Semi-supervised Nonnegative Matrix Factorization for Document Classification*

Jamie Haddock, *Dept. of Mathematics, Harvey Mudd College*
Lara Kassab, *Dept. of Mathematics, Colorado State Univ.*
Sixian Li, *Dept. of Mathematics, Univ. of Illinois Urbana-Champaign*
Alona Kryshchenko, *Dept. of Mathematics, Cal. State Univ. Channel Islands*
Rachel Grotheer, *Dept. of Mathematics, Wofford College*
Elena Sizikova, *Center for Data Science, New York Univ.*
Chuntian Wang, *Dept. of Mathematics, Univ. of Alabama*
Thomas Merkh, *Dept. of Mathematics, Univ. of Cal., Los Angeles*
R. W. M. A. Madushani, *Dept. of Infectious Diseases, Boston Medical Cent.*
Miju Ahn, *Dept. of EMIS, Southern Methodist Univ.*
Deanna Needell, *Dept. of Mathematics, Univ. of Cal., Los Angeles*
Kathryn Leonard, *Dept. of Computer Science, Occidental College*

***Abstract*—We propose new semi-supervised nonnegative matrix factorization (SSNMF) models for document classification and provide motivation for these models as maximum likelihood estimators. The proposed SSNMF models simultaneously provide both a topic model and a model for classification, thereby offering highly interpretable classification results. We derive training methods using multiplicative updates for each new model, and demonstrate the application of these models to single-label and multi-label document classification, although the models are flexible to other supervised learning tasks such as regression. We illustrate the promise of these models and training methods on document classification datasets (e.g., 20 Newsgroups, Reuters).**

***Index Terms*—semi-supervised nonnegative matrix factorization, maximum likelihood estimation, multiplicative updates**

I. Introduction

Frequently, one is faced with the problem of performing a classification task on high-dimensional data which contains redundant information. One such task is *document classification* in which one assigns a set of categorical labels to documents based upon their contents [2], [4]. Document data is often represented using a *bag-of-words* model, where the dimensionality of the representation of each document is linear in the number of unique words used in the document corpus and thus can be extremely large [16]. A common approach is to first apply a dimensionality-reduction technique (e.g., PCA [27]), and then train a model for the classification task on the new, learned representation of the data. One problematic aspect of this two-step approach is that the learned representation of the data may provide "good" fit, but could suppress data features which are integral to classification [13]. For this reason, supervision-aware dimensionality-reduction models have become increasingly important in data analysis. Such models aim to use supervision in the process of learning the lower-dimensional representation, or even learn this representation alongside the classification model [3], [28], [34].

In this work, we propose new semi-supervised nonnegative matrix factorization (SSNMF) formulations which provide a dimensionality-reducing topic model and a model for a supervised learning task. Our contributions are:

- we motivate these proposed SSNMF models and that of [23] as maximum likelihood estimators (MLE) given specific models of uncertainty in the observations;
- we derive multiplicative updates for the proposed models that allow for missing data and partial supervision; and
- we perform experiments on real data which illustrate the promise of these models in both topic modeling and supervised learning tasks relative to the performance of other relevant classifiers (e.g. Multinomial Naive Bayes).

A. Notation

Our models make use of two matrix similarity measures. The first is the standard Frobenius norm, $\|\mathbf{A}-\mathbf{B}\|_F$. The second is the *information divergence* or I-divergence, a measure defined between nonnegative matrices $\mathbf{A}$ and $\mathbf{B}$,

$$D(\mathbf{A}\|\mathbf{B}) = \sum_{i,j}\left(\mathbf{A}_{ij}\log\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}} - \mathbf{A}_{ij} + \mathbf{B}_{ij}\right), \qquad (1)$$

where $D(\mathbf{A}\|\mathbf{B}) \geq 0$ with equality if and only if $\mathbf{A} = \mathbf{B}$ [22].

In the following, $\mathbf{A}/\mathbf{B}$ indicates element-wise division, $\mathbf{A}\odot\mathbf{B}$ indicates element-wise multiplication, and $\mathbf{AB}$ denotes standard matrix multiplication. We denote the set of non-zero indices of a matrix by $\text{supp}(\mathbf{A}) := \{(i,j) : \mathbf{A}_{ij} \neq 0\}$. When an $n_1 \times n_2$ matrix is to be restricted to have only nonnegative entries, we write $\mathbf{A} \geq 0$ and $\mathbf{A} \in \mathbb{R}^{n_1\times n_2}_{\geq 0}$. We let $\mathbf{1_k}$ denote the length-k vector consisting of ones, $\mathbf{1_k} = \begin{bmatrix}1,\cdots 1\end{bmatrix}^\top \in \mathbb{R}^k$, and similarly $\mathbf{0_k}$ denotes the vector of all zeros, $\mathbf{0_k} = \begin{bmatrix}0,\cdots 0\end{bmatrix}^\top \in \mathbb{R}^k$.

