 2003 and references therein). It is straightforward to obtain Renyi's divergence and all the special cases outlined above via reparametrizations in (2.5). For example in (2.5), $\lambda \to 0$ corresponds to $\gamma \to 1$ in (2.3). Similarly, $\lambda = -\frac{1}{2}, -2$ and 1 correspond to $\gamma = \frac{1}{2}, -1$ and 2, respectively in (2.3).

Several other generalized divergence measures have been proposed in the machine learning literature recently within the context of NMF. These include Cichocki et al. (2006, 2008), Devarajan and Ebrahimi (2005) and Kompass (2007). Cichocki and Amari (2010) also discussed several divergence measures with potential applications in NMF. Unlike other measures, Renyi's divergence (2.3) and the power divergence family (2.5) are motivated by an underlying statistical model.

3 Methods

Text mining and document clustering are concerned with the recognition of patterns or similarities in natural language text. Consider a corpus of documents that is summarized as a $p \times n$ matrix V in which the rows represent the terms in the vocabulary and the columns correspond to the documents in the corpus. The entries of V denote the frequencies of words in each document. In document clustering studies, the number of terms p is typically in the thousands and the number of documents n is typically in the hundreds. The objective is to identify subsets of semantic categories and to cluster the documents based on their association

with these categories. To this end, we propose to find a small number of metaterms, each defined as a nonnegative linear combination of the p terms. This is accomplished via a decomposition of the frequency matrix V into two matrices W and H with nonnegative entries such that $V \sim WH$, where W has size $p \times k$, with each of k columns defining a metaterm and H has size $k \times n$, with each of n columns representing the metaterm frequency pattern of the corresponding document. The rank k of the factorization is chosen so that $(n + p)k < np$. Here, the entry w_{ia} in the matrix W is the coefficient of term i in metaterm a and the entry h_{aj} in the matrix H quantifies the influence of metaterm a in document j.

The metaterms and the metaterm frequency patterns have a sparse representation, potentially representing local hidden variables or clusters. These clusters are subgroups of terms that co-occur in subgroups of documents. The perception of the whole is simply a combination of the parts represented by these basis vectors. Since the data are presented as frequency of occurrence of terms for each document, NMF provides a more natural representation of the metaterms and metaterm frequency patterns unlike other dimension reduction methods. Moreover, the nonnegative coefficients in each metaterm are easily interpretable as the relative contribution of terms. A more thorough discussion of the interpretation of the factorization and the nonnegativity constraints in NMF can be found in Devarajan (2008). In this paper, our focus will be on clustering documents. A notable example of such an application is in biomedical informatics involving mining of the biomedical literature. Chagoyen et al. (2006) describe the application of NMF to create literature profiles from a corpus of documents relevant to large sets of genes and proteins using common semantic features extracted from the corpus.

3.1 A unified algorithm for NMF

In order to find an approximate factorization for the matrix V, we first need to define functions that quantify the quality of the approximation. In general, such a function can be constructed using some measure of distance between any two nonnegative matrices, say A and B. Examples of such measures include Euclidean distance and KL divergence, obtained based on the Gaussian and Poisson likelihoods, respectively. The latter can be derived based on reconstruction of an image represented by the matrix A from the matrix B by the addition of Poisson noise, i.e,

$$A = B + \epsilon \tag{3.1}$$

where ϵ is a Poisson random variable. This formulation was originally described in Lee and Seung (1999) for text mining applications involving count data as well as for facial pattern recognition. We generalize this approach by using Renyi's divergence $R_\gamma(f : g)$ related to the Poisson likelihood of generating A from B, as described in Sect. 2. Specifically, using (2.3), our measure is

$$D^*_\gamma(A||B) = \frac{1}{\gamma - 1} \sum_{i,j} \left[A_{ij}^\gamma B_{ij}^{1-\gamma} - \gamma A_{ij} - (1 - \gamma) B_{ij} \right] \tag{3.2}$$

which can be generalized by re-defining it as

$$D^*_\gamma(A||B) = \frac{1}{\gamma(\gamma - 1)} \sum_{i,j} \left[A_{ij}^\gamma B_{ij}^{1-\gamma} - \gamma A_{ij} - (1 - \gamma) B_{ij} \right] \tag{3.3}$$

where $\gamma \neq 1$ and $\gamma \neq 0$. This includes Renyi's divergence as defined by (2.3) and its special cases. If $\gamma \to 1$, then D_γ^* in (3.3) becomes KL divergence (Eq. 2.4),

$$K(A : B) = \sum_{i,j} A_{ij} \log\left(\frac{A_{ij}}{B_{ij}}\right) - A_{ij} + B_{ij}. \tag{3.4}$$

This coincides with the measure proposed by Lee and Seung (2001). If $\gamma \to 0$, we obtain dual KL divergence (Kullback 1959). For $\gamma \neq 1$ and $\gamma \neq 0$, one can ignore $\frac{1}{\gamma(\gamma-1)}$ in (3.3) and define the function

$$D_\gamma(A||B) = \begin{cases} \sum_{i,j} A_{ij}^\gamma B_{ij}^{1-\gamma} - \gamma A_{ij} - (1-\gamma)B_{ij}, & \gamma > 1 \\ \sum_{i,j} \gamma A_{ij} + (1-\gamma)B_{ij} - A_{ij}^\gamma B_{ij}^{1-\gamma}, & 0 < \gamma < 1. \end{cases} \tag{3.5}$$

Similarly for $\gamma \neq 1$, one can ignore $\frac{1}{\gamma - 1}$ in (3.2) and define the function

$$D_\gamma(A||B) = \sum_{i,j} -\gamma A_{ij} - (1-\gamma)B_{ij} + A_{ij}^\gamma B_{ij}^{1-\gamma}, \ \gamma < 0. \tag{3.6}$$

Thus, for any information measure which is proportional to Renyi's divergence we obtain Eq. (3.5). Both β-divergence (Cichocki et al. 2006) and its simplified version proposed by Kompass (2007) can be shown to be related to Renyi's divergence of order γ between matrices A and B via non-linear transformations of A, B and γ. These measures contain the Gaussian and Poisson models as special cases while the former also embeds the so called Itakura-Saito (IS) divergence that is suitable for modeling signal-dependent noise (Cheung and Tresch 2005; Devarajan and Cheung 2012). The relationship between these divergence measures provides a unified view of various algorithms for NMF from the perspective of different statistical models. In the case of non-normal data such as those arising in document clustering, Renyi's divergence is a flexible choice in decomposing a frequency matrix.

For a given document frequency matrix V, we now formally consider a method for finding nonnegative matrices W and H such that $V \approx WH$. In our setup, this is equivalent to minimizing $D_\gamma(V||WH)$ in (3.5) and (3.6) with respect to W and H, subject to the constraints $W, H \geq 0$. In this formulation, we observe that for a given γ, $D_\gamma(V||WH)$ is not convex in both variables (V and WH) together. Hence, the algorithm will only converge to a local minima. There are many techniques such as gradient descent and conjugate gradient from numerical optimization that can be applied to find the minima. In this paper, we use multiplicative update rules, similar to that in Lee and Seung (1999). For a given γ, we will start with random initial values for W and H and iterate until convergence, i.e, iterate until $|D_\gamma^{(i)}(V||WH) - D_\gamma^{(i-1)}(V||WH)| < \delta$ where δ is a pre-specified threshold between 0 and 1 and i denotes the iteration number.

Theorem 1 *For $\gamma \in \Re\backslash\{0,1\}$, the measure $D_\gamma(V||WH)$ is non-increasing under the multiplicative update rules for W and H given by*

$$H_{aj}^{t+1} = H_{aj}^{t} \left(\frac{\sum_i \left(\frac{V_{ij}}{\sum_b W_{ib} H_{bj}^t} \right)^\gamma W_{ia}}{\sum_i W_{ia}} \right)^{1/\gamma} \tag{3.7}$$

and

$$W_{ia}^{t+1} = W_{ia}^{t} \left(\frac{\sum_i \left(\frac{V_{ij}}{\sum_b W_{ib}^t H_{bj}} \right)^\gamma H_{aj}}{\sum_j H_{aj}} \right)^{1/\gamma}. \tag{3.8}$$

This measure is also invariant under these updates if and only if W and H are at a stationary point of the divergence.

A proof of the monotonicity of these updates is given in Cichocki et al. (2008). However, the auxiliary function used in their proof will be properly defined only if each element of V is assumed to be positive. This violates the non-negativity requirement on V. An alternate proof that overcomes this problem by considering different ranges of the parameter γ is provided in the Appendix.

3.2 Equivalence of NMF and PLSI: a generalization

PLSI is a method for modeling co-occurrence data arising in natural language processing such as text mining and document clustering. It is based on a statistical latent class model called the *aspect model* for the analysis of count data (Hoffman 2001). PLSI employs the likelihood principle and results in a factor representation of the data such as in NMF, thereby defining a proper generative model of the data. The relationship between NMF and PLSI has been described elsewhere in the literature (Buntine 2002; Ding et al. 2008; Gaussier and Goutte 2005). In this section, we generalize the equivalence of these methods within the unified framework provided by Renyi's divergence. This generalization allows us to view NMF as a family of probabilistic mixture models indexed by the parameter γ in Renyi's divergence.

