



HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2017 August 31.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2015 November ; 2015: 1658–1664. doi:10.1109/BIBM.2015.7359924.

A Hybrid Algorithm for Non-negative Matrix Factorization Based on Symmetric Information Divergence

Karthik Devarajan,

Department of Biostatistics & Bioinformatics, Fox Chase Cancer Center, Temple University Health System, Philadelphia, PA 19111

Nader Ebrahimi, and

Division of Statistics, Northern Illinois University, DeKalb, IL 60115

Ehsan Soofi

Lubar School of Business, University of Wisconsin-Milwaukee, Address P.O.Box 742, Milwaukee, WI 53201

Abstract

The objective of this paper is to provide a hybrid algorithm for non-negative matrix factorization based on a symmetric version of Kullback-Leibler divergence, known as *intrinsic information*. The convergence of the proposed algorithm is shown for several members of the exponential family such as the Gaussian, Poisson, gamma and inverse Gaussian models. The speed of this algorithm is examined and its usefulness is illustrated through some applied problems.

Keywords

non-negative matrix factorization; intrinsic information; Kullback-Leibler divergence; dual; exponential family

I. Introduction

Let V be a $p \times n$ matrix of nonnegative elements. Consider the approximate factorization of form

$$V \approx WH, \quad (1)$$

where W is $p \times k$ and H is $k \times n$ are matrices of nonnegative elements.

In molecular pattern discovery, for example, V is the gene expression matrix where the rows contain n expression levels, typically less than a hundred, of p genes, typically in the tens of thousands. The goal is to find a small number of metagenes each defined as a nonnegative linear combination of the p genes. This is accomplished by (1), where columns of W define

$k \ll n$ metagenes and columns of H represent metagene expression patterns corresponding to

the n samples. In (1), $v_{ij} \approx b_{ij} = \sum_{\ell=1}^k w_{i\ell} h_{\ell j}$, where $w_{i\ell}$ is the coefficient of gene $i = 1, \dots, p$ for the metagene $\ell = 1, \dots, k$ and $h_{\ell j}$ is the expression level of the metagene ℓ in sample $j = 1, \dots, n$.

The b_{ij} are found iteratively such that the error of approximation is controlled by the

divergence measure $D(V \| WH) = \sum_{i=1}^n \sum_{j=1}^p D(v_{ij} \| b_{ij}) < \varepsilon$. Lee and Seung [13] developed non-negative matrix factorization (NMF) algorithms based on the Euclidean distance $L_2(V \| WH) = \|V - WH\|^2$ and the following divergence measure

$$D(V \| WH) = \sum_{ij} \left(w_{ij} \log \frac{w_{ij}}{b_{ij}} - w_{ij} + b_{ij} \right) \geq 0, \quad (2)$$

where the equality holds if and only if $w_{ij} = b_{ij}$ for all i, j . This is the well-known Kullback-Leibler (KL) information divergence between two Poisson distributions where

$\sum_{ij} w_{ij} = \sum_{ij} b_{ij} = 1$. Unlike $L_2(V \| WH)$ which is symmetric, $D(V \| WH) \neq D(WH \| V)$, so Lee and Seung [13] referred to $D(V \| WH)$ as the divergence of V from WH . Asymmetric measures such as $D(V \| WH)$ are usually called “directed divergence” measures [12]. Other divergence measures used in the literature in this context include Rényi divergence and its special cases. This measure includes Lee & Seung’s KL information divergence (2) as a special case and assumes that elements w_{ij} and b_{ij} of V and $B = WH$ are means of independent Poisson random variables [6,8].

It should be noted that the term KL information divergence is used in a broader context in this paper, one that is defined by (3) below for any probability distribution. This paper provides an algorithm for NMF based on a symmetric version of KL divergence, known as *intrinsic information*, and illustrates its usefulness in some applied problems.

The organization of this paper is as follows. Section II defines the intrinsic information measure and illustrates its application to Gaussian, Poisson, gamma and inverse Gaussian models. Section III develops a hybrid NMF algorithm for these models and provides proof of its convergence. Section IV evaluates its performance through some examples while section V gives some concluding remarks.

II. Symmetric Information Divergence

The KL information divergence between two distributions F and G with density (mass) functions f and g is

$$K(f \| g) \equiv \int \log \frac{f(x)}{g(x)} dF(x). \quad (3)$$

$K(f \| g) = 0$, where the equality holds if and only if $f(x) = g(x)$ almost everywhere [12]. *KL* information divergence, also referred to as relative entropy, cross-entropy, and directed divergence, is the fundamental information measure with many desirable properties for developing probability and statistical methodologies.