JH and DN were partially supported by NSF DMS #2011140 and NSF BIGDATA #1740325. JH is also partially supported by NSF DMS #2111440. ES was supported by the Moore-Sloan Foundation. Funding from ICERM and the NSF-AWM ADVANCE grant initiated the collaboration.

We let $\mathcal{N}\left(z|\mu,\sigma^2\right)$ denote the Gaussian density function for a random variable z with mean μ and variance σ^2, and $\mathcal{PO}\left(z|\nu\right)$ denotes the Poisson density function for a random variable z with nonnegative intensity parameter ν.

B. Preliminaries

In this section, we give a brief overview of the NMF and SSNMF methods.

Nonnegative Matrix Factorization: Given a nonnegative matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$ and a target dimension $r \in \mathbb{N}$, NMF decomposes $\mathbf{X}$ into a product of two low-dimensional nonnegative matrices. The model seeks $\mathbf{A}$ and $\mathbf{S}$ so that $\mathbf{X} \approx \mathbf{AS}$, where $\mathbf{A} \in \mathbb{R}_{\geq 0}^{n_1 \times r}$ is called the dictionary matrix and $\mathbf{S} \in \mathbb{R}_{\geq 0}^{r \times n_2}$ is called the representation matrix. Several formulations for this nonnegative approximation, $\mathbf{X} \approx \mathbf{AS}$, have been studied [8], [21], [22], [36]; e.g.,

$$\underset{\mathbf{A}\geq 0,\mathbf{S}\geq 0}{\operatorname{argmin}} \|\mathbf{X}-\mathbf{AS}\|_F^2 \text{ and } \underset{\mathbf{A}\geq 0,\mathbf{S}\geq 0}{\operatorname{argmin}} D(\mathbf{X}\|\mathbf{AS}), \quad (2)$$

where $D(\cdot\|\cdot)$ is the information divergence defined in (1). In what follows, we refer to the left formulation of (2) as $\|\cdot\|_F$-NMF and the right formulation of (2) as $D(\cdot\|\cdot)$-NMF. We refer the reader to [8] for discussions of similarity measures and generalized divergences (where information divergence is a particular case), and [25], [31] for generalized nonnegative matrix approximations with Bregman divergences.

Semi-supervised NMF: SSNMF is a modification of NMF to jointly incorporate a data matrix and a (partial) class label matrix. Given a data matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$ and a class label matrix $\mathbf{Y} \in \mathbb{R}_{\geq 0}^{k \times n_2}$, $(\|\cdot\|_F, \|\cdot\|_F)$-SSNMF is defined by

$$\underset{\mathbf{A},\mathbf{S},\mathbf{B}\geq 0}{\operatorname{argmin}} \underbrace{\|\mathbf{W}\odot(\mathbf{X}-\mathbf{AS})\|_F^2}_{\text{Reconstruction Error}} + \lambda \underbrace{\|\mathbf{L}\odot(\mathbf{Y}-\mathbf{BS})\|_F^2}_{\text{Classification Error}}, \quad (3)$$

where $\mathbf{A} \in \mathbb{R}_{\geq 0}^{n_1 \times r}$, $\mathbf{B} \in \mathbb{R}_{\geq 0}^{k \times r}$, $\mathbf{S} \in \mathbb{R}_{\geq 0}^{r \times n_2}$, and the regularization parameter $\lambda > 0$ governs the relative importance of the supervision term [23]. The binary weight matrix $\mathbf{W}$ accommodates missing data by indicating observed and unobserved data entries. Similarly, $\mathbf{L} \in \mathbb{R}^{k \times n_2}$ is a weight matrix that indicates the presence or absence of a label. Multiplicative updates have been previously developed for SSNMF for the Frobenius norm, and the resulting performance of clustering and classification is improved by incorporating data labels into NMF [23].

C. Related Work

In this section, we describe related work most relevant to our own. This is not meant to be a comprehensive study of these areas.