Consider the corpus of documents summarized as a $p \times n$ co-occurrence matrix V described earlier. Let the v_{ij}th entry of V denote the frequency of occurrence of term i in document j. In the context of NMF, v_{ij} has a Poisson distribution with mean μ_{ij} and the v_{ij}s are independent. The log-likelihood can be shown to be equivalent to $\sum_{ij} \left\{ -v_{ij} \log \left(\frac{v_{ij}}{\mu_{ij}} \right) - \mu_{ij} + v_{ij} \right\}$. In contrast PLSI is based on multinomial sampling where the term frequencies are normalized by conditioning on their sum such that $\sum_{ij} v_{ij}$ is fixed; for example, by re-scaling to their sum $v_{ij} \leftarrow \frac{v_{ij}}{\sum_{ij} v_{ij}}$ where $\sum_{ij} v_{ij} = 1$. The normalized term frequencies v_{ij} are neither independent nor Poisson distributed, and the log-likelihood can be shown to be equivalent to

$$\mathcal{L} = \sum_{j=1}^{n} \sum_{i=1}^{p} v_{ij} \log P_{ij} \tag{3.9}$$

(Hoffman 2001). We generalize the likelihood (3.9) in the following lemma.

Lemma *The log-likelihood for PLSI (3.9) is a member of the family of λ-log-likelihoods given by*

$$\mathcal{L}_\lambda = -\frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{n} \sum_{i=1}^{p} \left\{ v_{ij} \left[\left(\frac{v_{ij}}{P_{ij}} \right)^{\lambda} - 1 \right] - \lambda(v_{ij} - P_{ij}) \right\} \tag{3.10}$$

where $\lambda \neq 0$ *and* $\lambda \neq -1$.

The proof is provided in the Appendix. Using this lemma, we generalize the equivalence of NMF and PLSI in the following theorem. Again, the proof is relegated to the Appendix.

Theorem 2 *Renyi's divergence,* $D_\gamma^*(A||B)$*, between two non-negative matrices A and B in (3.2) is equivalent to the negative λ-log-likelihood* $-\mathcal{L}_\lambda(A, B)$ *given by (3.10), and therefore generalizes the relationship between NMF and PLSI.*

4 Quantitative evaluation of clustering

In this section, we describe the implementation of our NMF algorithm and quantitatively evaluate its performance in grouping n documents into homogeneous classes based on the frequency of occurrence of p terms. The NMF algorithm may not converge to the same solution on each run due to the random nature of initial conditions. We exploited this feature to evaluate the consistency of its performance and to quantify the clustering accuracy for a benchmark data set where the true number of classes k is known. The algorithm is applied multiple times with random initial starting values for W and H; and it groups the documents into k clusters, where k is the pre-specified rank of the factorization.

In order to assess whether a given γ provides a meaningful decomposition of the data for a fixed (known) number of classes $k = K$, we applied consensus clustering to evaluate the clustering accuracy of the factorization. Consensus clustering (CC) (Monti et al. 2003; Brunet et al. 2004) evaluates the performance of any unsupervised clustering algorithm based on resampling methods. In our case, the stochastic nature of initial conditions in the NMF algorithm is utilized in the evaluation process. In this approach, class membership for each document is determined based on the highest metaterm frequency profile. Each run of the algorithm results in an $n \times n$ connectivity matrix C with an entry of 1 if documents i and j cluster together and 0 otherwise, where $i, j = 1, ..., n$. The consensus matrix $\bar{C}$ is simply the average connectivity matrix obtained over N runs of the algorithm. Final document assignments are based on the re-ordered consensus matrix obtained by hierarchical clustering (HC) using average linkage. In our studies (Devarajan and Ebrahimi 2005; Devarajan and Wang 2007), we found the performance of the method to be consistent across multiple runs and, in general, 50–200 runs were sufficient to provide stability to the clustering. For a given data set and pre-specified rank K factorization, we employed CC as the primary method for selecting the appropriate γ by evaluating the clustering accuracy for each of several choices of γ based on the measures described in the next section. In practice, however, this data-driven approach can be used for selecting the appropriate γ for a given rank k or the appropriate combination of γ and rank k (for a range of ranks) in analyzing a real data set.

4.1 Measures for evaluating clustering accuracy

We utilized four measures for evaluating clustering accuracy by combining the information across $N = 200$ runs of the NMF algorithm. These are the *misclassification rate* (ν), *adjusted*

Rand index (ARI), *normalized mutual information* (NMI), and the *cophenetic correlation coefficient* (ρ). These measures have been used in similar clustering applications Shahnaz and Berry (2004), Ding et al. (2008) and Brunet et al. (2004).

The misclassification rate, ν, is the proportion of documents that are classified incorrectly by the CC algorithm across all clusters based on the final cluster labels assigned by that algorithm. ν can be calculated only if the true number of classes K is known and thus provides us with an overall measure of agreement for the clustering. An equivalent measure is given by clustering accuracy defined as $1 - \nu$. ARI and NMI are commonly used measures to quantify the agreement between the true class labels X and the assigned cluster labels Y. ARI is the proportion of pairs of documents that are both in the same class and same cluster or that are both in a different class and different cluster, adjusted for chance (Monti et al. 2003). NMI is an information theoretic measure based on estimated entropies (Strehl and Ghosh 2002). The cophenetic correlation coefficient, ρ, is defined as the correlation between $1 - \bar{C}$ and the distance induced by HC using average linkage (Brunet et al. 2004).

Unlike ν, NMI, ARI and ρ can be computed even if the true number of classes K is not known. However, for our purpose of evaluating clustering accuracy the true K is known. The range of each measure is [0, 1] where the two extreme values correspond to random partitioning and perfect clustering, respectively. This enabled us to compare these three measures by correlating each with ν across the range of γ based on the true K.

4.2 Data normalization

We consider two different normalization schemes for the term frequency matrix for evaluating our proposed methods. These are term frequency normalization (*tf*) and term frequency-inverse document frequency normalization (*tfidf*). In each case, the input matrix is idealized by adding a small positive constant to the zero entries in order to provide numerical stability to the algorithm. A detailed account of these normalization methods can be found in Salton and Buckley (1988) and Salton and McGill (1983). Henceforth, we shall refer to these methods simply as *tf* and *tfidf*, respectively. For each dataset presented, we compare these two schemes based on clustering accuracy for each (k, γ) combination as well as their overall performance.

4.3 Algorithm implementation

For any real large-scale data set, the implementation of steps in the factorization for a given rank k (or for a range of ranks) and its evaluation using CC is computationally intensive due to the consideration of k, N and/or γ. The stochastic nature of the NMF algorithm, however, enables each step in this procedure to be run independently and simultaneously, thus lending itself easily to a parallel implementation that would increase speed and efficiency. A comprehensive parallel implementation of this algorithm on a Message-Passing Interface/C++ platform (http://www-unix.mcs.anl.gov/mpi/mpich2/) using high-performance computing (HPC) clusters was utilized in data analyses presented here (Devarajan and Wang 2007; Wang et al. Manuscript in preparation, http://devarajan.fccc.edu).

5 Real-life and simulated examples

We describe several real-life and simulated examples to illustrate the applicability of our algorithm as well as its performance. For this purpose, we considered the following choices

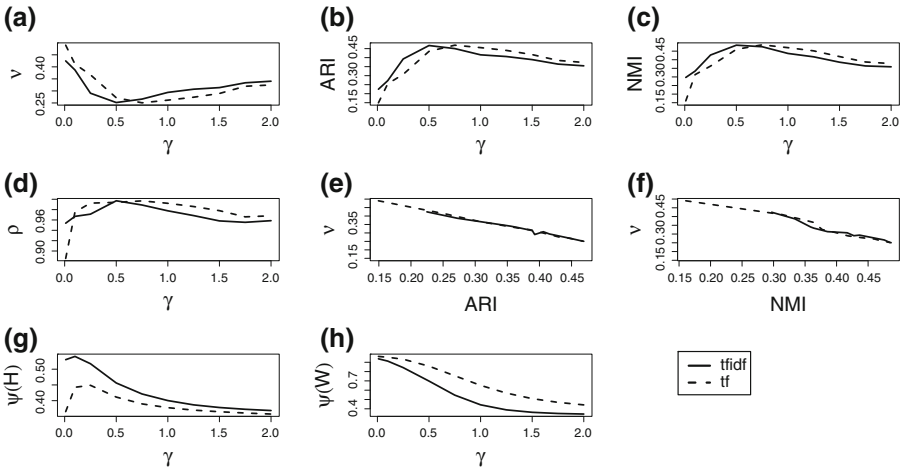

Fig. 1 Graphical illustration of the relationship between various measures of clustering accuracy, sparseness and the parameter γ for *tf* (*dashed*) and *tfidf* (*solid*) normalized WebKB data. **a**–**d** display, respectively, ν, ARI, NMI and ρ as a function of γ; **e**–**f** illustrate the relationship between ν and ARI and NMI, and **g**–**h** plot sparseness of the factored matrices W, H as a function of γ

of γ in the interval $(0, 2] : 0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2$, for various ranks k, each based on $N = 200$ runs. Note that $\gamma = 1$ represents KL divergence and $\gamma = 0.01$ approximates dual KL divergence described in (3.4). In addition, we considered well known algorithms based on ED and IS divergence.

5.1 WebKB data

The documents in the WebKB corpus are webpages collected from computer science departments of various universities by the World Wide Knowledge Base (WebKB) project and is available at http://web.ist.utl.pt/acardoso/datasets/. It consists of 2803 documents split into four classes, namely, project, course, faculty and student. We pre-processed this data set based on document and term frequencies alone. This resulted in 902 documents containing 1,338 terms across the four classes. We refer to this as the WebKB data set.

