Abstract

We propose several new models for semi-supervised non-negative matrix factorization (SSNMF) and provide motivation for SSNMF models as maximum likelihood estimators given specific distributions of uncertainty. We present multiplicative updates training methods for each new model, and demonstrate the application of these models to classification, although they are flexible to other supervised learning tasks. We illustrate the promise of these models and training methods on both synthetic and real data, and achieve high classification accuracy on the 20 Newsgroups dataset.

Keywords: semi-supervised nonnegative matrix factorization, maximum likelihood estimation, multiplicative updates

1 Introduction

Frequently, one is faced with the problem of performing a (semi-)supervised learning task on extremely high-dimensional data which contains redundant information. A common approach is to first apply a dimensionality-reduction or feature extraction technique (e.g., PCA [34]), and then train the model for the learning 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 the learning task [18]. 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 supervised learning model [13] [4] [36].

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 [28] 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;
• we perform experiments on synthetic and real data which illustrate the promise of these models in both topic modeling and supervised learning tasks; and
• we demonstrate the promise of SSNMF models for classification relative to the performance of other classifiers on a common benchmark data set.

1.1 Organization

The paper is organized as follows. We begin by briefly defining notation in Section 1.2 and provide a preliminary review of the fundamental models, nonnegative matrix factorization (NMF) [26] and SSNMF [28] in Section 1.3, other related work in Section 1.4, and the proposed models in Section 1.5. We motivate the proposed models and that of [28] via MLE in Section 2.1 and present the multiplicative update methods for training in Section 2.2 and present details of a framework for classification with these models in Section 2.3. We present experimental evidence illustrating the promise of the SSNMF models in Section 3, including experiments on synthetic data in Section 3.1 which are motivated by the MLE of Section 2.1 and experiments on the 20 Newsgroups dataset in Section 3.2. Finally, we end with some conclusions and discussion of future work in Section 4.
1.2 Notation

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

\[
D(A\|B) = \sum_{i,j} \left( A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right),
\]

where \( D(A\|B) \geq 0 \) with equality if and only if \( A = B \) \cite{27}. Because the information divergence reduces to the Kullback-Leibler divergence when A and B represent probability distributions, i.e., \( \sum A_{ij} = \sum B_{ij} = 1 \), it is often referred to as the generalized Kullback-Leibler divergence \cite{12}.

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

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

1.3 Preliminaries

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

Nonnegative Matrix Factorization

Given a nonnegative matrix \( X \in \mathbb{R}_{\geq 0}^{n_1 \times n_2} \) and a target dimension \( r \in \mathbb{N} \), NMF decomposes \( X \) into a product of two low-dimensional nonnegative matrices. The model seeks \( A \) and \( S \) so that \( X \approx AS \), where \( A \in \mathbb{R}_{\geq 0}^{n_1 \times r} \) is called the dictionary matrix and \( S \in \mathbb{R}_{\geq 0}^{r \times n_2} \) is called the representation matrix. Typically, \( r \) is chosen such that \( r < \min(n_1,n_2) \) to reduce the dimension of the original data matrix or reveal latent themes in the data. Data points are typically stored as columns of \( X \), thus \( n_1 \) represents the number of features, and \( n_2 \) represents the number of samples. The columns of \( A \) are generally referred to as topics, which are characterized by features of the data set. Each column of \( S \) provides the approximate representation of the respective column in \( X \) in the lower-dimensional space spanned by the columns of \( A \). Thus, the data points are well approximated by an additive linear combination of the latent topics.

Several formulations for this nonnegative approximation, \( X \approx AS \), have been studied \cite{9,26,27,40}; e.g.,

\[
\arg\min_{A \geq 0, S \geq 0} \|X - AS\|_F^2 \quad \text{and} \quad \arg\min_{A \geq 0, S \geq 0} D(X\|AS),
\]

where \( D(\cdot\|\cdot) \) is the information divergence defined in \cite{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 \cite{9} for discussions of similarity measures and generalized divergences (where information divergence is a particular case), and \cite{29,40} for generalized nonnegative matrix approximations with Bregman divergences.

Multiplicative update algorithms for both formulations of (2) have been proposed \cite{26,27}. These algorithms are widely adopted because they are easy to implement, do not require user-specified hyperparameters, preserve the nonnegativity constraints, and have desirable monotonicity properties \cite{27}. Other popular algorithms include projected gradient descent and alternating least-squares \cite{9,23,24,30}.

NMF has gained popularity recently due to large scale data demands of applications such as document clustering \cite{18,35,41,42}, image processing \cite{17,20,28}, feature extraction \cite{2,15,38,31}, financial data mining \cite{10}, audio processing \cite{8,16}, and genetics \cite{41}.