Statistical Motivation for NMF: The most common discrepancy measures for NMF $\|\cdot\|_F$-NMF and $D(\cdot\|\cdot)$-NMF correspond to the MLE given an assumed latent generative model and a Gaussian and Poisson model of uncertainty, respectively [5], [10], [32]. In [5], [32], the authors go further towards a Bayesian approach, introduce application-appropriate prior distributions on the latent factors, and apply *maximum a posteriori* (MAP) estimation. Additionally, under certain conditions, it is shown that $D(\cdot\|\cdot)$-NMF is equivalent to probabilistic latent semantic indexing [9].

Dimension Reduction and Learning: There has been much work developing dimensionality-reduction models that are supervision-aware. Semi-supervised clustering makes use of known label information or other supervision *and* the data features while forming clusters [1], [20], [33]. These techniques generally make use of label information in the cluster initialization or during cluster updating via must-link and cannot-link constraints; empirically, these approaches are seen to increase mutual information between computed clusters and user-assigned labels [1]. Semi-supervised feature extraction makes use of supervision information in the feature extraction process [12], [30]. These approaches are generally *filter*- or *wrapper*-based approaches, and distinguished by their underlying supervision type [30].

Semi-supervised and Joint NMF: Since the seminal work of Lee et al. [23], semi-supervised NMF models have been studied in a variety of settings. The works [6], [11], [18] propose models which exploit cannot-link or must-link supervision. In [7], the authors introduce a model with information divergence penalties on the reconstruction and on supervision terms which influence the learned factorization to approximately reconstruct coefficients learned before factorization by a support-vector machine (SVM). Several works [19], [35], [37] propose a supervised NMF model that incorporates Fisher discriminant constraints into NMF for classification. Furthermore, joint factorization of two data matrices, like that of SSNMF, is described more generally and denoted Simultaneous NMF in [8].

D. Overview of Proposed Models

We propose two SSNMF formulations for document classification, both of which utilize information divergence on the first (data reconstruction) term. This is a natural choice since many representations of document data (e.g., bag-of-words, n-grams, etc.) correspond to counts of word patterns in the data and are naturally modelled by Poisson distribution(s), which leads to the information divergence in the MLE model [5], [17], [32]. Our proposed models accept document data $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$, supervision matrix as $\mathbf{Y} \in \mathbb{R}_{\geq 0}^{k \times n_2}$, and target dimension r; we denote the models as $(D(\cdot,\cdot), \|\cdot\|_F)$-SSNMF,

$$\underset{\mathbf{A},\mathbf{S},\mathbf{B}\geq 0}{\operatorname{argmin}} \underbrace{D(\mathbf{W}\odot\mathbf{X}, \mathbf{W}\odot\mathbf{AS})}_{\text{Reconstruction Error}} + \lambda \underbrace{\|\mathbf{L}\odot(\mathbf{Y}-\mathbf{BS})\|_F^2}_{\text{Classification Error}}, \quad (4)$$

and $(D(\cdot,\cdot), D(\cdot,\cdot))$-SSNMF,

$$\underset{\mathbf{A},\mathbf{S},\mathbf{B}\geq 0}{\operatorname{argmin}} \underbrace{D(\mathbf{W}\odot\mathbf{X}, \mathbf{W}\odot\mathbf{AS})}_{\text{Reconstruction Error}} + \lambda \underbrace{D(\mathbf{L}\odot\mathbf{Y}, \mathbf{L}\odot\mathbf{BS})}_{\text{Classification Error}}. \quad (5)$$

In each model, the matrix $\mathbf{A} \in \mathbb{R}_{\geq 0}^{n_1 \times r}$ provides a basis for the lower-dimensional space, $\mathbf{S} \in \mathbb{R}_{\geq 0}^{r \times n_2}$ provides the coefficients representing the projected data in this space, and $\mathbf{B} \in \mathbb{R}_{\geq 0}^{k \times r}$ provides the supervision model which predicts the targets given the representation of points in the lower-dimensional space. We allow for missing data and labels

or confidence-weighted errors via the data-weighting matrix $\mathbf{W} \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$ and the label-weighting matrix $\mathbf{L} \in \mathbb{R}_{\geq 0}^{k \times n_2}$. Each resulting joint-factorization model is defined by the error functions applied to the reconstruction and supervision factorization terms.

II. SSNMF Models: Motivation and Methods

In this section, we present a statistical MLE motivation of several variants of the SSNMF model, introduce the general semi-supervised models, and provide a multiplicative updates method for each variant. While historically the focus of SSNMF studies have been on classification [23], we highlight that this joint factorization model can be applied quite naturally to regression tasks.