Given four major classes of documents in this corpus, we considered a rank $k = 4$ factorization for the choices of γ listed earlier. Figure 1a displays the misclassification rate ν plotted as a function of γ for normalized data based on *tf* and *tfidf*. These normalization methods are seen to perform similarly overall where $\gamma = 0.75$ and 0.5 result in the lowest misclassification rate of 25.06 and 25.17 %, respectively. This is corroborated by the relationship between ARI and γ as well as between NMI and γ, and is graphically presented in Fig. 1b, c. The strong negative correlation between each of these measures and ν is evidenced in Fig. 1e, f where, for each normalization method, the most homogeneous cluster corresponds to the value of γ that results in the smallest misclassification rate. The relationship between ρ and γ is similar to that observed for ARI and NMI (Fig. 1d). However, the correlation between ρ and ν is not as strong as that observed for the other measures (Table 2).

It is also worth noting from Fig. 1a that $\gamma = 1$ (KL divergence) results in higher misclassification rates of 26.16 and 29.38 % using *tf* and *tfidf*, respectively. Using simi-

lar measures of clustering accuracy, Ding et al. (2008) applied their NMF-based hybrid method to a different filtered version of this data set. They used only the top 1,000 terms in the corpus selected based on mutual information with class labels. On the other hand, our filtering scheme is completely blinded to the class labels and utilizes only the term and document frequencies. Nevertheless, our approach achieves a misclassification rate of only 25 % (clustering accuracy of 75 %) using both *tf* ($\gamma = 0.75$) and *tfidf* ($\gamma = 0.5$) normalization (Fig. 1a) and outperforms the 35.6 % misclassification rate (64.4 % clustering accuracy) achieved by their hybrid method. Tables 3 and 4 present a comparison of the performance of various algorithms based on the best performing model in terms of choice of γ (for Renyi divergence) and normalization method for real-life and simulated data sets used in this paper, respectively. The top row of Table 3 (WebKB) presents the results for this data set. There is clear evidence of the superior performance of the proposed algorithm based on Renyi divergence ($\nu = 25.06\,\%, \gamma = 0.75$) over algorithms based on ED ($\nu = 42.57\,\%$), IS divergence ($\nu = 31.15\,\%$) and KL divergence ($\nu = 26.16\,\%$).

Furthermore, we investigated the sparseness of the four metaterms and meteterm frequency profiles for each choice of γ using the sparseness measure defined in Hoyer (2004). For fixed γ, the mean sparseness of the metaterms (columns of W) was computed as the sparseness of each metaterm averaged across the four metaterms for each run and then averaged across the $N = 200$ runs. The mean sparseness of the metaterm frequency profiles (rows of H) was computed in a similar manner. For both normalization methods, sparseness of metaterms showed a monotonically decreasing trend with respect to γ (Fig. 1h) while sparseness of metaterm frequency profiles showed an initial surge for small γ before declining for higher values of γ (Fig. 1g). It is interesting to note that *tf* normalization resulted in uniformly sparser metaterms (across the range of γ) while *tfidf* normalization resulted in uniformly sparser metaterm frequency profiles.

Next, we repeated the above analysis on a larger version of the WebKB data set by preprocessing the original data using a less stringent filter, again based on document and term frequencies alone. By retaining documents containing a large number of very low frequency terms, this approach resulted in 2,606 out of the 2,803 documents in the original corpus, split into four classes, for the same number (1,338) of terms. We refer to this as the WebKB 2 data set. The purpose of this analysis was twofold: (1) to evaluate the robustness of the proposed methods relative to existing methods in terms of clustering accuracy and (2) to aid in the assessment of computational performance of the proposed methods for large-scale problems. Results are summarized in the second row of Table 3 (WebKB 2) for this data set as described before. Despite the threefold increase in the number of documents to be clustered based on the same number of terms, there is clear evidence of the superior performance of the proposed algorithm based on Renyi divergence ($\nu = 31.08\,\%, \gamma = 0.75$) over algorithms based on ED ($\nu = 45.01\,\%$), IS divergence ($\nu = 39.41\,\%$), KL divergence ($\nu = 33.11\,\%$) as well as the hybrid method of Ding et al. (2008) ($\nu = 35.6\,\%$). Given the increase in data set size, it is not entirely surprising that the misclassification rate attained by each method for this set has increased relative to that of the previous set. However, it is important to note that the overall performance of the various algorithms relative to one another was unchanged. The overall performance of the NMF algorithm based on Renyi divergence was similar between the two sets across values of γ, as evidenced by the correlations between various measures of clustering accuracy shown in Table 2. Furthermore, the same value of $\gamma = 0.75$ attained the best performance for both sets. These results establish the robust performance of the proposed methods in large-scale applications. The assessment of computational performance is devoted to §6.5.

Table 1 Subsets of Reuters data

Subset	K	# of documents	# of terms
1	3	77	1, 969
2	4	68	1, 105
3	5	120	1, 527
4	6	139	1, 639
5	8	169	1, 706
6	10	195	1, 800
7	20	276	1, 969
8	4	99	1, 407
9	3	55	828

5.2 Reuters data

The Reuters data is one of the most widely used benchmark datasets in text mining. We utilize the pre-processed data consisting of the frequencies of 1,969 terms from 276 different documents presented by Shahnaz and Berry (2004) and Shahnaz et al. (2006). These documents belong to a total of 20 different categories. For the purpose of illustrating our methodology, we created various subsets of this dataset where the known, true number K, of classes varied anywhere from 3 to 20. This allowed us to evaluate the performance of our method for various models, each determined by the true number of classes of documents. In each case, the appropriate rank of factorization was used.

Table 1 presents a summary of the various subsets used in our analysis. Subsets 1–6 were created based on the K most frequently occurring classes of documents in the corpus where the corresponding K is specified in this table. Subset 7 represents the complete data set. The other two subsets were created based on different combinations of more and less frequently occurring classes. The subset numbers in the first column of Table 1 are used to refer to these subsets in subsequent Tables 2, 3, 5 and 6.

For most subsets, *tf* was observed to perform at least as well as or better than *tfidf* in delineating the true classes. In all subsets, there was at least one value of γ that outperformed $\gamma = 1$ (KL divergence). Perfect clustering ($\nu = 0$) was achieved for subsets 1 and 8 for at least one normalization method and for at least one choice of γ (Table 3). A decreasing trend was observed in the metaterms and metaterm frequency profiles with respect to γ, similar to that seen for the WebKB data (data not shown). Strong negative correlations between ARI and ν, and between NMI and ν are clearly seen for the best performing normalization method (Table 2) across all subsets, with ARI showing a stronger correlation with ν relative to other measures.

The various subsets in this example allowed us to perform a sensitivity analysis with real data whereby we have assessed performance of our method for various true models. Shahnaz and Berry (2004) and Shahnaz et al. (2006) adopted a similar approach for evaluating their penalized NMF (PNMF) algorithm using this dataset. They considered ranks (subsets) ranging from $K = 2$ to 20 and assessed the clustering accuracy of their method for various choices of their penalty parameter λ. For more details, the interested reader is referred to their paper referenced above. It is not clear exactly how the subsets were chosen in their approach, nevertheless, it provides us with a basis for comparing the two methods for this dataset.

Table 2 Correlation between measures of clustering accuracy

Dataset	ARI versus ν	NMI versus ν	ρ versus ν
WebKB	−1.00	−0.98	−0.81
WebKB 2	−0.99	−0.97	−0.55
Reuters 1	−1.00	−0.98	−0.70
Reuters 2	−0.49	−0.55	−0.50
Reuters 3	−0.82	0.05	0.36
Reuters 4	−0.96	−0.41	−0.66
Reuters 5	−0.87	−0.96	−0.33
Reuters 6	−0.98	−0.97	−0.80
Reuters 7	−0.98	−0.96	−0.88
Reuters 8	−1.00	−0.99	0.33
Reuters 9	−0.70	−0.45	−0.16
Page Blocks[a]	−0.85	−0.19	0.03
Example 1	−0.97	−0.95	0.53
Example 2	−0.99	−1.00	−0.47

Results reported for best performing normalization method
[a] No normalization was required

Table 3 Comparison of methods based on misclassification rate for real-life data

Dataset	K	PNMF (λ)[a]	ED ($\lambda = 0$)	IS	KL ($\gamma = 1$)	Renyi ($\gamma \neq 1$)[a]
WebKB	4	–	42.57	31.15	26.16	25.06 (0.75)
WebKB 2	4	–	45.01	39.41	33.11	31.08 (0.75)
Reuters 1	3	–	2.60	37.66	1.30	0 (0.25, 0.5, 0.75)
Reuters 2	4	–	36.76	36.76	29.41	23.53 (1.25)
Reuters 3	5	–	35.00	41.67	17.50	17.50 (0.5)
Reuters 4	6	27.38 (0.001)	25.90	40.29	28.06	20.86 (1.75)
Reuters 5	8	42.75 (0.1)	35.50	49.70	36.69	33.14 (0.5)
Reuters 6	10	32.64 (0.01)	45.64	52.31	28.72	30.26 (1.25)
Reuters 7	20	42.86 (0.001)	55.07	56.16	42.75	41.67 (1.5)
Reuters 8	4	22.25 (0.001)	21.21	31.31	2.02	0 (0.5)
Reuters 9	3	–	30.91	30.91	27.27	23.64 (1.5)
Page Blocks	5	–	34.61	47.05	41.79	29.64 (0.25)

[a] Best-performing λ (PNMF) or γ (Renyi) shown within parentheses

Table 3 presents the misclassification rates achieved by different algorithms for various ranks. In each case, the best performing model determined by the choice(s) of γ or λ and normalization method is listed. It is evident from these results that our approach not only outperforms PNMF but also algorithms based on ED, IS and KL divergences throughout the range of K considered. This table also presents results for several additional subsets of the Reuters data for these algorithms where similar improvements in performance are seen. There is also a notable improvement in performance for smaller ranks where our algorithm achieved near-perfect or perfect clustering. While the misclassification rate tends to increase

with K for all algorithms, the gain in clustering accuracy is substantial for Renyi divergence especially in comparison to ED and IS divergence.