Two points pertaining $K(f \| g)$ may be less than desirable in some problems. First, (3) is defined only when F is absolutely continuous with respect to G , denoted as $F \leq G$. The absolute continuity requirement limits utilization of $K(f \| g)$ in some problems such as comparison of two binomial distributions when $n_1 \neq n_2$, evaluation of Poisson approximation to binomial, and continuous approximations of discrete distributions such as the normal approximations to binomial and Poisson distributions.

The second issue pertaining to $K(f \| g)$ is that, apart from some exceptional cases such as $F = N(\mu_1, \sigma^2)$ and $G = N(\mu_2, \sigma^2)$, $K(f \| g)$ is not symmetric in F and G ; the latter is referred to as the reference distribution. This lack of symmetry may be of no concern or even desirable in many situations where a natural or ideal reference is at hand; e.g., when G is uniform, a natural, or an ideal distribution for a problem. Historically, the lack of symmetry has been dealt with by using Jeffreys divergence

$$J(f \| g) = K(f \| g) + K(g \| f). \quad (4)$$

Here, $K(g \| f)$ is referred to as dual *KL* divergence. Jeffreys divergence is defined only when both distributions are absolutely continuous with respect to each other; i.e., $F \simeq G$, a more stringent requirement than needed for $K(f \| g)$.

Bernardo and Rueda [2] and Bernardo [1], in the context of Bayesian reference analysis, defined the *intrinsic information* between F and G as

$$\vartheta(f \| g) = \min \{K(f \| g), K(g \| f)\}. \quad (5)$$

Clearly, $\vartheta(f \| g)$ is symmetric in F and G and bypasses the absolute continuity requirement. If, for example, when $F \not\leq G$, we have $K(f \| g) = \infty$ and $\vartheta(f \| g) = K(g \| f)$.

Definition 1

Let $F(x|\theta)$ denote a family of distributions indexed by a parameter $\theta \in \mathfrak{R}$, and let $F_i = F(x|\theta_i)$, $i = 1, 2$. The family is said to be intrinsic information ordered by θ if $\vartheta(f_1 \| f_2) = K(f_1 \| f_2)$ whenever $\theta_1 \leq \theta_2$.

The following mathematical relation will be used in the examples that follow.

Lemma 1—For any $a > 0$, $a - \frac{1}{a} \geq (\leq) 2 \log a$ if and only if $a \geq (<) 1$.

Proof: For $a \geq 1$ use the following well-known inequality. For any $a > 0$, $a \geq 1$,

$$\frac{\log a}{a-1} \leq \frac{1}{\sqrt{a}}. \text{ For } a=1 \text{ take the limit.}$$

Example 1 (Gaussian model)—Let F_i , $i=1, 2$ be the Gaussian distribution with mean μ_i and variance σ^2 . Then,

$$K(f_1 \| f_2) = K(f_2 \| f_1) = \frac{1}{2\sigma^2} (\mu_1 - \mu_2)^2, \quad (6)$$

and, hence, $\vartheta(f_1 \| f_2) = K(f_1 \| f_2) = K(f_2 \| f_1)$.

Example 2 (Poisson model)—Let F_i , $i=1, 2$ be the Poisson distribution with mean λ_i . Then,

$$K(f_1 \| f_2) = \lambda_1 \log \frac{\lambda_1}{\lambda_2} - \lambda_1 + \lambda_2 = \lambda_1 (\phi - \log \phi - 1), \quad (7)$$

where $\phi = \frac{\lambda_2}{\lambda_1}$. Now, using Lemma 1, $\vartheta(f_1 \| f_2) = K(f_1 \| f_2)$ if and only if $\phi \geq 1$. Thus, Poisson distributions are intrinsic information ordered according to the mean.

Remark 1—Rényi information divergence between probability densities f and g is defined by

$$K_r(f \| g) = \frac{1}{r-1} \log \int f^r(x) [g(x)]^{1-r} dx, \quad (8)$$

for $r \geq 1$, $r > 0$, where $K_1(f \| g) = K(f \| g)$ by continuity. For $r=1/2$ Rényi divergence is symmetric, $K_{1/2}(f \| g) = K_{1/2}(g \| f)$. This is the well-known Bhattacharya distance [10] and has been used for NMF [8]. Let F_i , $i=1, 2$ be the Poisson distribution with mean λ_i . Then, using (8),

$$K_{1/2}(f_1 \| f_2) = K_{1/2}(f_2 \| f_1) = \left(\sqrt{\lambda_1} - \sqrt{\lambda_2} \right)^2. \quad (9)$$