A. Maximum Likelihood Estimation

In this section, we demonstrate that specific forms of our proposed variants of SSNMF are maximum likelihood estimators for given models of uncertainty or noise in the data matrices $\mathbf{X}$ and $\mathbf{Y}$. These different uncertainty models have their likelihood function maximized by different error functions chosen for reconstruction and supervision errors, R and S. We summarize these results next; each MLE derived is a specific instance of a general model discussed in Section II-B, in [23], or in [15]. As mentioned previously, models which make use of the information-divergence objective are a natural choice since many representations of document data (e.g., bag-of-words, n-grams, etc.) are naturally modelled by Poisson distribution(s), which leads to the information divergence in the MLE model [5], [17], [32].

Maximum Likelihood Estimators *Suppose that the observed data* $\mathbf{X}$ *and supervision information* $\mathbf{Y}$ *have entries given as the sum of random variables,*

$$\mathbf{X}_{\gamma,\tau} = \sum_{i=1}^{r} x_{\gamma,i,\tau} \quad \textit{and} \quad \mathbf{Y}_{\eta,\tau} = \sum_{i=1}^{r} y_{\eta,i,\tau},$$

and that the set of $\mathbf{X}_{\gamma,\tau}$ *and* $\mathbf{Y}_{\eta,\tau}$ *are statistically independent conditional on* $\mathbf{A}, \mathbf{B}$*, and* $\mathbf{S}$*.*

1) *When* $x_{\gamma,i,\tau}$ *and* $y_{\eta,i,\tau}$ *have distributions*

$$\mathcal{N}\left(x_{\gamma,i,\tau} | \mathbf{A}_{\gamma,i}\mathbf{S}_{i,\tau}, \sigma_1\right) \textit{ and } \mathcal{N}\left(y_{\eta,i,\tau} | \mathbf{B}_{\eta,i}\mathbf{S}_{i,\tau}, \sigma_2\right)$$

respectively, the maximum likelihood estimator is

$$\underset{\mathbf{A},\mathbf{B},\mathbf{S}\geq 0}{\operatorname{argmin}} \ \|\mathbf{X} - \mathbf{AS}\|_F^2 + \frac{\sigma_1}{\sigma_2}\|\mathbf{Y} - \mathbf{BS}\|_F^2.$$

2) *When* $x_{\gamma,i,\tau}$ *and* $y_{\eta,i,\tau}$ *have distributions*

$$\mathcal{N}\left(x_{\gamma,i,\tau} | \mathbf{A}_{\gamma,i}\mathbf{S}_{i,\tau}, \sigma_1\right) \textit{ and } \mathcal{PO}\left(y_{\eta,i,\tau} | \mathbf{B}_{\eta,i}\mathbf{S}_{i,\tau}\right)$$

respectively, the maximum likelihood estimator is

$$\underset{\mathbf{A},\mathbf{B},\mathbf{S}\geq 0}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{AS}\|_F^2 + 2r\sigma_1 D(\mathbf{Y}\|\mathbf{BS}).$$

3) *When* $x_{\gamma,i,\tau}$ *and* $y_{\eta,i,\tau}$ *have distributions*

$$\mathcal{PO}\left(x_{\gamma,i,\tau} | \mathbf{A}_{\gamma,i}\mathbf{S}_{i,\tau}\right) \textit{ and } \mathcal{N}\left(y_{\eta,i,\tau} | \mathbf{B}_{\eta,i}\mathbf{S}_{i,\tau}, \sigma_2\right)$$

respectively, the maximum likelihood estimator is

$$\underset{\mathbf{A},\mathbf{B},\mathbf{S}\geq 0}{\operatorname{argmin}} D(\mathbf{X}\|\mathbf{AS}) + \frac{1}{2r\sigma_2}\|\mathbf{Y} - \mathbf{BS}\|_F^2.$$

4) *When* $x_{\gamma,i,\tau}$ *and* $y_{\eta,i,\tau}$ *have distributions*

$$x_{\gamma,i,\tau} \sim \mathcal{PO}\left(x_{\gamma,i,\tau} | \mathbf{A}_{\gamma,i}\mathbf{S}_{i,\tau}\right) \textit{ and } \mathcal{PO}\left(y_{\eta,i,\tau} | \mathbf{B}_{\eta,i}\mathbf{S}_{i,\tau}\right)$$

respectively, the maximum likelihood estimator is

$$\underset{\mathbf{A},\mathbf{B},\mathbf{S}\geq 0}{\operatorname{argmin}} D(\mathbf{X}\|\mathbf{AS}) + D(\mathbf{Y}\|\mathbf{BS}).$$

We note that 4 follows from [5], [10], [32], but the others are distinct from previous MLE derivations due to the difference in the distributions assumed on data $\mathbf{X}$ and supervision $\mathbf{Y}$.