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 \( X \in \mathbb{R}_{\geq 0}^{n_1 \times n_2} \) and a class label matrix \( Y \in \mathbb{R}_{\geq 0}^{k \times n_2} \), SSNMF is defined by

\[
\arg\min_{A,S,B \geq 0} \left\{ (W \odot (X - AS))^2_F + \lambda (L \odot (Y - BS))^2_F \right\}.
\]

1.4 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. We focus on work in three main areas: statistical motivation for NMF models, models for simultaneous dimension reduction and supervised learning, and semi-supervised and joint NMF models.
Figure 1: Given the number of classes \( k \), and a desired dimension \( r \), SSNMF is formulated as a joint factorization of a data matrix \( X \in \mathbb{R}^{n_1 \times n_2} \) and a label matrix \( Y \in \mathbb{R}^{k \times n_2} \), sharing representation factor \( S \in \mathbb{R}^{r \times n_2} \).

**Statistical Motivation for NMF**

The most common discrepancy measures for NMF are the Frobenius norm and the information divergence. One reason for this popularity is that \( ||·||_F \)-NMF and \( D(·||·)\)-NMF correspond to the MLE given an assumed latent generative model and a Gaussian and Poisson model of uncertainty, respectively [5, 12, 41]. In [5, 11], 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, \( D(·||·)\)-NMF is equivalent to probabilistic latent semantic indexing [11].

**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 [11, 25, 42]. 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 [14, 52]. These approaches are generally *filter* or *wrapper*-based approaches, and distinguished by their underlying supervision type [29].

**Semi-supervised and Joint NMF**

Since the seminal work of Lee et al. [28], semi-supervised NMF models have been studied in a variety of settings. The works [6, 13, 21] propose models which exploit cannot-link or must-link supervision. In [21], 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 [22, 13, 47] 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 [9].

### 1.5 Overview of Proposed Models

Our proposed models generalize NMF to supervised learning tasks and provide a topic model which simultaneously provides a lower dimensional representation of the data and a predictive model for targets. We denote the data matrix as \( X \in \mathbb{R}^{n_1 \times n_2} \) and the supervision matrix as \( Y \in \mathbb{R}^{k \times n_2} \). Following along with [28], the data observations are the columns of \( X \) and the associated targets (e.g., labels) are the columns of \( Y \). Our models seek \( A \in \mathbb{R}^{n_1 \times r} \), \( S \in \mathbb{R}^{r \times n_2} \), and \( B \in \mathbb{R}^{k \times r} \) which jointly factorize \( X \) and \( Y \); that is \( X \approx AS \) and \( Y \approx BS \). We point out the simple fact that these joint factorizations can be stacked into a single NMF (visualized in Figure 1)

\[
\begin{bmatrix}
X \\
y
\end{bmatrix}
\approx
\begin{bmatrix}
A \\
B
\end{bmatrix}
S. 
\]

(4)

In each model, the matrix \( A \in \mathbb{R}^{n_1 \times r} \) provides a basis for the lower-dimensional space, \( S \in \mathbb{R}^{r \times n_2} \) provides the coefficients representing the projected data in this space, and \( B \in \mathbb{R}^{k \times r} \) 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 \( W \in \mathbb{R}^{n_1 \times n_2} \) and the label-weighting matrix \( L \in \mathbb{R}^{k \times n_2} \). Each resulting joint-factorization model is defined by the error functions applied to the reconstruction and supervision factorization terms. We denote the model

\[
\arg\min_{A,S,B \geq 0} \left( R(W \odot X, W \odot AS) + \lambda S(L \odot Y, L \odot BS) \right)
\]

(5)

as \( (R(\cdot, \cdot), S(\cdot, \cdot))-SSNMF \) for specific choices of \( R \) and \( S \). Here, \( R(\cdot, \cdot) \) and \( S(\cdot, \cdot) \) are the error functions applied to the reconstruction term and supervision term, respectively. For clarity, we include Table 1 below which summarizes existing and proposed models, where each proposed model is of the form (4) for specific choices of error functions \( R \) and \( S \).