Furthermore, Ding et al. (2008) applied their NMF-based hybrid method to the subset of this dataset containing the ten most frequently occurring categories. Once again, they used an informative filter that utilized the top 1,000 terms based on mutual information with class labels. However, our approach performed significantly better by achieving misclassification rates of 28.72 and 30.76 % (or clustering accuracies of 71.28 and 69.24 %), respectively, using *tf* and *tfidf* in clustering the documents over their hybrid approach which achieved a much higher misclassification rate of 47.9 % (52.1 % clustering accuracy).

5.3 Page Blocks data

This dataset was described by Esposito et al. (1994) and represents a unique example in document analysis. Here, we are interested in classifying all the blocks of the page layout of a document that have been detected by a segmentation process. This is an important step in document analysis that is necessary for separating text from non-text areas. The original dataset consists of 5,473 blocks from 54 distinct documents. Each block represents an observation and there are five classes of blocks, namely, text, horizontal line, picture, vertical line and graphic. The following variables are measured for each block-number of black pixels per unit area, mean number of white-black transitions, total number of black pixels, number of white-black transitions in the original bitmap of the block, height, length, area and eccentricity (ratio of length to height). In addition, the dataset also contains the number of black pixels per unit area and the total number of black pixels obtained after the application of a smoothing algorithm. This data set is available http://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification. For more details on this dataset, the interested reader is referred to Esposito et al. (1994) and Malerba et al. (1995).

We reduced the dimensionality of this dataset by removing blocks with a relatively small number of black pixels per unit area *and* mean number of white-black transitions. This resulted in 1407 blocks across the five classes. Also since some variables have been normalized with respect to other variables in this dataset, no further normalization (i.e., *tf* or *tfidf*) was deemed appropriate for this data. A rank $k = 5$ factorization was applied to this dataset for each γ under consideration. $\gamma = 0.25$ resulted in the most homogeneous grouping of blocks based on the measured variables. In particular, it is worth noting from Table 3 that $\gamma = 0.25$ outperformed all four known algorithms that are embedded within Renyi divergence, namely, KL divergence ($\gamma = 1, \nu = 41.79\,\%$), approximation to dual KL divergence ($\gamma = 0.01, \nu = 44.28\,\%$), Bhattacharya distance ($\gamma = 0.5, \nu = 36.74\,\%$) and the Pearson Chi-squared statistic ($\gamma = 2, \nu = 45.98\,\%$)—by a wide margin (see Fig. 2). These results emphasize the need to incorporate different choices of γ in the factorization, beyond the commonly known metrics. Furthermore, $\gamma = 0.25$ also outperforms algorithms based on ED ($\nu = 34.61\,\%$) and IS divergence ($\nu = 47.05\,\%$).

5.4 Simulating nested classes

We further investigate the performance of our NMF algorithm via extensive simulations involving a correlated structure. In particular, we illustrate its ability to recover documents into the true underlying classes when there exists a sub-structure (or a dependent structure) between different classes. This is more realistic in real-life data especially when the number of classes exceeds two, and there is a hierarchical or nested structure of the classes. To this end, we construct two examples involving simulated frequencies of $p = 1{,}000$ terms for

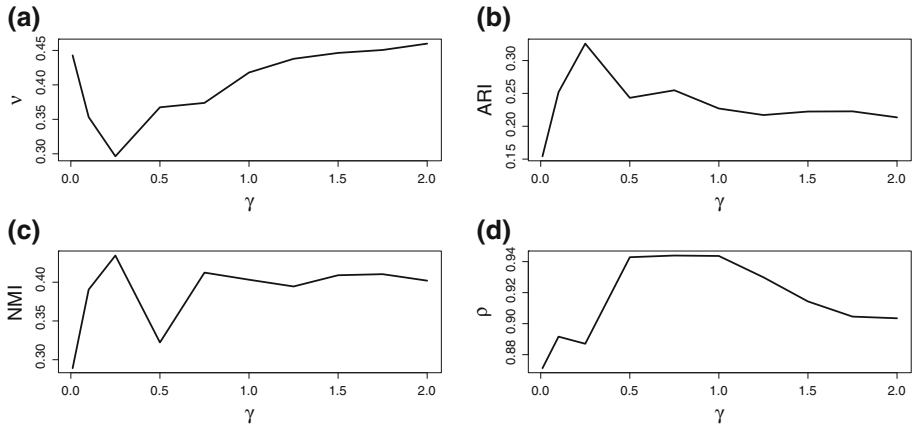

Fig. 2 Illustration of the relationship between various measures of clustering accuracy and γ for the Page Blocks data. **a–d** display, respectively, ν, ARI, NMI and ρ as a function of γ. The best performing model was obtained for $\gamma = 0.25$. This is indicated by the observed trend in ARI and NMI as γ increases (**b**, **c**) and by the smallest misclassification rate, ν, of 29.64 % (**a**). On the other hand, ρ exhibits an inconsistent change as γ increases, peaking at $\gamma = 1$ (corresponding to $\nu = 41.79\,\%$) (**d**)

each of $n = 60$ documents. We first describe their construction followed by their analyses based on our methods.

Example 1 We generated the term-document frequencies as follows: Let documents 1–20, 21–40 and 41–60 denote classes A, B and C respectively. For the first 50 terms, frequencies for documents in classes A, B and C were generated from a Poisson distribution with means 10, 1 and 1 respectively. For terms $51-100$, frequencies for documents in class B were generated as $Y \sim \min(X_1, X_2)$ where $X_1 \sim Poisson(mean = \lambda_1)$ and $X_2 \sim Poisson(mean = \lambda_2)$; and frequencies for documents in class C were generated from a Poisson distribution with mean λ_3. Documents in classes B and C have a dependent structure for terms 51–100 while for terms 1–50, documents in class A are independent of those in classes B and C. For the remainder of the terms, all documents are generated from a Poisson distribution with unit mean. We set $\lambda_1 = 20$ and considered various choices of λ_2 in the range (20, 40].

Example 2 For this example, we generated toy data based on the same setup as Example 1 above except for the following: For the first 50 terms, frequencies for documents in classes A, B and C were generated from a Poisson distribution with means 20, 1 and 1 respectively. In this set-up, documents in classes B and C have a dependent structure for terms 51–100 while documents in classes A and B have a dependent structure among the first 100 terms. In this structure, class B is dependent on both classes A and C. We set $\lambda_1 = 20$ and considered various choices of λ_2 in the range [25, 40].

The rationale behind this scheme is to generate data with a dependent and/or a hierarchical structure. In the first example, there are two major classes where one class has two subclasses while the second example represents a unique dependent structure between all three classes. Each dataset was normalized using *tf* and *tfidf* and then a rank $k = 3$ factorization was applied using our method. The overall performance of our algorithm on the simulated datasets paralleled that on the real-life data presented. The results were also insensitive to the choice of the Poisson mean parameters in each example. For the sake of brevity, we present results only for the case $\lambda_1 = 20$ and $\lambda_2 = 25$ in these examples.

Table 4 Comparison of methods based on misclassification rate for simulated data

Dataset	K	ED	IS	KL ($\gamma = 1$)	Renyi ($\gamma \neq 1$)[a]
Example 1	3	31.67	31.67	26.67	3.33 (1.5)
Example 1a	3	31.67	31.67	0	0[b]
Example 1b	3	21.67	28.33	8.33	0 (1.25,1.75)
Example 1c	3	33.33	31.67	28.33	16.67 (0.25)
Example 2	3	33.33	31.67	6.67	5.00 (1.25,1.75,2)

[a] Best-performing γ shown within parentheses
[b] All choices of γ except $\gamma = 0.01$

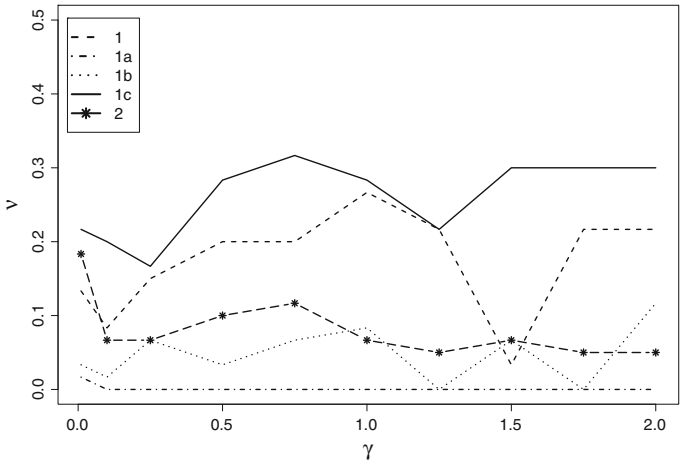

Fig. 3 Graphical illustration of the relationship between the misclassification rate ν and γ for simulated data in Examples 1 and 2 in §6.4. Examples are identified using different line types in the legend. There is at least one value of γ that outperforms KL divergence ($\gamma = 1$) in each example. In Example 1 (*dashed lines*) there is a significant improvement in clustering for $\gamma = 1.5$ ($\nu = 3.33\,\%$) compared to $\gamma = 1$ (26.67 %). Results are displayed for various values of the parameter λ_2 in this example. As λ_2 is decreased from 40 (Example 1a) to a value closer to 20 (Example 1b, $\lambda_2 = 30$; Example 1, $\lambda_2 = 25$; Example 1c, $\lambda_2 = 22$), i.e., as classes B and C became more and more similar, a gradual increase is observed in ν across γ. Note that as $\lambda_2 \to 20$, classes B and C merge into a single, larger class. In Example 2 (*dashed lines* connected by *asterisk*), $\gamma = 1.25, 1.75, 2$ achieve the lowest misclassification rate of 5 %