B. Multiplicative Updates

The multiplicative updates method for all methods can be derived as follows [23]. Suppose that the gradient of the objective function F with respect to one of the variables Θ has a decomposition that is of the form:

$$\nabla_\Theta F = [\nabla_\Theta F]^+ - [\nabla_\Theta F]^-,$$

where $[\nabla_\Theta F]^+ > 0$ and $[\nabla_\Theta F]^- > 0$. Then multiplicative update for Θ has the form

$$\Theta \leftarrow \Theta \odot \frac{[\nabla_\Theta F]^-}{[\nabla_\Theta F]^+}.$$

We provide multiplicative updates for all three methods. The pseudocodes for these methods are provided in [15].

Implementation of these methods and code for experiments is available in the Python package `SSNMF` [14]. Finally, we note that the behavior of these models and methods are dependent on the hyperparameters r, λ, and N. One can select the parameters according to *a priori* information or use a heuristic selection technique; we use both and indicate selected parameters and method of selection.

C. Classification Framework

Here we describe a framework for using any of the SSNMF models for classification tasks. Given training data $\mathbf{X}_{\text{train}}$ (with any missing data indicated by matrix $\mathbf{W}_{\text{train}}$) and labels $\mathbf{Y}_{\text{train}}$, and testing data $\mathbf{X}_{\text{test}}$ (with unknown data indicated by matrix $\mathbf{W}_{\text{test}}$), we first train our $(R(\cdot\|\cdot), S(\cdot\|\cdot))$-SSNMF model to obtain learned dictionaries $\mathbf{A}_{\text{train}}$ and $\mathbf{B}_{\text{train}}$, where $R(\cdot\|\cdot)$ and $S(\cdot\|\cdot))$ denote specific metrics. We then use these learned matrices to obtain the representation of test data in the subspace spanned by $\mathbf{A}_{\text{train}}$, $\mathbf{S}_{\text{test}}$, and the predicted labels for the test data $\mathbf{Y}_{\text{test}}$.

Single-label Classification. This process is:

1) Compute $\mathbf{A}_{\text{train}}, \mathbf{B}_{\text{train}}, \mathbf{S}_{\text{train}}$ as

$$\underset{\mathbf{A},\mathbf{B},\mathbf{S}\geq 0}{\operatorname{argmin}} R(\mathbf{W}_{\text{train}} \odot \mathbf{X}_{\text{train}}, \mathbf{W}_{\text{train}} \odot \mathbf{AS}) + \lambda S(\mathbf{Y}_{\text{train}}, \mathbf{BS}).$$

2) Solve $\mathbf{S}_{\text{test}} = \underset{\mathbf{S}\geq 0}{\operatorname{argmin}} R(\mathbf{W}_{\text{test}} \odot \mathbf{X}_{\text{test}}, \mathbf{W}_{\text{test}} \odot \mathbf{A}_{\text{train}}\mathbf{S})$.

3) Compute predicted labels as $\hat{\mathbf{Y}}_{\text{test}} = \text{label}(\mathbf{B}_{\text{train}}\mathbf{S}_{\text{test}})$, where label$(\cdot)$ assigns the largest entry of each column to 1 and all other entries to 0.

In step 1, we compute $\mathbf{A}_{\text{train}}, \mathbf{B}_{\text{train}}$, and $\mathbf{S}_{\text{train}}$ using implementations of the multiplicative updates methods described above. In step 2, we use either a nonnegative least-squares method (if $R = \|\cdot\|_F$) or one-sided multiplicative updates only updating $\mathbf{S}_{\text{test}}$ (if $R = D(\cdot\|\cdot)$). We note that this framework is significantly different than the classification framework proposed in [23]; in particular, we use the classifier $\mathbf{B}$ learned by SSNMF, rather than independent SVM trained on the SSNMF-learned lower-dimensional representation to allow for an additional layer of interpretability.

Multi-label Classification. This framework generalizes to multi-label classification simply. One first applies only steps (1) and (2) of the process above, forms $\hat{\mathbf{Y}}_{\text{test}} = \mathbf{B}_{\text{train}}\mathbf{S}_{\text{test}}$, and then applies a thresholding technique to decide the set of predicted labels for each data point; values above the threshold correspond to predicted labels, and those below to unpredicted labels. There are many ways to do this; we instead vary the threshold uniformly between the minimum and maximum output values for each data point and report the highest encountered model metric.

III. Experimental Data and Results

In this section, we quantitatively evaluate the proposed methods on several document classification datasets to illustrate the promise of SSNMF models in both topic modeling and classification.