Example 3 (Generalized gamma and related models)—Consider the generalized gamma family, $GG(\alpha, \tau, \lambda_i)$, with density

$$f(x | \alpha, \tau, \lambda_i) = \frac{\tau \lambda_i^{\alpha\tau}}{\Gamma(\alpha)} x^{\alpha\tau-1} e^{-(\lambda_i x)^\tau},$$

where $x > 0$ and $\alpha, \tau, \lambda_i > 0$. The GG family includes several well-known models as subfamilies (see [11]). For the $GG(\alpha, \tau, \lambda_i)$ family,

$$K(f_1 \| f_2) = \alpha (\phi - \log \phi - 1), \quad (10)$$

where $\phi = \frac{\lambda_2^\tau}{\lambda_1^\tau} = \frac{\mu_1^\tau}{\mu_2^\tau}$, and $\mu_i = E_i(X) = \frac{\Gamma(\alpha+1/\tau)}{\Gamma(\alpha)\lambda_i}$, $i = 1, 2$. This information divergence is of the form in (7), and thus $\mathcal{I}(f_1 \| f_2) = K(f_1 \| f_2)$ if and only if $\phi = 1$. Hence, the GG family is intrinsic information ordered according to the mean. When τ is known, GG is a member of the exponential family. The subfamilies of GG are gamma ($\tau = 1$), Weibull ($\alpha = 1$), exponential ($\alpha = \tau = 1$), and generalized normal ($\tau = 2$). The generalized normal is itself a flexible family and includes Half-normal ($\alpha = 1/2$), Rayleigh ($\alpha = 1$), Maxwell-Boltzmann ($\alpha = 3/2$), and Chi ($\alpha = k/2$; $k = 1, 2, \dots$). Thus, all these distributions are intrinsic information ordered according to the scale parameter and the mean.

Example 4 (Inverse gaussian model)—Consider the inverse gaussian distribution, $IG(\lambda, \mu)$, with density

$$f(x|\lambda, \mu) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\lambda(x-\mu)^2}{2x\mu^2}\right),$$

where $x > 0$ and $\lambda, \mu > 0$. For this distribution,

$$K(f_1 \| f_2) = \frac{\lambda(\phi - 1)^2}{2\mu_1}, \quad (11)$$

where $\phi = \frac{\mu_1}{\mu_2}$ and $E(X_i) = \mu_i$, $i = 1, 2$. It is evident from $K(f_1 \| f_2)$ that $\mathcal{I}(f_1 \| f_2) = K(f_1 \| f_2)$ if and only if $\phi = 1$. Thus, IG distributions are intrinsic information reverse ordered according to the mean.

III. Non-negative Matrix Factorization

Let v_{ij} and $b_{ij} = \sum_{\ell=1}^k w_{i\ell} h_{\ell j}$ be means of independent Poisson distributions $F_{ij,m}$, $m = 1, 2$. Lee and Sueng [13] have shown that the KL divergence,

$$K(V \| WH) \equiv \sum_{ij} \left\{ V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \right\}, \quad (12)$$

is non-increasing under the following multiplicative sequential updating

$$H_{aj}^{t+1} = H_{aj}^t \left[\frac{\sum_i W_{ia} (V_{ij} / \sum_a W_{ia} H_{aj}^t)}{\sum_i W_{ia}} \right], \quad (13)$$

$$W_{ia}^{t+1} = W_{ia}^t \left[\frac{\sum_j H_{aj} (V_{ij} / \sum_a W_{ia}^t H_{aj}^t)}{\sum_j H_{aj}} \right]. \quad (14)$$

The following theorem provides multiplicative sequential updates for dual KL divergence, $K(WH \| V)$, for the Poisson model given by

$$K(WH \| V) = \sum_{ij} \left\{ (WH)_{ij} \log \frac{(WH)_{ij}}{V_{ij}} - (WH)_{ij} + V_{ij} \right\}. \quad (15)$$

This divergence has been described in [5] and can be shown to be a special case of Rényi divergence [5,8].