Table 4 presents a comparison of various algorithms using simulated data. In both examples, *tf* outperformed *tfidf* in delineating the true classes, and its performance was seen to be uniformly better than that of *tfidf* throughout the range of γ. Therefore, results are shown only for *tf* normalized data. The improvement afforded by our algorithm over those based on ED, IS and KL divergences is evidenced by the results in this table. Figure 3 presents the misclassification rates for each example plotted against γ. While Renyi divergence improves upon every other algorithm in both examples, it is particularly significant to note that there is at least one choice of γ that outperforms ED, IS and KL divergences by a wide margin. ARI, NMI and ρ displayed relationships with γ that were similar to those observed for real-life data. The sparseness of both metaterms and metaterm frequency profiles exhibited a decreasing trend with respect to γ in a similar fashion to that seen in real data.

Next, we investigated the sensitivity of our method in delineating similar clusters. In our simulation studies, the similarity between any two clusters can be simply determined by the

data generating mechanism. To this end, we utilized Example 1 and varied the Poisson mean parameter λ_2 that determines the degree of closeness between classes B and C. It is natural that our method performs very well for $\lambda_2 \geq 30$, however, the utility of our method lies in its ability to capture subtle differences between classes B and C (i.e., as λ_2 approaches the limiting value of 20). A summary of the performance of various algorithms for Examples 1, 1a, 1b and 1c is given in Table 4. As noted above, Examples 1a and 1b represent situations where classes B and C are more dissimilar (where $\lambda_2 = 40$ and 30, respectively). In both cases, perfect clustering is achieved by Renyi divergence for multiple choices of γ. For instance in Example 1a, the performance of Renyi divergence is uniformly excellent throughout the range of γ where perfect clustering is achieved for all values of γ ($\nu = 0$), including KL divergence, with the exception of $\gamma = 0.01$ ($\nu = 1.67\,\%$). This deviates considerably from the performance of ED and IS divergence where both algorithms perform poorly in both cases (Table 4). When $\lambda_2 = 25$ (Example 1), $\gamma = 1.5$ is the best-performing algorithm with $\nu = 3.33\,\%$ while ED and IS are the worst performing algorithms, each with $\nu = 31.67\,\%$. KL divergence performed slightly better with a $\nu = 26.67\,\%$. Finally, for the extreme case of $\lambda_2 = 22$ (Example 1c), $\gamma = 0.25$ is the best performer ($\nu = 16.67\,\%$) and results in significant improvements over ED, KL and IS divergence (Table 4). This phenomenon was also observed in other examples in our simulation studies as the parameter values were varied (data not shown) and it emphasizes the need for a broader approach.

5.5 Computational performance of algorithms

We made systematic comparisons of the computational performance of the various algorithms developed and tested in this paper. Data were analyzed using Intel Xeon processors running at 2.40 Ghz and 24 Gb of RAM using the parallel implementation of algorithms on HPC clusters described in §5.4. Data set size was determined by the product of the number of rows (terms) and number of columns (documents) contained in it. Two objective measures were used to quantify computational speed. The first measure is the number of updates required until convergence for each run, averaged across $N = 200$ runs of a particular algorithm for a pre-specified rank K factorization for a given data set. As noted earlier, each run is allowed a maximum of 2000 iterations for convergence. A run that fails to converge within this pre-specified limit stops after 2000 iterations and the resulting W and H matrices are utilized in further quantitative evaluation of clustering. Faster algorithms typically will have a lower mean number of updates until convergence for a fixed rank K factorization, dependent on data set size.

The second measure is the more traditional CPU time required to complete a single run, averaged across $N = 200$ runs of a fixed rank K factorization of a given data set. Again, faster algorithms will require fewer updates until convergence per run and the mean number of updates required will have a bearing on computational time contingent on data set size. In addition to CPU time, we also computed wall clock times required for completion of a single run, averaged across $N = 200$ runs of a particular algorithm for a fixed rank K factorization of a given data set. Wall clock time is relevant in this application due to parallel implementation of our NMF algorithms on HPC clusters. It gives a more realistic assessment of computational time required for a particular job depending on the data set size, rank K of the factorization, number of runs N and number of choices of γ in Renyi divergence. It is also dependent on the number of cores (processors) chosen or available for a particular job on the cluster. Since CPU time may be affected by the number of available nodes and the demands on the HPC cluster, wall clock time provides a measure of the expected duration for a particular job subject to constraints on the cluster and job size. In this regard, the mean number of

Table 5 Mean number of updates until convergence

Dataset	K	ED	IS	KL ($\gamma = 1$)	Renyi ($\gamma \neq 1$)[a]
WebKB	4	214.7	647.1	162.1	104.9 (0.75)
WebKB 2	4	213.1	312.4	364.3	211.2 (0.75)
Reuters 1	3	34.6	392.1	54.6	37.2 (0.5), 42.1 (0.75), 29.3 (0.25)
Reuters 2	4	45.5	292.9	63.0	78.5 (1.25)
Reuters 3	5	43.5	677.9	85.6	61.9 (0.5)
Reuters 4	6	49.5	780.1	101.2	144.8 (1.75)
Reuters 5	8	51.6	976.5	119.2	93.7 (0.5)
Reuters 6	10	56.3	1, 131.1	127.1	154.6 (1.25)
Reuters 7	20	75.4	1, 494.5	163.2	189.1 (1.5)
Reuters 8	4	41.2	542.8	71.8	50.7 (0.5)
Reuters 9	3	40.4	303.7	69.6	68.4 (1.5)
Page Blocks	5	2,000	645.3	1, 983.2	1,984 (0.25)
Example 1	3	146.7	510.5	27.3	37.2 (1.5)
Example 2	3	176.3	517.8	20.5	23.3 (1.25), 36.2 (1.75), 39.2 (2)

[a] Best-performing γ shown within parentheses

updates can be considered to be a more objective measure of speed that is independent of cluster occupancy. As outlined in §5.4, the parallel implementation allows us to pre-specify all the aforementioned parameters prior to initiation of a particular job. This implementation was described in our earlier work (Devarajan and Wang 2007) and a more detailed account of it is provided in Wang et al. (Manuscript in preparation).

Table 5 summarizes the performance of various algorithms in terms of the mean number of updates until convergence for each data set presented in this paper. For each combination of algorithm and data set, the mean number of updates listed in the table corresponds to the best performing model whose misclassification rate is listed in Table 3 (for real data sets) and Table 4 (for simulated data sets). The NMF algorithm based on ED was observed to converge faster than most other algorithms in general, and for higher rank factorizations in particular. However the overall performance of ED has been poor in delineating the clusters of documents as evidenced by the results shown in Tables 3 and 4. This poor performance was observed throughout the range of K and for all data sets, both real life and simulated, considered. Computational speed has not translated to better performance in this case. An interesting observation in Table 5 is that the mean number of updates for ED has virtually remained the same between the smaller and larger WebKB data sets while its performance has worsened, and remains the overall worst performer across all algorithms tested (Table 3). On the other hand, the best performing model using Renyi divergence required twice as many updates until convergence for the larger WebKB set while continuing to remain the best overall performer among all algorithms. Although the NMF algorithm based on Renyi divergence has been relatively slow in terms of mean number of updates, particularly for higher rank factorizations, there have been significant gains in clustering accuracy as seen by our experimental results on both real-life and simulated data. In some cases, Renyi divergence converges faster than any other algorithm while in other cases, KL divergence converges faster. The difference is striking in the simulated data sets where Renyi divergence is uniformly the fastest algorithm. It is also the best performing algorithm in terms of classification accuracy (Table 3). IS

Table 6 Computational speed: mean clock (and CPU)[b] time per run (seconds)

Dataset	K	ED	IS	KL ($\gamma = 1$)	Renyi ($\gamma \neq 1$)[a]	Renyi (all γ)
WebKB	4	5.89	23.10	3.98	5.58 (0.75)	76.71
		[36.41]	[161.11]	[22.34]	[40.20]	[539.6]
		[[10.39]]				
WebKB 2	4	24.49	35.98	27.06	35.55 (0.75)	593.56
		[123.36]	[238.98]	[147.48]	[272.25]	[4147.60]
		[[27.33]]				
Reuters 4	6	0.38	6.41	0.25	1.34 (1.75)	8.22
		[1.32]	[43.62]	[2.80]	[10.59]	[66.88]
		[[0.38]]				
Reuters 5	8	0.63	13.39	1.02	0.72 (0.5)	13.55
		[2.74]	[90.65]	[5.10]	[4.66]	[108.50]
		[[0.65]]				
Reuters 6	10	0.98	24.34	1.70	2.65 (1.25)	19.77
		[6.40]	[159.03]	[8.43]	[23.30]	[154.62]
		[[0.95]]				
Reuters 7	20	3.71	98.00	6.63	7.20 (1.5)	62.81
		[33.56]	[646.56]	[36.38]	[57.55]	[468.52]
		[[2.80]]				
Page Blocks	5	0.59	0.43	0.51	0.99 (0.25)	7.69
		[2.44]	[2.08]	[3.09]	[9.79]	[77.46]
		[[3.67]]				

[a] Best-performing γ shown within parentheses
[b] CPU time indicated within [.] for all algorithms and within [[.]] for HALS algorithm for ED

divergence is the slowest among all algorithms considered and it is also the worst performing algorithm overall, for both real and simulated data sets, often exhibiting poorer classification accuracy than ED in our studies.