A. 20 Newsgroups Data Experiments

We first present our experiment on a subset of the 20 Newsgroups dataset [29], summarized in Table I, where highlight the advantages of our SSNMF models and framework over benchmark methods.

We compute the term frequency–inverse document frequency representation for the documents, treat the groups as classes and assign them labels, and treat the subgroups as (unlabeled) latent topics in the data. We compare to the linear Support Vector Machine (SVM) and Multinomial Naive Bayes (NB) (see e.g., [26]) classifiers, where the groups are treated as classes. We also apply SVM as a classifier to the low-dimensional representation obtained from the NMF models, where (for both NMF and SSNMF models) we consider rank equal to 13 reflecting the number of subgroups in the dataset. We consider all SSNMF models with the training process described in Section II-C with the maximum number of iterations (number of multiplicative updates) $N = 50$; our stopping criterion is the earlier of N iterations or relative errorbelow tolerance tol. We select the hyperparameters tol and λ for the models by searching over different values and selecting those with the highest average classification accuracy on the validation set.

We report in Table II the average test classification accuracy for each of the models over 11 trials. We define the test classification accuracy as $\sum_{i=1}^{n} \delta(\mathbf{Y}_i, \hat{\mathbf{Y}}_i)/n$, where $\delta(u, v) = 1$ for $u = v$, and 0 otherwise, and where $\mathbf{Y}_i$ and $\hat{\mathbf{Y}}_i$ are true and predicted labels, respectively. We observe that the accuracy of $(D(\cdot\|\cdot), \|\cdot\|_F)$-SSNMF is comparable to Multinomial NB which performs classification in the high-dimensional space. Note that the SSNMF models, which provide both dimensionality-reduction and classification in that lower-dimensional space, do not suffer great accuracy loss which suggests that the simultaneously learned low-dimensional representation serves the classification task well. The SSNMF framework provides an intermediate layer that allows for additional interpretability by representing the data points in the low-dimensional topics space, where we learn about the shared and discriminative topics between classes. This serves the purpose of topic modeling (dimensionality reduction and clustering) and classification. Further, we observe that $(D(\cdot\|\cdot), \|\cdot\|_F)$-SSNMF performs significantly better than $D(\cdot\|\cdot)$-NMF + SVM in terms of accuracy emphasizing the importance of learning simultaneously a linear classifier and a low-dimensional representation.

Here, we consider the "typical" decomposition for the $(D(\cdot\|\cdot), \|\cdot\|_F)$-SSNMF by selecting the decomposition corresponding to the median test classification accuracy. We display in Figure 1 the column-sum normalized $\mathbf{B}_{\text{train}}$ matrix of the decomposition, where each column illustrates the distribution of topic association to classes. We display in Table III the top

TABLE I
SUBSET OF THE 20 NEWSGROUPS DATASET [29] CONSISTING OF 5 GROUPS AND 13 SUBGROUPS PARTITIONED ROUGHLY ACCORDING TO SUBJECTS.

Groups	Subgroups
Computers	graphics, mac.hardware, windows.x
Sciences	crypt(ography), electronics, space
Politics	guns, mideast
Religion	atheism, christian(ity)
Recreation	autos, baseball, hockey

TABLE II
MEAN (AND STD. DEV.) OF TEST CLASSIFICATION ACCURACY FOR EACH OF THE MODELS ON THE SUBSET OF THE 20 NEWSGROUPS DATASET DESCRIBED IN TABLE I.

Model	Class. accuracy % (sd)
$(\|\cdot\|_F, \|\cdot\|_F)$	79.37 (0.47)
$(\|\cdot\|_F, D(\cdot\|\cdot))$	79.51 (0.38)
$(D(\cdot\|\cdot), \|\cdot\|_F)$	**81.88** (0.44)
$(D(\cdot\|\cdot), D(\cdot\|\cdot))$	81.50 (0.47)
$\|\cdot\|_F$-NMF + SVM	70.99 (2.71)
$D(\cdot\|\cdot)$-NMF + SVM	74.75 (2.50)
SVM	80.70 (0.27)
Multinomial NB	**82.28**