Theorem 1

Let v_{ij} and $b_{ij} = \sum_{\ell=1}^k w_{i\ell} h_{\ell j}$ be means of independent Poisson distributions $F_{ij,m}$ $m = 1, 2$. $K(WH \| V)$ in (15) is non-increasing under the following multiplicative sequential updating

$$H_{aj}^{t+1} = H_{aj}^t \exp \left[\frac{\sum_i W_{ia} \log \left(\frac{V_{ij}}{\sum_a W_{ia} H_{aj}^t} \right)}{\sum_i W_{ia}} \right], \quad (16)$$

$$W_{ia}^{t+1} = W_{ia}^t \exp \left[\frac{\sum_j H_{aj} \log \left(\frac{V_{ij}}{\sum_a W_{ia}^t H_{aj}^t} \right)}{\sum_j H_{aj}} \right]. \quad (17)$$

Proof—A detailed proof is provided in [15].

Using equation (6), the Gaussian model can be seen to be the trivial case for which $K(V \| WH) = K(WH \| V) = \mathcal{D}(V \| WH) = L_2(V \| WH) = \sum_{ij} (V_{ij} - (WH)_{ij})^2$. Heuristic as

well as rigorous Majorization-Minimization (MM) algorithms for NMF for the exponential family have been proposed [4,5,9]. These algorithms are based on KL divergence, $K(V \| WH)$, and embed the Gaussian, Poisson, gamma and inverse Gaussian models as special cases. The MM algorithms outlined in §4.1 and §4.2 of [9] provide sequential updates for W and H for these models and establish conditions for monotonicity of $K(V \| WH)$. Lee & Seung's Poisson NMF algorithm stated in equations (13) and (14) is a special case of this family of algorithms. Similarly, rigorous Expectation-Maximization (EM) algorithms for the gamma and inverse Gaussian models based on dual KL divergence, $K(WH \| V)$, have been recently proposed [7]. Theorems 1 and 3 of [7] provide sequential updates for W and H for these models and establish conditions for monotonicity of $K(WH \| V)$. These existing results can be used in conjunction with the result in Theorem 1 above (equations (16) and (17)) to obtain the generalization stated in Theorem 2 below. Theorem 2 provides conditions for monotonicity of $\mathcal{A}(V \| WH) = \min\{K(V \| WH), K(WH \| V)\}$ using sequential updates for W and H based on $K(V \| WH)$ and $K(WH \| V)$ for the Gaussian, Poisson, gamma and inverse Gaussian models.

Theorem 2

Let $F_{ij,m}$, $m = 1, 2$ be independent distributions with means v_{ij} and $b_{ij} = \sum_{\ell=1}^k w_{i\ell} h_{\ell j}$ from one of the following member models: Gaussian, Poisson, gamma and inverse Gaussian. Then

$\vartheta(V \| WH) \equiv \sum_{i=1}^n \sum_{j=1}^p \vartheta(f_{i,j,1} \| f_{i,j,2})$ is non-increasing under any non-increasing sequential updates of $K(V \| WH)$ and $K(WH \| H)$ where these quantities represent, respectively, the KL and dual KL divergence for that member.

Proof—Sequential updates for W and H based, separately, on $K(V \| WH)$ and $K(WH \| V)$ are available for the Gaussian, Poisson, gamma and inverse Gaussian models as outlined above. In order to differentiate the factorizations obtained using each divergence measure, let W, H and \tilde{W}, \tilde{H} represent the factored matrices obtained by minimizing $K(V \| WH)$ and $K(WH \| V)$, respectively, for each specified model and given rank k . At iteration t , compute $K(V \| W^t H^t)$ and $K(\tilde{W}^t \tilde{H}^t \| V)$. Let

$$\vartheta_t = \min\{K(V \| W^t H^t), K(\tilde{W}^t \tilde{H}^t \| V)\}.$$

Suppose that $\vartheta_t = K(V \| W^t H^t)$ (similar arguments can be used for the case

$\vartheta_t = K(\tilde{W}^t \tilde{H}^t \| V)$). Using the update rules for $K(V \| WH)$ and $K(WH \| V)$ we obtain W^{t+1} , H^{t+1} and

$$\vartheta_{t+1} = \min\{K(V \| W^{t+1} H^{t+1}), K(\tilde{W}^{t+1} \tilde{H}^{t+1} \| V)\}.$$

It is known that $K(V \| W^{t+1} H^{t+1}) < K(V \| W^t H^t)$. Thus if $\vartheta_{t+1} = K(V \| W^{t+1} H^{t+1})$, then

$\vartheta_{t+1} = K(V \| W^t H^t) = \vartheta_t$. If $\vartheta_{t+1} = K(\tilde{W}^{t+1} \tilde{H}^{t+1} \| V)$, then

$$\vartheta_{t+1} < K(V \| W^{t+1} H^{t+1}) < K(V \| W^t H^t) = \vartheta_t.$$

This implies that ϑ_t decreases with increasing iteration number t and completes the proof.