**Table 1: Overview of NMF and SSNMF models.**

| Model | Objective |
|-------|-----------|
| \( \|\cdot\|_F \)-NMF | \( \arg\min_{A \geq 0} \|X - AS\|_F^2 \) |
| \( \|\cdot\|_F \)-SSNMF | \( \arg\min_{A \geq 0} \|W \odot (X - AS)\|_F^2 + \lambda \|L \odot (Y - BS)\|_F^2 \) |
| \( (\|\cdot\|_F, \|\cdot\|_F) \)-SSNMF | \( \arg\min_{A \geq 0} \|W \odot (X - AS)\|_F^2 + \lambda \|L \odot (Y - BS)\|_F^2 \) |
| \( (\|\cdot\|_F, \|\cdot\|_F) \)-SSNMF | \( \arg\min_{A \geq 0} \|W \odot (|X - AS| + |Y - BS|)\|_F^2 \) |
| \( \|\cdot\|_F \)-NMF | \( \arg\min_{A \geq 0} \|X - AS\|_F^2 \) |
| \( (\|\cdot\|_F, \|\cdot\|_F) \)-NMF | \( \arg\min_{A \geq 0} \|W \odot (X - AS)\|_F^2 + \lambda \|L \odot (Y - BS)\|_F^2 \) |
| \( (\|\cdot\|_F, \|\cdot\|_F) \)-SSNMF | \( \arg\min_{A \geq 0} \|W \odot (X - AS)\|_F^2 + \lambda \|L \odot (Y - BS)\|_F^2 \) |
2 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 [28], we highlight that this joint factorization model can be applied quite naturally to regression tasks.

2.1 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 $X$ and $Y$. 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 2.2 or in [28].

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

$X_{\gamma,\tau} = \sum_{i=1}^{r} x_{\gamma,i,\tau}$ and $Y_{\eta,\tau} = \sum_{i=1}^{r} y_{\eta,i,\tau},$

and that the set of $X_{\gamma,\tau}$ and $Y_{\eta,\tau}$ are statistically independent conditional on $A$, $B$, and $S$.

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

$N(x_{\gamma,i,\tau}|A_{\gamma,i,S_{i,\tau}},\sigma_1)$ and $N(y_{\eta,i,\tau}|B_{\eta,i,S_{i,\tau}},\sigma_2)$ respectively, the maximum likelihood estimator is

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

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

$N(x_{\gamma,i,\tau}|A_{\gamma,i,S_{i,\tau}},\sigma_1)$ and $\PO(y_{\eta,i,\tau}|B_{\eta,i,S_{i,\tau}})$ respectively, the maximum likelihood estimator is

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

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

$\PO(x_{\gamma,i,\tau}|A_{\gamma,i,S_{i,\tau}})$ and $N(y_{\eta,i,\tau}|B_{\eta,i,S_{i,\tau}},\sigma_2)$ respectively, the maximum likelihood estimator is

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

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

$x_{\gamma,i,\tau} \sim \PO(x_{\gamma,i,\tau}|A_{\gamma,i,S_{i,\tau}})$ and $\PO(y_{\eta,i,\tau}|B_{\eta,i,S_{i,\tau}})$ respectively, the maximum likelihood estimator is

$$\arg\min_{A,B,S \geq 0} D(X||AS) + D(Y||BS).$$

We note that [1] follows from [5] [12] [11], but the others are distinct from previous MLE derivations due to the difference in the distributions assumed on data $X$ and supervision $Y$. Here, we provide only the MLE derivation for [2] as the other derivations are similar; these are included in the appendix for completeness. We demonstrate that the MLE, in the case that the uncertainty on $X$ is Gaussian distributed and on $Y$ is Poisson distributed, is a specific instance of the $(|| \cdot ||_F, D(||\cdot||))$-SSNMF model.

Our models for the distribution of observed entries of $X$ and $Y$ assume that the mean is given by $E[X] = AS$ and $E[Y] = BS$, and the uncertainty in the set of observations in $X$ is governed by a Gaussian distribution while the set in $Y$ is governed by a Poisson distribution. That is, we consider hierarchical models for $X$ and $Y$ where

$X_{\gamma,\tau} = \sum_{i=1}^{r} x_{\gamma,i,\tau}$ and $y_{\eta,i,\tau} \sim N(x_{\gamma,i,\tau}|A_{\gamma,i,S_{i,\tau}},\sigma_1),$

$Y_{\eta,\tau} = \sum_{i=1}^{r} y_{\eta,i,\tau}$ and $y_{\eta,i,\tau} \sim \PO(y_{\eta,i,\tau}|B_{\eta,i,S_{i,\tau}}).$

Note then that

$X_{\gamma,\tau} \sim N\left(\sum_{i=1}^{r} A_{\gamma,i,S_{i,\tau}}, r\sigma_1\right)$, and

$Y_{\eta,\tau} \sim \PO\left(\sum_{i=1}^{r} B_{\eta,i,S_{i,\tau}}\right).$

due to the summable property of Gaussian and Poisson random variables. We note that this assumes different distributions on the two collections of rows of the NMF [1] with Gaussian and Poisson models of uncertainty.