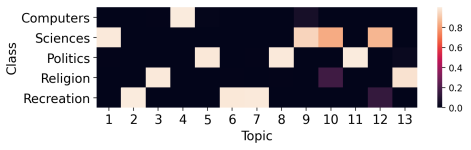

Fig. 1. The normalized $\mathbf{B}_{\text{train}}$ matrix for the $(D(\cdot\|\cdot), \|\cdot\|_F)$ SSNMF decomposition corresponding to the median test classification accuracy equal to 81.78% showcasing the topic distribution over classes.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
would	game	god	x	would	game	players	people	would	one	israel	like	god
space	team	would	thanks	armenian	one	team	israel	chip	us	guns	anyone	people
government	car	one	anyone	one	like	car	gun	key	get	people	available	church
use	games	jesus	graphics	people	car	last	right	algorithm	could	gun	key	one
key	engine	think	know	fbi	baseball	year	government	use	like	well	probably	christians

TABLE III

TOP KEYWORDS REPRESENTING EACH TOPIC OF THE $(D(\cdot\|\cdot), \|\cdot\|_F)$-SSNMF MODEL REFERRED TO IN FIGURE 1. WE (QUALITATIVELY) OBSERVE FOR EXAMPLE THAT TOPIC 5, TOPIC 8, AND TOPIC 11 CAPTURE THE SUBJECTS OF MIDDLE EAST AND GUNS ("ISRAEL", "GOVERNMENT", "GUN"). ALL THREE TOPICS ARE ASSOCIATED WITH CLASS POLITICS; SEE FIGURE 1. WE ALSO OBSERVE THAT TOPIC 1 AND TOPIC 9 RELATE TO ELECTRONICS/CRYPTOGRAPHY. BOTH ARE ASSOCIATED TO CLASS SCIENCES.

5 keywords (i.e. those that have the highest weight in topic column of $\mathbf{A}_{\text{train}}$) for each topic of the $(D(\cdot\|\cdot), \|\cdot\|_F)$-SSNMF of Figure 1.

B. Reuters Data Experiments

We next present our experiment on the Reuters Corpus [24]. This corpus, which we download via `NLTK`, contains 10,788 news documents totaling 1.3 million words. The documents are classified into 90 classes (each document can have multiple labels and most do), and are grouped into two fixed sets, called "training" and "test."

We compute the term frequency–inverse document frequency representation for the documents, and apply the training process described in Section II-C (steps (1) and (2) for this multi-label classification task) with the maximum number of multiplicative updates iterations $N = 10$. We set the hyperparameters $k = 200$ and $\lambda = 1$ for all models in this experiment. In Table IV, we present the mean and standard deviation of the micro-F1 score calculated on the test set over 100 trials. In each trial, we compute the matrix $\hat{\mathbf{Y}}_{\text{test}} = \mathbf{B}_{\text{train}}\mathbf{S}_{\text{test}}$ and vary the applied threshold for predicting labels uniformly (between data points) over the interval from the minimum entry to the maximum entry (in each data point). That is, for each $\alpha \in [0, 1]$, label i is predicted for data point j if the (i, j)th entry of $\hat{\mathbf{Y}}_{\text{test}} \geq \min \hat{\mathbf{Y}}_{\text{test},j} + \alpha[\max \hat{\mathbf{Y}}_{\text{test},j} - \min \hat{\mathbf{Y}}_{\text{test},j}]$ where $\hat{\mathbf{Y}}_{\text{test},j}$ is the jth column of $\hat{\mathbf{Y}}_{\text{test}}$. In each trial, we compute this thresholded prediction for all $\alpha \in [0, 1]$ and choose the largest micro-F1 score encountered.

TABLE IV

MEAN (AND STD. DEV.) OF TEST MICRO-F1 SCORE FOR EACH OF THE MODELS ON THE SUBSET OF THE REUTERS DATASET.

Model	Micro-F1 Score % (sd)
$(\|\cdot\|_F, \|\cdot\|_F)$	**41.59 (1.37)**
$(\|\cdot\|_F, D(\cdot\|\cdot))$	34.46 (1.76)
$(D(\cdot\|\cdot), \|\cdot\|_F)$	38.15 (2.08)
$(D(\cdot\|\cdot), D(\cdot\|\cdot))$	36.86 (2.51)

IV. CONCLUSION

In this work, we have have proposed several SSNMF models, and have demonstrated that these models and that of [23] are MLE in the case of specific distributions of uncertainty assumed on the data and labels. We provided multiplicative update training methods for each model, and demonstrated the ability of these models to perform classification.

In future work, we plan to take a Bayesian approach to SSNMF by assuming data-appropriate priors and performing maximum *a posteriori* estimation. Furthermore, we will form a general framework of MLE models for exponential family distributions of uncertainty, and study the class of models where multiplicative update methods are feasible.

ACKNOWLEDGEMENTS

The authors are appreciative of useful conversations with William Swartworth, Joshua Vendrow, and Liza Rebrova.