Using the result in Theorem 2, we propose the following hybrid algorithm for NMF.

A. A Hybrid NMF Algorithm based on Symmetric Information Divergence

- Let $K_1^{(t)}(\cdot)$ and $K_2^{(t)}(\cdot)$ denote the reconstruction errors based on $K(V \| WH)$ and $K(WH \| V)$, respectively, at iteration t . Update rules for W, H are available separately based on $K(V \| WH)$ and $K(WH \| V)$. Consider the Poisson model as an example. These updates are given, respectively, by (13,14) and (16,17).
- *Iteration 0:* Initialize W, H . Denote the initial values by $W^{(0)}, H^{(0)}$. Compute $K_1^{(0)}(\cdot)$ and $K_2^{(0)}(\cdot)$ using $W^{(0)}, H^{(0)}$. Let $m_0 = \min(K_1^{(0)}(\cdot), K_2^{(0)}(\cdot))$.
- *Iteration 1:* If $K_1^{(1)}(\cdot) < K_2^{(1)}(\cdot)$, then update $W^{(0)}, H^{(0)}$ using $K_1^{(1)}(\cdot)$ updates; else update $W^{(0)}, H^{(0)}$ using $K_2^{(1)}(\cdot)$ updates. Denote these updates by $W^{(1)}, H^{(1)}$. Go to iteration 2. Let $\vartheta(1) = \min(K_1^{(1)}(\cdot), K_2^{(1)}(\cdot))$.
- *Iteration 2:* Compute $K_1^{(2)}(\cdot)$ and $K_2^{(2)}(\cdot)$ using $W^{(1)}, H^{(1)}$. If $K_1^{(2)}(\cdot) < K_2^{(2)}(\cdot)$, then update $W^{(1)}, H^{(1)}$ using $K_1^{(2)}(\cdot)$ updates; else update $W^{(1)}, H^{(1)}$ using $K_2^{(2)}(\cdot)$ updates. Denote these updates by $W^{(2)}, H^{(2)}$. Go to iteration 3. Let $\vartheta(2) = \min(K_1^{(2)}(\cdot), K_2^{(2)}(\cdot))$.
- *Iteration t :* Let $W^{(t)}, H^{(t)}$ denote the updates and $\vartheta(t) = \min(K_1^{(t)}(\cdot), K_2^{(t)}(\cdot))$.
- Repeat the above steps. If $|\vartheta(t) - \vartheta(t-1)| < \epsilon$ where $t = \text{maxiter}$, then stop. $W^{(t)}$ and $H^{(t)}$ are the final converged values of W, H . The final value of the reconstruction error is then $\vartheta(t)$ depending on whether $K_1^{(t)}(\cdot)$ or $K_2^{(t)}(\cdot)$ is smaller at the converged iteration t . *maxiter* is the maximum number of iterations per run.

For the Gaussian model, the hybrid algorithm is identical to that based on KL or dual KL divergence. Theorem 2 generalizes the applicability of the hybrid algorithm and extends its utility to several important members of the exponential family. This makes our proposed approach robust. Furthermore, the hybrid algorithm is generalizable to other NMF algorithms as long as separate update rules for W and H , based on KL and dual KL divergences, are available. The hybrid algorithm allows the simultaneous application of KL divergence and its dual by alternating between the two measures based on their smaller value attained at any given iteration using sequential updates for W and H . At the converged iterate, the algorithm attains the minimum achievable value for the objective function based

on these measures and in some cases this value is seen to be smaller than that attained by either algorithm separately (see examples).

IV. Applications

A. Simulated Data

We investigated the performance of our hybrid NMF algorithm within the context of document clustering using toy data similar to that outlined in [8]. The algorithm based on the Poisson model was used for this purpose. Specifically, we constructed examples involving simulated frequencies of $p = 1000$ terms for each of $n = 60$ documents in order to illustrate the ability of the algorithm in reconstructing the data when the number of classes exceeds two, and there is a hierarchical or nested structure of the classes. Term-document frequencies were generated 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 mean (λ) 20, 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 $X_1 \sim \text{Pois}(\lambda = 70)$ and $X_2 \sim \text{Pois}(\lambda = 50)$; and frequencies for documents in class C were generated as $Z \sim \max(X_3, X_2)$ where $X_3 \sim \text{Pois}(\lambda = 30)$. For the remainder of the terms, all documents are generated from $\text{Pois}(\lambda = 1)$. In this set up, there are two major classes where one class has two sub-classes.