Assuming that the set of $X_{\gamma,\tau}$ and $Y_{\eta,\tau}$ are statistically independent conditional on $A$, $B$, and $S$, we have that the likelihood $p(X,Y|A,B,S)$ is

$$\prod_{\gamma,\tau} \sum_{i=1}^{r} A_{\gamma,i,S_{i,\tau}} r\sigma_1) \prod_{\eta,\tau} \PO\left(\sum_{i=1}^{r} B_{\eta,i,S_{i,\tau}}\right).$$

We apply the monotonic natural logarithmic function to the likelihood and ignore terms that are invariant with the factor matrices. This transforms the likelihood into a $(|| \cdot ||_F, D(||\cdot||))$-SSNMF objective while preserving the maximizer. That is, the log likelihood (excluding additive terms which are constant with respect to $A$, $B$, and $S$) is

$$\ln p(X,Y|A,B,S) = + \frac{1}{2r\sigma_1} \sum_{\gamma,\tau} \left( X_{\gamma,\tau} - \sum_{i=1}^{r} A_{\gamma,i,S_{i,\tau}} \right)^2$$

$$- \sum_{\eta,\tau} \left[ (BS)_{\eta,\tau} - Y_{\eta,\tau} \log (BS)_{\eta,\tau} + \log (Y_{\eta,\tau} + 1) \right]$$

$$= + \frac{1}{2r\sigma_1} \left[ ||X - AS||_F^2 + 2r\sigma_1 D(Y||BS) \right].$$

Here, =+ denotes suppression of additive terms that do not depend upon $A$, $B$, and $S$. Thus, the maximum likelihood estimators for $A$, $B$, and $S$ are given by

$$\arg\min_{A,B,S \geq 0} ||X - AS||_F^2 + 2r\sigma_1 D(Y||BS).$$
We see that the MLE in the case of Gaussian uncertainty on the observations in $X$ and Poisson uncertainty on the observations in $Y$, is a specific instance of the $(\|\cdot \|_F, D(\cdot))$-SSNMF objective where the regularization parameter $\lambda$ is a multiple of the variance of the Gaussian distribution. The other MLEs are derived similarly; see Appendix A.

An instance of each of the models in Table 1 are MLE for a given model of uncertainty in the observed data $X$ and supervision $Y$. While this motivates our exploration of these models, we present them in more general context next and provide training methods for the general form.

### 2.2 General Models and Mult. Updates

Recall that $(\|\cdot \|_F, \|\cdot \|_F)$-SSNMF is defined by [26] and multiplicative updates are derived in [25]. Now, we propose the general form of $(\|\cdot \|_F, D(\cdot))$-SSNMF, $(D(\cdot)\|\cdot \|_F)$-SSNMF, and $(D(\cdot), D(\cdot))$-SSNMF and present multiplicative updates methods for each model. These three models are novel forms of SSNMF, and besides their statistical motivation via MLE, we demonstrate their promise experimentally in Section 3.

As in [25], our multiplicative updates methods allow for missing (or certainty-weighted) data and missing (or certainty-weighted) supervision information via matrices $W$ and $L$, which represent our knowledge or certainty of the corresponding entries of $X$ and $Y$, respectively. When $W$ is a matrix of all ones (or more generally has all equal entries) and $L$ is a matrix of all zeros, the SSNMF models reduce to either the $\|\cdot \|_F$-NMF or $D(\cdot)$-NMF. The SSNMF model is fully supervised when $\text{supp}(Y) \subset \text{supp}(L)$ and $Y$ contains supervision information for each element in $X$.

The first proposed semi-supervised NMF model is $(\|\cdot \|_F, D(\cdot))$-SSNMF, which is defined by the solution to

$$\arg \min_{A, B, S \geq 0} \|W \odot (X - AS)\|_F^2 + \lambda D(L \odot Y | L \odot BS). \quad (7)$$

We denote this objective function as $F_2(A, B, S; X, Y)$. Similar to the previous SSNMF model, this model seeks a joint factorization of the data matrix $X$ and target matrix $Y$; however, the error functions applied to the reconstruction and classification terms in the objective differ.

The multiplicative updates for this model are provided in Algorithm 1. We provide intuition for the derivation of only this method as the others are similar. The multiplicative updates for $A$, $B$, and $S$ which minimize $F_2$ are derived as follows. The gradient of the objective function of $F_2$ with respect to $A$, $B$, and $S$ are, respectively,

$$\nabla_A F_2 = -2\|W \odot (X - AS)\|_F^\top,$$

$$\nabla_B F_2 = LS^\top - \left[\frac{L \odot Y}{L \odot BS} \odot L\right] S^\top,$$

and

$$\nabla_S F_2 = \lambda B^\top L - \lambda B^\top \left[\frac{L \odot Y}{L \odot BS} \odot L\right] - 2A^\top W \odot (X - AS).$$

### Algorithm 1 $(\|\cdot \|_F, D(\cdot))$-SSNMF mult. updates

**Input:** $X, W \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}, Y, L \in \mathbb{R}_{\geq 0}^{k \times n_2}, r, \lambda, N$