REFERENCES

[1] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Proc. Int. Conf. Mach. Learn.* Citeseer, 2002.

[2] Michael W Berry, Nicolas Gillis, and François Glineur. Document classification using nonnegative matrix factorization and underapproximation. In *2009 IEEE International Symposium on Circuits and Systems*, pages 2782–2785. IEEE, 2009.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3(Jan):993–1022, 2003.

[4] Harold Borko and Myrna Bernick. Automatic document classification. *Journal of the ACM (JACM)*, 10(2):151–162, 1963.

[5] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Comput. Intel. Neurosc.*, 2009, 2008.

[6] Y. Chen, M. Rege, M. Dong, and J. Hua. Non-negative matrix factorization for semi-supervised data clustering. *Knowl. Inf. Syst.*, 17(3):355–379, 2008.

[7] Y. Cho and L. K. Saul. Nonnegative matrix factorization for semi-supervised dimensionality reduction. *arXiv preprint arXiv:1112.3714*, 2011.

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

[9] C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data An.*, 52(8):3913–3927, 2008.

[10] P. Favaro and S. Soatto. *3-d shape estimation and image restoration: Exploiting defocus and motion-blur*. Springer Science & Business Media, 2007.

[11] W. Fei, L. Tao, and Z. Changshui. Semi-supervised clustering via matrix factorization. In *Proc. SIAM Int. Conf. on Data Mining*, 2008.

[12] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, 5(Jan):73–99, 2004.

[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(Mar):1157–1182, 2003.

[14] J. Haddock, L. Kassab, and S. Li. SSNMF, 2020.

[15] Jamie Haddock, Lara Kassab, Sixian Li, Alona Kryshchenko, Rachel Grotheer, Elena Sizikova, Chuntian Wang, Thomas Merkh, RWMA Madushani, Miju Ahn, et al. Semi-supervised nmf models for topic modeling in learning tasks. *arXiv preprint arXiv:2010.07956*, 2020.

[16] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[17] Le Thi Khanh Hien and Nicolas Gillis. Algorithms for nonnegative matrix factorization with the kullback-leibler divergence. *arXiv preprint arXiv:2010.01935*, 2020.

[18] Y. Jia, S. Kwong, J. Hou, and W. Wu. Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization. *IEEE T. Neur. Net. Lear.*, 2019.

[19] Y. Jia, Y. Wang, C. Turk, and M. Hu. Fisher non-negative matrix factorization for learning local features. In *Proc. Asian Conf. Comp. Vis.*, pages 27–30. Citeseer, 2004.

[20] D. Klein, S. D. Kamvar, and C. D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. Technical report, Stanford, 2002.

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

[22] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proc. Adv. Neur. In.*, pages 556–562, 2001.

[23] H. Lee, J. Yoo, and S. Choi. Semi-supervised nonnegative matrix factorization. *IEEE Signal Proc. Let.*, 17(1):4–7, 2009.

[24] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[25] L. Li, G. Lebanon, and H. Park. Fast Bregman divergence NMF using Taylor expansion and coordinate descent. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 307–315, 2012.

[26] C. D. Manning, H. Schütze, and P. Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[27] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[28] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[29] J. Rennie. 20 Newsgroups, 2008.

[30] R. Sheikhpour, M. Sarram, S. Gharaghani, and M. Chahooki. A survey on semi-supervised feature selection methods. *Pattern Recogn.*, 64:141–158, 2017.

[31] S. Sra and I. S. Dhillon. Generalized nonnegative matrix approximations with Bregman divergences. In *Proc. Adv. Neur. In.*, pages 283–290, 2006.

[32] T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. IEEE Int. Conf. on Acoust., Speech and Sig. Process.*, pages 1825–1828. IEEE, 2008.

[33] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al. Constrained k-means clustering with background knowledge. In *Proc. Int. Conf. Mach. Learn.*, volume 1, pages 577–584, 2001.

[34] W. Wang and M. A. Carreira-Perpinán. The role of dimensionality reduction in classification. In *Proc. AAAI Conf. on Artif. Intel.*, pages 2128–2134, 2014.

[35] Y. Xue, C. S. Tong, W. Chen, W. Zhang, and Z. He. A modified non-negative matrix factorization algorithm for face recognition. In *Proc. Int. Conf. on Pattern Recognition*, volume 3, pages 495–498. IEEE, 2006.

[36] Z. Yang, H. Zhang, Z. Yuan, and E. Oja. Kullback-Leibler divergence for nonnegative matrix factorization. In *Proc. Int. Conf. on Artif. Neural Networks*, pages 250–257. Springer, 2011.

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