B. Reuters Data

The Reuters data is a widely used benchmark data set in text mining consisting of the frequencies of nearly 2000 terms in multiple categories from a document corpus [8,14]. For the purpose of illustrating our method, we selected a subset consisting of 1163 terms from 82 documents in five categories normalized using term frequencies.

C. Saccharomyces Genome Database (SGD)

Chagoyen et al. [3] utilized a corpus of 7080 articles relevant to a large set of genes and proteins from the SGD and created a data set consisting of 2365 terms across 575 yeast genes. NMF was applied to create literature profiles using common semantic features extracted from the corpus. Genes are then represented as additive linear combinations of the semantic features which can be further used for studying their functional associations.

D. Illustration of Examples

In all examples, factorizations of ranks $k = 2-6$ were considered. The hybrid algorithm was compared with the following algorithms - Lee & Seung [13] which is based on KL divergence (KL), dual KL divergence in Theorem 1 (dKL) and the symmetric special case of Renyi divergence ($symRenyi$) (see Remarks 1 and 2) [8]. Comparisons were made in terms of minimum reconstruction error (RE) at the converged iteration and the corresponding number of iterations required for convergence (N) across 20 runs of each algorithm. RE was calculated based on the particular algorithm employed. A maximum of 3000 iterations per run was used. These results are displayed in Figures 1–3.

Unlike other algorithms, the hybrid algorithm requires alternating between updates based on the KL or dKL algorithm as outlined in §III A. Starting with random initial values for W and H , sequential updates are based on the smaller of the RE based on KL and dKL at each iterate. This hybrid approach efficiently utilizes both algorithms simultaneously in this fashion and does not require the separate use of each algorithm. It provides better reconstruction of the original data while utilizing fewer iterations to convergence relative to the use of both KL and dKL , a phenomenon observed across all ranks considered (see Figures 1–3). We noticed that the hybrid algorithm is able to settle on a solution relatively quickly after alternating between the two asymmetric information divergence algorithms. In the simulated example it settled on KL (Figure 1(a)) while for the Reuters and SGD data sets it settled on dKL (Figures 2(a) and 3(a), respectively). The feasibility of the hybrid algorithm is empirically illustrated in Figure 4 for ranks $k = 2 - 5$. In our investigations, it was observed that this algorithm exhibited an average 2.10-fold (ranging from 1.45 to 3.12) reduction in the number of iterations to convergence relative to KL and dKL . In addition, it is also seen to be superior to $symRenyi$ in terms of RE across all ranks.

V. Conclusions and Future Work

In summary, a hybrid NMF algorithm that is applicable to several members of the exponential family of models has been presented. The algorithm relies on intrinsic information, a symmetric version of KL information divergence and its dual. A rigorous proof of monotonicity of updates for the algorithm is provided. Numerical experiments using simulated and real data demonstrate faster convergence of the algorithm relative to the asymmetric versions as well as better reconstruction of the original data. The proposed approach is particularly useful in applications where there is *a priori* knowledge or empirical evidence of signal-dependence in noise.

The basic principle underlying this fundamental algorithm is broadly extensible to frameworks involving penalty, kernel and discriminant functions. We are currently working on generalizing this algorithm to include all members of the exponential family, such as those that lie in the continuum between well-known models. A more detailed comparison of the performance of the hybrid approach in relation to existing methods for a given application, such as unsupervised clustering, will form the core of future work on this topic.

Acknowledgments

Work of KD was supported in part by NIH grant P30 CA 06927. NE was supported in part by NSF grant DMS 1208273. KD would like to thank Prof. Michael Berry at the University of Tennessee, Knoxville for providing the Reuters data.

References

1. Bernardo JM. Reference Analysis. Handbook of Statistics. 2004;25.
2. Bernardo JM, Rueda R. Bayesian hypothesis testing: a reference approach. International Statistics Review. 2002; 70:351–372.
3. Chagoyen M, Carmona-Saez P, Shatkay H, Carazo JM, Pascual-Montano A. Discovering semantic features in the literature: a foundation for building functional associations. BMC Bioinformatics. 2006; 7:41. [PubMed: 16438716]