1. Initialize $A \in \mathbb{R}_{\geq 0}^{n_1 \times r}, S \in \mathbb{R}_{\geq 0}^{r \times n_2}, B \in \mathbb{R}_{\geq 0}^{k \times r}$

2. for $i = 1, ..., N$

3. $A \leftarrow A \odot \left[\frac{(W \odot X)S^\top}{(W \odot AS)S^\top}\right]$

4. $B \leftarrow \frac{B}{LS^\top} \odot \left[\frac{(L \odot Y)S^\top}{(L \odot BS)S^\top}\right] L^\top S^\top$

5. $S \leftarrow S \odot \frac{2A^\top W \odot (X - AS) + \lambda B^\top (L \odot BS)S^\top}{2A^\top(W \odot AS) + \lambda B^\top L}$

The multiplicative updates method, Algorithm 1 can be viewed as an entrywise gradient descent method, where the stepsizes are chosen individually for each entry of the updating matrix to ensure nonnegativity. That is, the updates in Algorithm 1 are given by

$$A \rightarrow A - \Gamma \odot \nabla_A F_2 \quad \text{when} \quad \Gamma = \frac{A}{2(W \odot AS)S^\top},$$

$$B \rightarrow B - \Gamma \odot \nabla_B F_2 \quad \text{when} \quad \Gamma = \frac{B}{LS^\top},$$

and

$$S \rightarrow S - \Gamma \odot \nabla_S F_2 \quad \text{when} \quad \Gamma = \frac{S}{2A^\top(W \odot AS) + \lambda B^\top L}.$$

The next proposed semi-supervised NMF model is $(D(\cdot), \|\cdot \|_F)$-SSNMF, defined by the solution to

$$\arg \min_{A, B, S \geq 0} D(W \odot X | W \odot AS) + \lambda \|L \odot (Y - BS)\|_F^2. \quad (8)$$

We denote this objective function as $F_3(A, B, S; X, Y)$. Again, this model seeks a joint factorization of the data matrix $X$ and target matrix $Y$; here the reconstruction error is penalized by the information divergence, while the supervision error is penalized by the Frobenius norm.

Multiplicative updates for this model are provided in Algorithm 2.

The third, and final, proposed semi-supervised NMF model is $(D(\cdot), D(\cdot))$-SSNMF, defined by the solution to

$$\arg \min_{A, B, S \geq 0} D(W \odot X | W \odot AS) + \lambda D(L \odot Y | L \odot BS). \quad (9)$$

We denote this objective function as $F_4(A, B, S; X, Y)$. Again, this model seeks a joint factorization of the data matrix $X$ and target matrix $Y$; here both the reconstruction error and supervision error are penalized by the information divergence error function. The multiplicative updates for this model are provided in Algorithm 3.

As previously stated in Section 2.2, an instance of each family of models, $(\|\cdot \|_F, \|\cdot \|_F)$-SSNMF, $(\|\cdot \|_F, D(\cdot))$-SSNMF, $(D(\cdot), \|\cdot \|_F)$-SSNMF, and $(D(\cdot), D(\cdot))$-SSNMF, correspond to the MLE in the case that the data $X$ and supervision $Y$ are sampled from specific distributions with mean given by a latent lower-dimensional factorization model. One might expect that each model is most appropriately applied when the associated model of uncertainty is a reasonable assumption (i.e., one has a
### 3 Numerical Experiments

In this section, we present numerical experiments of the proposed models applied to both synthetic and real data.

#### 3.1 Synthetic Data Experiments

We expect that, for each pair of distributions of uncertainty, the MLE model derived in Section 2.1 will produce larger likelihood (i.e., smaller relative error) than any other SSNMF model. To confirm this hypothesis, we generate synthetic data according to the four distributions in Section 2.1, train the SSNMF models, and measure the relative error (error is negative likelihood).

Through all experiments, we use the same factor matrices \( A \in \mathbb{R}^{500 \times 5}, S \in \mathbb{R}^{5 \times 500}, \) and \( B \in \mathbb{R}^{500 \times 5} \). Here \( A \) is generated with all entries sampled from the uniform distribution on \([0,1]\), and \( S \) and \( B \) are generated as sparse random matrices with density 0.5 (support sampled uniformly amongst matrix entries) with nonzero entries sampled from the uniform distribution on \([0,1]\).