Table 6 summarizes the performance of various algorithms in terms of mean clock and CPU times for selected, large data sets discussed in the paper. CPU times are listed within brackets and represent single core values. Various criteria such as the presence of a large number of clusters ($K = 6, 8, 10, 20$ for the Reuters data sets) or a large number of documents to be clustered ($n = 902, 2{,}606$ and $1{,}407$, respectively, for the WebKB, WebKB 2 and Page Blocks data sets) were used to select these data sets. The NMF algorithm using Renyi's divergence was run for ten choices of the parameter γ for the rank K factorization specified in Table 6 for each data set (last column). In addition, the hierarchical alternating least squares (HALS) algorithm (Cichocki et al. 2009; Gillis and Glineur 2012) was applied to these data sets and its computational performance was compared to our proposed methods. HALS has been demonstrated to considerably improve convergence speed; however, this approach is limited to the use of ED as the divergence measure. In terms of clustering accuracy, this algorithm was found to have similar performance to that of the regular multiplicative algorithm for ED (see also Zhou et al. 2012; Gillis and Glineur 2012). CPU times for the HALS algorithm are listed within double brackets for each data set along with those of our approach for ED (third column of Table 6). For each combination of algorithm and data set, the mean

computational time across $N = 200$ runs is listed. For NMF algorithms based on ED, IS, KL and Renyi divergence (columns 3–6), these correspond to the best performing model whose misclassification rate is listed in Table 3. For each algorithm, a strong correlation was observed between computational time (both CPU and wall clock times) and the mean number of updates after accounting for data set size, with Spearman's rank correlations ranging from 0.83–1.00. An average sevenfold improvement in computational time was observed overall due to the parallel implementation. In half the cases, Renyi divergence (including KL) is the fastest algorithm while ED is faster for the remaining cases. It is also worth noting that in terms of CPU time, the HALS algorithm is significantly faster than the regular multiplicative algorithm for ED for all data sets with the exception of the Page Blocks data. In terms of mean clock time due to parallel implementation, a considerable gain is noted for the WebKB and Page Blocks data sets.

From a practical point of view, slower computational speed is mitigated by our parallel implementation of these algorithms. With the advent of HPC clusters and their widespread utilization as part of computational infrastructure in recent years, such a parallel implementation should be readily accessible to researchers in a variety of scenarios. The implementation of the HALS algorithm and its refinements (Cichocki et al. 2009; Gillis and Glineur; Zhou et al. 2012) in this setting are expected to significantly increase computational speed, but, as alluded to earlier, it is limited to the use of ED as the divergence measure. One possible avenue for future research is to extend these algorithms for the generalized divergence measures described in this paper.

6 Summary and discussion

In summary, we have described a unified algorithm for NMF and PLSI based on Renyi's divergence stemming from the Poisson likelihood. We proved convergence of our algorithm using an auxiliary function analogous to that used for proving convergence of the EM algorithm. This approach provides a unique and generalized statistical framework for NMF and PLSI and includes well-known divergence measures as special cases. It is also related to some recently proposed divergence measures via transformations. Furthermore, we generalized the relationship between NMF and PLSI using a Box–Cox transformation in the multinomial likelihood for PLSI. This generalization embeds PLSI within the larger framework of the λ-log-likelihood and enables us to utilize some useful properties of PLSI within a broader class of models. Last but not least, we demonstrated the applicability of our methods using simulated as well as real-life document clustering data.

One of the objectives of this paper has been to demonstrate the need for a generalized metric for modeling high-dimensional data in the context of text mining and document clustering. The generalized metric presented here retains the distributional assumption on the data while providing modeling flexibility via the choice of the parameter γ. In that regard, one could arguably view our approach from the perspective of penalized likelihood where the choice of γ in Renyi's divergence determines the joint penalty on the metaterms and metaterm frequency profiles, or alternatively, on the reconstructed matrix WH. Furthermore, the application of consensus clustering to select γ is analogous to the use of cross-validation for choosing the penalty parameter in penalized likelihood methods. Our real-life examples and simulation studies suggest an underlying effect due to the distribution of the term frequencies (across documents) on the performance of the clustering algorithm. This is determined by the choice of γ. Perfect clustering is indeed achievable with the appropriate choice of γ for some datasets, as demonstrated in our examples. The approach emphasizes the need for a data-driven choice

of γ and, hence, of the divergence measure itself used in the decomposition. In practice, we recommend the use of several values of γ for evaluating the homogeneity of clustering for a given factorization rank k. Our studies indicate that values of γ in the $(0, 2]$ range work very well in practice for any real data set as evidenced by our real-life and simulated examples. We also found that no significant improvement was afforded by the choice of higher values of γ (data not shown). This range also contains several well-known divergence measures thus providing interpretability of the NMF objective function. Our parallel implementation has distinct advantages in terms of computational speed and allows one to simultaneously evaluate several factorization ranks for multiple choices of γ.

Several computational algorithms have been suggested in the literature recently for improving the speed and efficiency in NMF (for example, Zhou et al. 2012; Gillis and Glineur 2010, 2012; Cichocki et al. 2007, 2009; Phan and Cichocki 2011; Cichocki and Phan 2009; Lin 2007; Févotte and Idier 2011; Wang and Li 2010). The theoretical approach presented in this paper paves the way for potentially extending these algorithms by incorporating the generalized Renyi divergence for count data. It should be emphasized, however, that computational algorithm development is not the focus of our work and that such an extension could form the core of future work on this topic. In particular, extending the fast NMF algorithms proposed by Gillis and Glineur (2012) and Zhou et al. (2012) to non-Gaussian models would broaden the applicability of NMF in different areas.

An important observation from the analytical results presented is that the best performing model is not necessarily the sparsest, either in terms of the metaterms or the metaterm frequency profiles. The simulations highlight the ability of our approach to delineate classes based on subtle differences between them. The overall performance of *tf* normalization was found to be superior to that of *tfidf*. Our results demonstrated that both ARI and NMI were better measures of clustering accuracy than ρ. The problems associated with ρ have been well documented in the literature (Hastie et al. 2001; Holgersson 1978). Moreover, ρ typically has too narrow a range to be useful in many applications and, unlike ARI and NMI, can only be used with consensus clustering in conjunction with hierarchical clustering. In particular, for a real dataset with unknown true number of classes, we recommend the use of ARI or NMI on *tf* normalized data.

While Renyi's divergence is applicable for modeling count data, it has been shown to closely approximate data from skewed distributions in large-scale gene expression studies (Devarajan and Ebrahimi 2005, 2008; Devarajan 2006, 2008). Applications of NMF in the domain of computational biology are abundant in the literature. An extensive list of such applications is presented in Devarajan (2008). Thus the approach presented here provides the generalizability and flexibility in modeling such large-scale biological data as well, and further broadens the usefulness and applicability of our method.

Acknowledgments KD was supported in part by NIH Grant P30 CA 06927 and an appropriation from the Commonwealth of Pennsylvania. NE was partially supported by NSF Grant DMS 1208273. The work of GW was done while he was a member of the High-Performance Computing Facility at Fox Chase Cancer Center. The authors thank Prof. Michael Berry at the University of Tennessee, Knoxville for kindly providing the Reuters data set. The authors also wish to acknowledge the assistance of Joseph Anlage of the High-Performance Computing Facility at Fox Chase Cancer Center.

7 Appendix

Proof of Thoerem 1 As noted earlier, a proof of the monotonicity of the updates in Theorem 1 is given in Cichocki et al. (2008). However, the auxiliary function used to derive updates

for W and H and to prove their monotonicity contains elements of V in the denominator. In order for this formulation of the auxiliary function to be properly defined, each element of V is implicitly required to be positive. However, for Poisson distributed data, it is possible to have exact zeroes in V and accordingly the auxiliary function may not be defined. Moreover, the non-negativity constraint imposed on V is fundamental to NMF and is independent of the data generating mechanism. Here we provide a more general proof of the monotonicity of updates that satisfies the non-negativity assumption on V. It is based on splitting the domain $\Re\backslash\{0, 1\}$ of the parameter γ into three disjoint regions and considering each separately. The update rules obtained under all cases, however, are the same.