4. Cheung VCK, Tresch MC. Nonnegative matrix factorization algorithms modeling noise distributions within the exponential family. *Proceedings of the 2005 IEEE Engineering in Medicine & Biology 27th Annual Conference*. 2005:4990–4993.
5. Cichocki, A., Zdunek, R., Amari, S. *Lecture Notes in Computer Science, Independent Component Analysis & Blind Signal Separation*. Springer; 2006. Csiszar's Divergences for Non-negative Matrix Factorization: Family of New Algorithms.
6. Cichocki A, Lee H, Kim YD, Choi S. Nonnegative matrix factorization with α -divergence. *Pattern Recognition Letters*. 2008; 29(9):1433–1440.
7. Devarajan K, Cheung VCK. On nonnegative matrix factorization algorithms for signal-dependent noise with application to electromyography data. *Neural Computation*. 2014; 26(6):1128–1168. [PubMed: 24684448]
8. Devarajan K, Wang G, Ebrahimi N. A unified statistical approach to nonnegative matrix factorization & probabilistic latent semantic indexing. *Machine Learning*. 2015; 99(1):137–163. [PubMed: 25821345]
9. Févotte C, Idier J. Algorithms for nonnegative matrix factorization with the β -divergence. *Neural Computation*. 2011; 23(9):2421–2456.
10. Freeman MF, Tukey JW. Transformations related to the angular & the square root. *Annals of Mathematical Statistics*. 1950; 21:607–611.
11. Johnson, NL., Kotz, S., Balakrishnan, N. *Continuous univariate distributions: Volume 1, Second Edition*. Wiley; New York: 1994.
12. Kullback, S. *Information Theory and Statistics*. New York: Wiley; 1959.
13. Lee, DD., Seung, HS. *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press; 2001. Algorithms for non-negative matrix factorization; p. 556-562.
14. Shahnaz F, Berry M, Pauca VP, Plemmons RJ. Document clustering using nonnegative matrix factorization. *Information Processing & Management: An International Journal*. 2006; 42(2):373–386.
15. Soofi E, Devarajan K, Ebrahimi N. A generalized hybrid algorithm for non-negative matrix factorization for the exponential family. *Manuscript in preparation*.

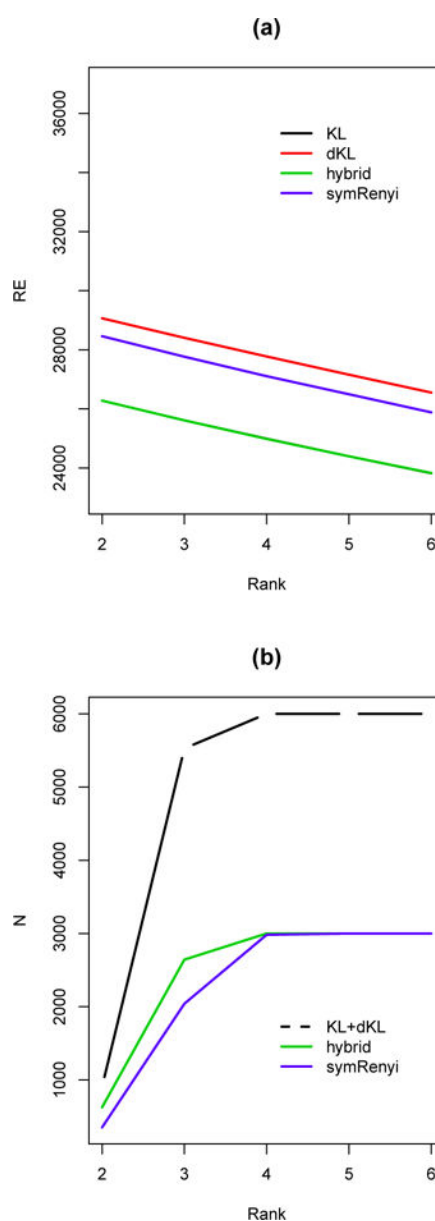


Figure 1. Comparison of algorithms: simulated data (figure legends apply to respective panels of Figures 1–3)