In each experiment, we train each of the SSNMF models with rank \( r = 5 \) and \( \lambda = 1 \) by running \( N = 100000 \) iterations of the multiplicative updates of [23] and Algorithms 2 and 3 using all data and supervision; each algorithm is initialized with the same matrices \( A^{(0)}, B^{(0)}, \) and \( S^{(0)} \) and we denote the resulting approximate factor matrices as \( A^{(N)}, B^{(N)}, \) and \( S^{(N)} \). We present relative errors averaged over five trials (independent initializations of \( A^{(0)}, B^{(0)}, \) and \( S^{(0)} \)). Here, we let the relative error for objective function \( F \) be

\[
\frac{F(A^{(N)}, B^{(N)}, S^{(N)}; AS, BS)}{F(A^{(0)}, B^{(0)}, S^{(0)}; AS, BS)}.
\]

In our first experiment, we generate data \( X_{ij} \sim N(X_{ij}(AS)_{ij}, 1) \) and supervision \( Y_{ij} \sim \).
Table 2: Experimental results with synthetic data.

| Experiment | 1 | 2 | 3 | 4 |
|------------|---|---|---|---|
| SSNMF      | F₁ err. | F₂ err. | F₃ err. | F₄ err. |
| (∥·∥ₚ, ∥·∥ₚ) | 0.2917 | 0.0081 | 0.0078 | 0.0204 |
| (∥·∥ₚ, D(∥·∥)) | 0.2927 | 0.0065 | 0.0109 | 0.0199 |
| (D(∥·∥), ∥·∥ₚ) | 0.2930 | 0.0093 | 0.0071 | 0.0210 |
| (D(∥·∥), D(∥·∥)) | 0.2922 | 0.0076 | 0.0075 | 0.0172 |

Table 3: 20 Newsgroups and subgroups.

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

We consider all SSNMF models with the training process described in Section 2.3 with the maximum number of iterations (number of multiplicative updates) \( N = 50 \); our stopping criterion is the earlier of \( N \) iterations or relative error \( 10^{-6} \) below tolerance \( tol \).

We also apply SVM as a classifier to the low-dimensional representation obtained from NMF as follows. We consider the default implementation \( \text{[35]} \) of \( \| \cdot \|_{2} \text{-NMF} \) with multiplicative updates, random initialization, and maximum number of iterations \( N = 400 \). We apply NMF on the train data to obtain a vocabulary dictionary matrix \( A_{\text{train}} \) and a document representation \( S_{\text{train}} \). Next, we train an SVM classifier using \( S_{\text{train}} \) and the labels of the train set. We test our model by (i) computing the document representation of the test data \( S_{\text{test}} \) from the learned dictionary \( A_{\text{train}} \) (i.e., step 2 of Section 2.3), then (ii) applying the trained SVM classifier on \( S_{\text{test}} \) to obtain the test predicted labels.

For both NMF and all four SSNMF models, we consider rank (the number of topics) \( v \) equal to 13. We select the hyperparameters \( tol \) and \( \lambda \) for the models by searching over different values and selecting those with the highest average classification accuracy on the validation set; see Appendix C.

Table 4: Mean (and std. dev.) of test classification accuracy for each of the models on 20 Newsgroups data.

| Model                   | Class. accuracy % (sd) |
|-------------------------|------------------------|
| (∥·∥ₚ, ∥·∥ₚ)            | 79.37 (0.47)           |
| (∥·∥ₚ, D(∥·∥))         | 79.51 (0.38)           |
| (D(∥·∥), ∥·∥ₚ)         | 81.88 (0.44)           |
| (D(∥·∥), D(∥·∥))      | 81.50 (0.47)           |
| \( \| \cdot \|_{2} \text{-NMF} + \text{SVM} \) | 70.99 (2.71)           |
| SVM                     | 80.70 (0.27)           |
| Multinomial NB          | 82.28                  |

We report in Table 4 the mean and standard deviation of the test classification accuracy for each of the models over 11 trials. We define the test classification accuracy as \( \sum_{i=1}^{n} \delta(\hat{Y}_{i}, Y_{i})/n \), where \( \delta(u, v) = 1 \) for \( u = v \), and 0 otherwise, and where \( Y_{i} \) and \( \hat{Y}_{i} \) are true and predicted labels, respectively. We observe that \( (D(∥·∥), ∥·∥ₚ)\text{-SSNMF} \) produces the highest average classification accuracy, and is comparable to Multinomial NB.

\(^1\)Our results present this data in its raw form; in particular, we do not capitalize words to reflect common usage. Results are in no way meant to be a political statement.

\(^2\)A larger choice of rank could be made to learn hidden topics within subgroups.
representation in this significantly lower-dimensional space is able to achieve nearly the same accuracy as the higher-dimensional multinomial NB model. We expect that this is due to the relevance of the topic modeling and classification tasks on this data set. We note that while \( (D(\|\cdot\|), \|\cdot\|) \)-SSNMF outperforms the other SSNMF models in terms of classification accuracy, this does not imply that a different model would not produce a lower overall relative error for the objective function. Additionally, the strong performance of \( (D(\|\cdot\|), \|\cdot\|) \)-SSNMF could be due to a number of factors including the choice of hyperparameters. Ongoing and future work will further investigate this phenomenon.