First, we derive the update for H and prove its monotonicity for $0 < \gamma < 1$. Then we show how similar arguments can be used to prove the result for $\gamma > 1$ and for $\gamma < 0$. We will make use of an auxiliary function similar to the one used in the EM algorithm (Dempster et al. 1977; Lee and Seung 2001). Note that for h real, $G(h, h')$ is an auxiliary function for $F(h)$ if $G(h, h') \geq F(h)$ and $G(h, h) = F(h)$ where G and F are scalar valued functions. Also, if G is an auxiliary function, then F is non-increasing under the update $h^{t+1} = \arg\min_h G(h, h^t)$. Using the second Eq. in (3.5), we define

$$F(H_{aj}) = \gamma \sum_i V_{ij} + (1-\gamma) \sum_{ia} W_{ia} H_{aj} - \sum_i V_{ij}^{\gamma} \left[\sum_a W_{ia} H_{aj} \right]^{1-\gamma},$$

where H_{aj} denotes the ajth entry of H. Then the auxiliary function for $F(H_{aj})$ is

$$G\left(H_{aj}, H_{aj}^t\right) = \gamma \sum_i V_{ij} + (1-\gamma) \sum_{ia} W_{ia} H_{aj} - \sum_{ia} V_{ij}^{\gamma} (W_{ia} H_{aj})^{1-\gamma} \left(\frac{W_{ia} H_{aj}^t}{\sum_b W_{ib} H_{bj}^t} \right)^{\gamma}.$$

It is straightforward to show that $G(H_{aj}, H_{aj}) = F(H_{aj})$. To show that $G(H_{aj}, H_{aj}^t) \geq F(H_{aj})$, we use the convexity of $-x^{1-\gamma}$ and the fact that for any convex function f, $f\left(\sum_{i=1}^n r_i x_i\right) \leq \sum_{i=1}^n r_i f(x_i)$ for rational nonnegative numbers $r_1, \ldots, r_n$ such that $\sum_{i=1}^n r_i = 1$. We then obtain

$$-\left(\sum_a W_{ia} H_{aj}\right)^{1-\gamma} \leq -\sum_a \gamma_a \left(\frac{W_{ia} H_{aj}}{\gamma_a} \right)^{1-\gamma} = -\sum_a (W_{ia} H_{aj})^{1-\gamma} \left(\frac{W_{ia} H_{aj}^t}{\sum_b W_{ib} H_{bj}^t} \right)^{\gamma},$$

where $\gamma_a = \dfrac{W_{ia} H_{aj}^t}{\sum_b W_{ib} H_{bj}^t}$. From this inequality it follows that $F(H_{aj}) \leq G(H_{aj}, H_{aj}^t)$. The minimizer of $F(H_{aj})$ is obtained by solving

$$\frac{dG\left(H_{aj}, H_{aj}^t\right)}{dH_{aj}}$$

$$= (1-\gamma)\left(\sum_i W_{ia} - \sum_i V_{ij}^{\gamma}\left(W_{ia}^{1-\gamma}\right)\frac{W_{ia}H_{aj}^t}{\sum_b W_{ib}H_{bj}^t}\left(\frac{W_{ia}H_{aj}^t}{\sum_b W_{ib}H_{bj}^t}\right)^{\gamma-1} H_{aj}^{-\gamma}\right) = 0.$$

The update rule for H thus takes the form given in (3.7). For $\gamma > 1$, using the first Eq. in (3.5) we define

$$F(H_{aj}) = -\gamma\sum_i V_{ij} - (1-\gamma)\sum_{ia} W_{ia}H_{aj} + \sum_i V_{ij}^{\gamma}\left[\sum_a W_{ia}H_{aj}\right]^{1-\gamma},$$

and the auxiliary function for $F(H_{aj})$ as

$$G\left(H_{aj}, H_{aj}^t\right) = -\gamma\sum_i V_{ij} - (1-\gamma)\sum_{ia} W_{ia}H_{aj} + \sum_{ia} V_{ij}^{\gamma}(W_{ia}H_{aj})^{1-\gamma}\left(\frac{W_{ia}H_{aj}^t}{\sum_b W_{ib}H_{bj}^t}\right)^{\gamma}.$$

It is easy to see that $G(H_{aj}, H_{aj}) = F(H_{aj})$. By using the convexity of $x^{1-\gamma}$ for $\gamma > 1$, we can show that $F(H_{aj}) \leq G(H_{aj}, H_{aj}^t)$ and proceed to obtain the update rule for H as described above. The update rule for this case is exactly as those specified for the case $0 < \gamma < 1$. Finally, the proof for the case $\gamma < 0$ is obtained by using (3.6) and defining

$$F(H_{aj}) = -\gamma\sum_i V_{ij} - (1-\gamma)\sum_{ia} W_{ia}H_{aj} + \sum_i V_{ij}^{\gamma}\left[\sum_a W_{ia}H_{aj}\right]^{1-\gamma},$$

and the auxiliary function for $F(H_{aj})$ to be

$$G(H_{aj}, H_{aj}^t) = -\gamma\sum_i V_{ij} - (1-\gamma)\sum_{ia} W_{ia}H_{aj} + \sum_{ia} V_{ij}^{\gamma}(W_{ia}H_{aj})^{1-\gamma}\left(\frac{W_{ia}H_{aj}^t}{\sum_b W_{ib}H_{bj}^t}\right)^{\gamma}.$$

Again, it is easy to verify that $G(H_{aj}, H_{aj}) = F(H_{aj})$. Using the convexity of x^{γ} for $\gamma < 0$, we can show that $F(H_{aj}) \leq G(H_{aj}, H_{aj}^t)$ and proceed to obtain the update rule for H as shown above. By using symmetry of the decomposition $V \sim WH$ and by reversing the arguments on W, one can easily obtain the update rule for W given in (3.8) in the same manner as H.

Proof of Lemma Without loss of generality, we re-write the log-likelihood (3.9) by adding a constant term such that $\mathcal{L} = \sum_{j=1}^{n}\sum_{i=1}^{p} -v_{ij}\log\left(\frac{v_{ij}}{P_{ij}}\right)$. Using the Box–Cox family of transformations (Box and Cox 1964), we can generalize it as

$$\mathcal{L}_{\lambda} = \sum_{j=1}^{n}\sum_{i=1}^{p} -\frac{v_{ij}}{\lambda}\left[\left(\frac{v_{ij}}{P_{ij}}\right)^{\lambda} - 1\right] \tag{7.1}$$

where $\lambda \neq 0$. In the limit $\lambda \to 0$, we obtain the log-likelihood given in (3.9). This is similar in principle to the α-log-likelihood approach outlined in Matsuyama (2003). Since $\sum_{ij} v_{ij} = \sum_{ij} P_{ij} = 1$, we have $\sum_{ij}(v_{ij} - P_{ij}) = 0$. Adding this term to $\mathcal{L}_{\lambda}$ in (7.1) and multiplying throughout by the constant $\frac{2}{\lambda+1}$ does not alter the meaning and interpretation

of $\mathcal{L}_\lambda$, and results in $\mathcal{L}_\lambda$ defined in Eq. (3.10). We refer to $\mathcal{L}_\lambda = \mathcal{L}_\lambda(V, P)$ in (3.10) as the λ-log-likelihood where $V = [v_{ij}]$, $P = [P_{ij}]$, $\lambda \neq -1$ and $\lambda \neq 0$.

Proof of Thoerem 2 In the context of NMF, the power-divergence family of statistics (2.5) can be re-written as

$$\phi_\lambda(A, B) = \frac{2}{\lambda(\lambda+1)} \sum_{i,j} A_{ij} \left[\left(\frac{A_{ij}}{B_{ij}} \right)^\lambda - 1 \right] - \lambda(A_{ij} - B_{ij}) \tag{7.2}$$

for $\lambda \neq -1$ and $\lambda \neq 0$ since $\sum_{ij}(A_{ij} - B_{ij}) = 0$. Note that $\phi_\lambda(A, B) = -\mathcal{L}_\lambda(A, B)$, and if we reparametrize (7.2) such that $A_{ij} = \frac{\gamma}{2}\tilde{A}_{ij}$, $B_{ij} = \frac{\gamma}{2}\tilde{B}_{ij}$ and $\lambda = \gamma - 1$, we obtain the quantity $D^*_\gamma(\tilde{A}||\tilde{B})$ defined in (3.2). Hence the negative λ-log-likelihood is equivalent to Renyi's divergence between the matrices A and B. This equivalence thus generalizes the relationship between NMF and PLSI.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Behnke, S. (2003). Discovering hierarchical speech features using convolutional non-negative matrix factorization. In textitProceedings of the international joint conference on neural networks (Vol. 4, pp. 2758–2763). International joint conference on neural network; July 20–24, Portland, Oregon.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, *52*(1), 155–173.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*(2), 211–252.

Brunet, J.-P., Tamayo, P., Golub, T., & Mesirov, J. (2004). Metagenes and molecular pattern discovery using nonnegative matrix factorization. *Proceedings of the National Academy of Sciences*, *101*, 4164–4169.

Buciu, I., & Pitas, I. (2004). Application of non-negative and local non negative matrix factorization to facial expression recognition. In: *Proceedings of the 17th international conference on pattern recognition. 17th international conference on pattern recognition*. August 23–26, 2004; Cambridge, UK.

Buntine, W. (2002). Variational extensions to EM and multinomial PCA. In *Proceedings of ECML'02*.

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

Cheung, V. C. K., & Tresch, M. C. (2005). Nonnegative matrix factorization algorithms modeling noise distributions within the exponential family. In *Proceedings of the 2005 IEEE engineering in medicine and biology 27th annual conference* (pp. 4990–4993).

Cho, Y.-C., Choi, S., & Bang, S.-Y. (2003). Non-negative component parts of sound for classification. In *Proceedings of the 3rd IEEE international symposium on signal processing and information technology. 3rd IEEE international symposium on signal processing and information technology* (pp. 633–636). December 14–17, 2003, Darmstadt, Germany.

Cichocki, A., & Amari, S. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, *12*, 1532–1568.

Cichocki, A., Cruces, S., & Amari, S. (2011). Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, *13*, 134–170.

Cichocki, A., Lee, H., Kim, Y.-D., & Choi, S. (2008). Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, *29*(9), 1433–1440.

Cichocki, A., & Phan, H. A. (2009). Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E92–A*(3), 708–721.