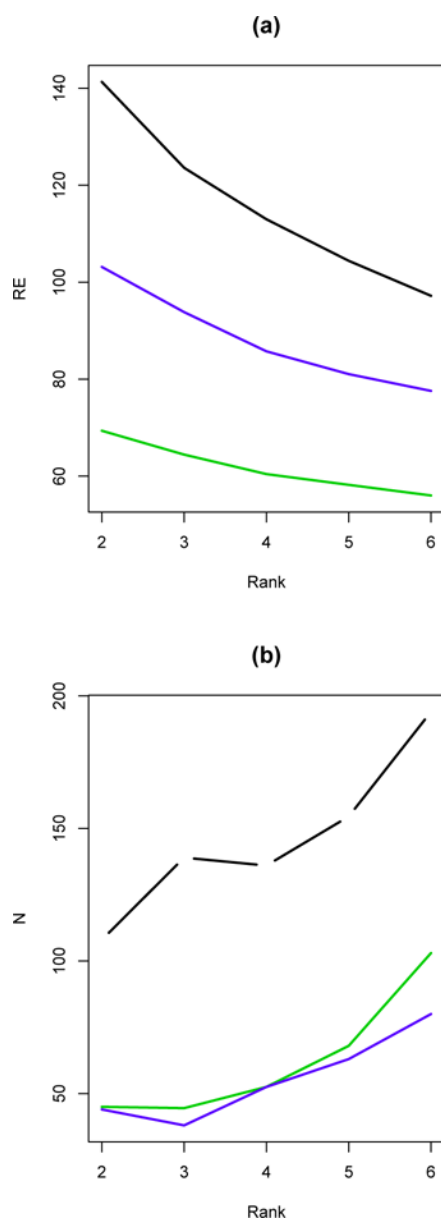


Figure 2.
Comparison of algorithms: Reuters data

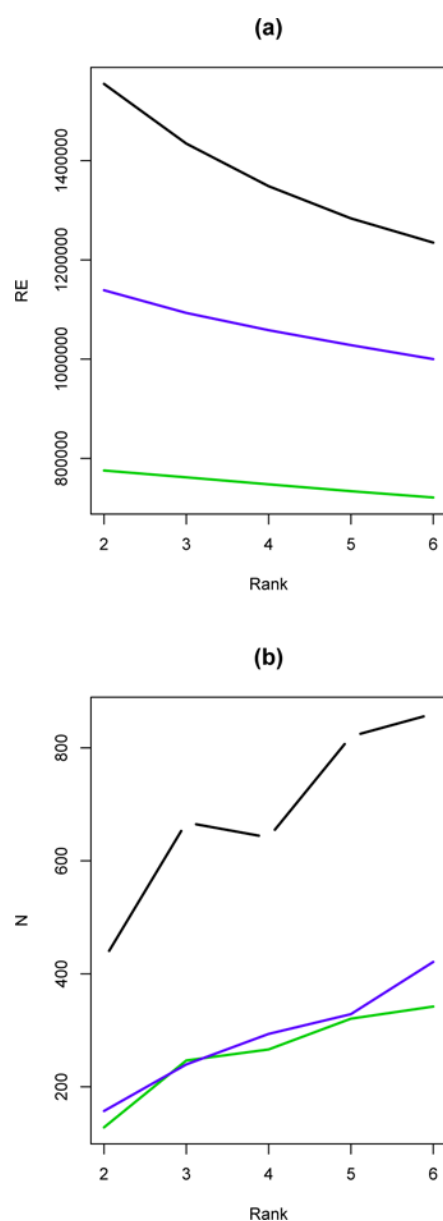


Figure 3.
Comparison of algorithms: SGD data

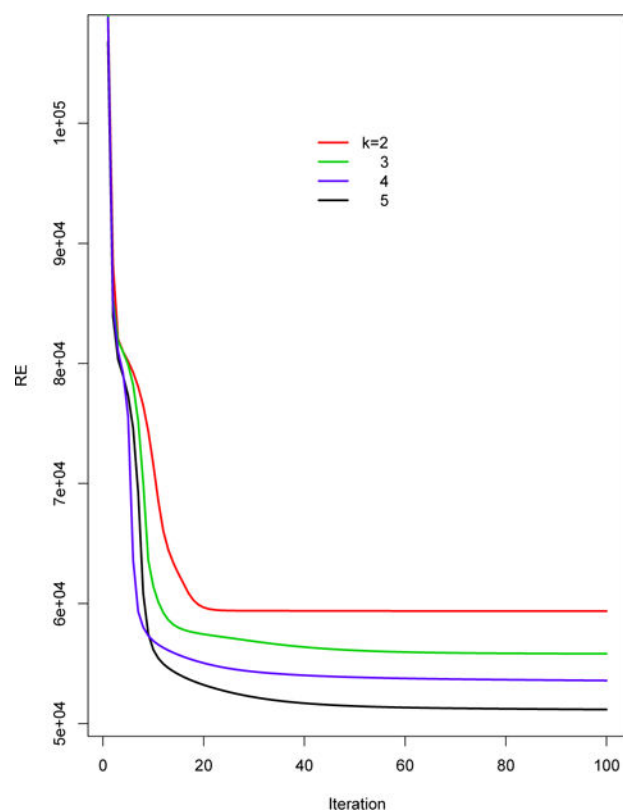


Figure 4.
Convergence of the hybrid algorithm