4 Conclusion

In this work, we have have proposed several SSNMF models, and have demonstrated that these models and that of [28] 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 \textit{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 \textit{Proc. Int. Conf. Mach. Learn.} Citeseer, 2002.

[2] M. W. Berry and M. Browne. Email surveillance using non-negative matrix factorization. \textit{Comput. Math. Organ. Th.}, 11(3):249–264, 2005.

[3] S. Bird, E. Klein, and E. Loper. \textit{Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit}. O’Reilly Media, Inc., 2009.

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

[5] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. \textit{Comput. Intel. Neurosc.}, 2009, 2008.
Table 5: Top keywords representing each topic of the $(D(\cdot | \cdot), \| \cdot \|_F)$-SSNMF model referred to in Figure 2.

| 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 |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| world   | game    | team    | would   | x       | armenian | one     | players | team    | people   | would    | israel   | some     |
| space   | get     | thanks  | armenian | one     | like     | game    | people  | chip    | us       | gun      | right    | one      |
| government | use   | games   | one     | anyone  | people   | like    | gun     | key     | get      | last     | government| could  |
|          | key     | engine  | year    | bible   | baseball | year    | game    | right   | algorithm| use      | earth    | probably |
|          | chip    | year    | believe | mac     | israel   | get     | hockey  | say     | bit      | space    | know     | one      |
|          | get     | like    | christian| know   | would    | fire    | would   | season  | would    | would    | system   | see      |
|            | clipper | one     | espn    | get     | say      | get     | jews    | data    | would    | israel   | anyone   | would    |
|            | could   |         |         |         |          |         |         |         |          | bush     |         |         |

[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, and S. Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Proc. Int. Conf. Acoust. Spe. Sig. Process.*, volume 5, pages V–V. IEEE, 2006.

[9] 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.

[10] R. de Fréin, K. Drakakis, S. Rickard, and A. Cichocki. Analysis of financial data using non-negative matrix factorization. In *Proc. Int. Mathematical Forum*, volume 3(38), pages 1853–1870. Hikari, 2008.

[11] 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.

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

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

[14] 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.

[15] E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *Proc. ACM SIGIR Conf. on Research and Development in Inform. Retrieval*, pages 601–602, 2005.

[16] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vannrumste, et al. An exemplar-based NMF approach to audio event detection. In *Proc. IEEE Workshop on Appl. Sig. Process. to Audio and Acoust.*, pages 1–4. IEEE, 2013.

[17] D. Guillamet and J. Vitria. Non-negative matrix factorization for face recognition. In *Proc. Catalanian Conf. on Artif. Intel.*, pages 336–344. Springer, 2002.

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

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

[20] P. O. Hoyer. Non-negative sparse coding. In *Proc. IEEE Workshop on Neural Networks for Sig. Process.*, pages 557–565. IEEE, 2002.

[21] 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.

[22] 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.

[23] D. Kim, S. Sra, and I. S. Dhillon. Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Stat. Anal. Data Min.*, 1(1):38–51, 2008.

[24] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. A.*, 30(2):713–730, 2008.

[25] 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.

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

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

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

[29] 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.
[30] C. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.*, 19(10):2756–2779, 2007.

[31] J. Liu, D. Wang, Y. Gao, C. Zheng, Y. Xu, and J. Yu. Regularized non-negative matrix factorization for identifying differentially expressed genes and clustering samples: a survey. *IEEE/ACM T. Comput. Bio. Bioinf.*, 15(3):974–987, 2017.

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

[33] V. Pauca, F. Shahnaz, M. Berry, and R. Plemmons. Text mining using non-negative matrix factorizations. In *Proc. SIAM Int. Conf. on Data Mining*, pages 452–456. SIAM, 2004.

[34] 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.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

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

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

[38] F. Shahnaz, M. Berry, V. Pauca, and R. Plemmons. Document clustering using nonnegative matrix factorization. *Inform. Process. Manag.*, 42(2):373–386, 2006.

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

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

[41] 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.

[42] 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.

[43] 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.

[44] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. ACM SIGIR Conf. on Research and Development in Inform. Retrieval*, pages 267–273, 2003.

[45] 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.

[46] 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.

[47] 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.
5 Appendix

In this appendix, we provide the remaining MLE derivations (1, 3, and 4) in Appendix A the intuition for the derivation of multiplicative updates for (\(D(\|\cdot\|,\|\cdot\|_F)\)-SSNMF and \((D(\|\cdot\|), D(\|\cdot\|))\)-SSNMF (Algorithms 2 and 3) in Appendix B and additional experimental results on the 20 Newsgroups data set in Appendix C.