Cichocki, A., Zdunek, R., & Amari, S. (2006). *Csiszar's divergences for non-negative matrix factorization: Family of new algorithms, lecture notes in computer science, independent component analysis and blind signal separation*. Berlin: Springer.

Cichocki, A., Zdunek, R., & Amari, S. (2007). *Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization, lecture notes in computer science* (Vol. 4666, pp. 169-176). Berlin: Springer.

Cichocki, A., Zdunek, R., Phan, A.-H., & Amari, S. (2009). *Nonnegative matrix and tensor factorizations: Applications to Exploratory multi-way data analysis*. Hoboken: Wiley.

Cooper, M., & Foote, J. (2002). Summarizing video using nonnegative similarity matrix factorization. In *Proceedings of the IEEE workshop on multimedia signal processing. IEEE workshop on multimedia signal processing* (pp. 25–28). December 9–11, 2002. St. Thomas, U.S. Virgin Islands.

Cressie, N., Pardo, L., & Pardo, M. (2003). Size and power considerations for testing log-linear models using ϕ-divergence test statistics. *Statistica Sinica*, *13*, 555–570.

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

Devarajan, K. (2006). Nonnegative matrix factorization—A new paradigm for large-scale biological data analysis. In: *Proceedings of the joint statistical meetings*. Seattle, Washington.

Devarajan, K. (2008). Nonnegative matrix factorization—An analytical and interpretive tool in computational biology. *PLoS Computational Biology*, *4*(7), E1000029. doi:10.1371/journal.pcbi.1000029.

Devarajan, K. (2011a). *Matrix and tensor decompositions, Problem solving handbook in computational biology and bioinformatics, part 5*. Berlin: Springer.

Devarajan, K. (2011b). *Statistical methods for the analysis of next-generation sequencing data*. Joint Statistical Meetings Miami Beach, Florida.

Devarajan, K., & Cheung, V. C. K. (2012). *On the relationship between non-negative matrix factorization and generalized linear modeling*. Joint Statistical Meetings, San Diego, California.

Devarajan, K., & Ebrahimi, N. (2005). Molecular pattern discovery using nonnegative matrix factorization based on Renyi's information measure. In *Proceedings of the XII SCMA international conference*. December 2–4, 2005. Auburn, Alabama. http://Atlas-Conferences.Com/C/A/Q/T/98.Htm

Devarajan, K., & Ebrahimi, N. (2008). Class discovery via nonnegative matrix factorization. *American Journal of Management and Mathematical Sciences*, *28*(3&4), 457–467.

Devarajan, K., & Wang, G. (2007). Parallel implementation of non-negative matrix algorithms using high-performance computing cluster. In: *Proceedings of the 39th symposium on the interface: Computing science and statistics. Theme: Systems biology*. May 23–26, 2007. Temple University, Philadelphia, Pennsylvania.

Dhillon, I. S., & Sra, S. (2005). *Generalized nonnegative matrix approximations with Bregman divergences. Advances in neural information processing systems. Vol. 18*. Cambridge: MIT Press.

Ding, C., Li, T., & Peng, W. (2008). On the equivalence between nonnegative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, *52*, 3913–3927.

Ding, N., Qi, Y., Xiang, R., Molloy, I., & Li, N. (2010). Nonparametric Bayesian matrix factorization by power-EP. *Journal of Machine Learning Research, W&CP 9*, 169–176.

Ebrahimi, N., & Soofi, E. (2004). Information functions for Reliability. In R. Soyer, T. A. Mazzuchi, & N. D. Singpurwalla (Eds.), *Mathematical reliability, an expository perspective* (pp. 127–159). Boston: Kluwer Academic Publishers

Esposito, F., Malerba, D., & Semeraro, G. (1994). Multistrategy learning for document recognition. *Applied Artificial Intelligence*, *8*, 33–84.

Févotte, C., & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, *23*(9), 2421–2456.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, *21*, 607–611.

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

Gaussier, E., & Goutte, C. (2005). Relation between PLSA and NMF and implications. In *Procedings of SIGIR'05*.

Gillis, N., & Glineur, F. (2010). Using underapproximations for sparse nonnegative matrix factorization. *Pattern Recognition*, *43*(4), 1676–1687.

Gillis, N., & Glineur, F. (2012). Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Computation*, *24*(4), 1085–1105.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

Hoffman, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177–196.

Holgersson, M. (1978). The limited value of cophenetic correlation as a clustering criterion. *Pattern Recognition*, *10*(4), 287–295.

Hoyer, P.O. (2002). Nonnegative sparse coding. In: *Neural networks for signal processing. IEEE workshop on neural networks for signal processing* (Vol. XII, pp. 557–565). September 4–6, 2002. Martigny, Switzerland.

Hoyer, P. O. (2003). Modeling receptive fields with nonnegative sparse coding. *Neurocomputing*, *52–54*, 547–552.

Hoyer, P. O. (2004). Nonnegative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, *5*, 1457–1469.

Kompass, R. (2007). A generalized divergence measure for nonnegative matrix factorization. *Neural Computation*, *19*, 780–791.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Lee, D. D., & Seung, S. H. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, *401*, 788–791.

Lee, D. D., & Seung, S. H. (2001). Algorithms for nonnegative matrix factorization. *Advances In Neural Information Processing Systems*, *13*, 556–562.

Li, SZ., Hou, X., Zhang, H., & Cheng, Q. (2001). Learning spatially localized, partsbased representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 207–212). December 8–14 2001, Kauai, Hawaii.

Lin, C.-J. (2007). Projected gradient methods for non-negative matrix factorization. *Neural Computation*, *19*, 2756–2779.

Liu, W., Zheng, N., & Lu, X. (2003). Non-negative matrix factorization for visual coding. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing. IEEE international conference on acoustics, speech and signal processing* (Vol. 3, pp. 293–296). April 6–10, 2003, Taiwan, China.

Lu, J., Xu, B., & Yang, H. (2003). Matrix dimensionality reduction for mining Web logs. In *Proceedings of the IEEE/WIC international conference on web intelligence. IEEE/WIC international conference on web intelligence* (pp. 405–408). October 13, 2003, Nova Scotia, Canada.

Malerba, D., Esposito, F., & Semeraro, G. (1995). A further comparison of simplification methods for decision-tree induction. In D. Fisher & H. Lenz (Eds.), *Learning from data: Artificial intelligence and statistics V, Lecture notes in statistics*. Berlin: Springer.

Mao, Y., & Saul, L. K. (2004). Modeling distances in large-scale networks by matrix factorization. In *Proceedings of the ACM internet measurement conference. ACM internet measurement conference* (pp. 278–287). October 25–27, 2004, Sicily, Italy.

Matsuyama, Y. (2003). The α-EM algorithm: Surrogate likelihood maximization using α-logarithmic information measures. *IEEE Transactions on Information Theory*, *49*(3), 692–706.

Matusita, K. (1954). On estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics*, *5*, 59–65.

Monti, S., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*, 91–118.

Neyman, J. (1949). Contributions to the theory of the χ^2 test. In *Proceedings of the first Berkeley symposium on mathematical statistics and probability*. Berkeley, University of California Press.

Okun, O., & Priisalu, H. (2006). Fast nonnegative matrix factorization and its application for protein fold recognition. *EURASIP Journal ol Applied Signal Processing*, Article ID 71817.

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

Pauca, P., Shahnaz, F., Berry, M., & Plemmons, R. (2004). Text mining using nonnegative matrix factorizations. In *Proceedings of the fourth SIAM international conference on data mining. Fourth SIAM international conference on data mining*. April 22–24, 2004, Lake Buena Vista, Florida.

Phan, A.-H., & Cichocki, A. (2011). Extended HALS algorithm for nonnegative Tucker decomposition and its applications for multiway analysis and classification. *Neurocomputing*, *74*(11), 1956–1969.

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

Renyi, A. (1970). *Probability theory*. Amsterdam: North Holland.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill, ISBN 0070544840.

Shahnaz, F., & Berry, M. (2004). *Document clustering using nonnegative matrix factorization*. Technical report 2004–2007, Department of Mathematics, Wake Forest University, North Carolina.

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

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for ccombining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.

Tsuge, S., Shishibori, M., Kuroiwa, S., & Kita, K. (2001). Dimensionality reduction using non-negative matrix factorization for information retrieval. In *Proceedings of the IEEE international conference on systems, man, and cybernetics* (Vol. 2, pp. 960–965). October 7–10, 2001, Tucson, Arizona.

Wang, G., Anlage, J. P., & Devarajan, K. *hpcNMF: A high-performance software package for non-negative matrix factorization*. URL:http://devarajan.fccc.edu (manuscript in preparation).

Wang, G., Kossenkov, A. V., & Ochs, M. F. (2006). LS-NMF: A modified nonnegative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, *7*, 175.

Wang, F., & Li, P. (2010). Efficient non-negative matrix factorization with random projections, In *Proceedings of the 10th SIAM international conference on data mining* (pp. 281–292).

Xu, B., Lu, J., & Huang, G. (2003). A constrained non-negative matrix factorization in information retrieval. In Proceedings of the IEEE international conference on information reuse and integration. IEEE international conference on information reuse and integration (pp. 273–277). October 27–29, Las Vegas, Nevada.

Zhang, S., Li, Q., Liu, J., & Zhou, X. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, *27*(13), i401–i409.

Zhou, G., Cichocki, A., & Xie, S. (2012). Fast nonnegative matrix/tensor factorization based on low-rank approximation. *IEEE Transaction on Signal Processing*, *60*(6), 2928–2940.