A MLE Derivations

We begin by demonstrating that the MLE, in the case that the uncertainty on the \(X\) and \(Y\) observations is Gaussian distributed, is a specific instance of \((D(\|\cdot\|,\|\cdot\|_F)\)-SSNMF of [22]. Our models for the distribution of the observed entries of \(X\) and \(Y\) will assume that the mean is given by an exact factorization, \(E[X] = AS\) and \(E[Y] = BS\), and the uncertainty in each set of observations is governed by a Gaussian distribution. That is, we consider the hierarchical models for \(X\) and \(Y\) in which

\[
X_{\gamma,\tau} = \sum_{i=1}^{r} x_{\gamma,i,\tau} \sim \mathcal{N} \left( \sum_{i=1}^{r} A_{\gamma,i} s_{i,\tau}, \sigma_1 \right),
\]

\[
Y_{\eta,\tau} = \sum_{i=1}^{r} y_{\eta,i,\tau} \sim \mathcal{N} \left( \sum_{i=1}^{r} B_{\eta,i} s_{i,\tau}, \sigma_2 \right).
\]

Here and throughout, \(\gamma\) and \(\eta\) are row indices of \(X\) and \(Y\) respectively, \(\tau\) is a column index of \(X\) and \(Y\), and \(i\) indexes the random variable summands which form \(X_{\gamma,\tau}\) and \(Y_{\eta,\tau}\). Note then that

\[
X_{\gamma,\tau} \sim \mathcal{N} \left( \sum_{i=1}^{r} A_{\gamma,i} S_{i,\tau}, r\sigma_1 \right),
\]

and

\[
Y_{\eta,\tau} \sim \mathcal{N} \left( \sum_{i=1}^{r} B_{\eta,i} S_{i,\tau}, r\sigma_2 \right).
\]

due to the summable property of Gaussian random variables. We note that this assumes different Gaussian models of uncertainty on the two collections of rows of the NMF [4].

Assuming that the set of \(X_{\gamma,\tau}\) and \(Y_{\eta,\tau}\) are statistically independent conditional on \(A\), \(B\), and \(S\), we have that the likelihood \(p(X, Y|A, B, S)\) is

\[
\prod_{\gamma,\tau} \mathcal{N} \left( X_{\gamma,\tau} | \sum_{i=1}^{r} A_{\gamma,i} S_{i,\tau}, r\sigma_1 \right) \prod_{\eta,\tau} \mathcal{N} \left( Y_{\eta,\tau} | \sum_{i=1}^{r} B_{\eta,i} S_{i,\tau}, r\sigma_2 \right).
\]

(11)

We apply the monotonic natural logarithmic function to the likelihood, and ignore terms that do not vary with the factor matrices. This transforms the likelihood function into a \((\|\cdot\|_F,\|\cdot\|)\)-SSNMF objective, while preserving the maximizer. That is, the log likelihood (excluding additive terms which are constant with respect to \(A\), \(B\), and \(S\) is

\[
\ln p(X, Y|A, B, S) = -\frac{1}{2\sigma_1} \sum_{\gamma,\tau} \left( X_{\gamma,\tau} - \sum_{i=1}^{r} A_{\gamma,i} S_{i,\tau} \right)^2 - \frac{1}{2\sigma_2} \sum_{\eta,\tau} \left( Y_{\eta,\tau} - \sum_{i=1}^{r} B_{\eta,i} S_{i,\tau} \right)^2 - \frac{\lambda}{2\sigma_2} \sum_{\eta,\tau} \left( Y_{\eta,\tau} - \sum_{i=1}^{r} B_{\eta,i} S_{i,\tau} \right)^2.
\]

Thus, the maximum likelihood estimators for \(A\), \(B\), and \(S\) are given by

\[
\arg \min_{A,B,S} ||X - AS||^2_F + \frac{\sigma_1}{\sigma_2} ||Y - BS||^2_F.
\]

We see that the MLE in the case of Gaussian uncertainty on both sets of observations, \(X\) and \(Y\), is a specific instance of \((\|\cdot\|_F,\|\cdot\|)\)-SSNMF objective where the regularization parameter \(\lambda\), which defines the relative weighting of the supervision term, is given as a ratio of the variances of the distributions.

Next, we demonstrate that the MLE, in the case that the uncertainty on \(X\) is Poisson distributed and \(Y\) is Gaussian distributed, is a specific instance of the \((D(\|\cdot\|_2), \|\cdot\|_{2})\)-SSNMF model. This MLE derivation follows from that of 2 by swapping the roles of \(X\) and \(Y\), and rescaling the resulting log likelihood; however, we include a sketch of the derivation to be thorough.

Again, our models for observed \(X\) and \(Y\) assume that the mean is given by an exact factorization, \(E[X] = AS\) and \(E[Y] = BS\), with the uncertainty in \(X\) governed by a Poisson distribution and the uncertainty in \(Y\) governed by a Gaussian distribution. That is, we consider the hierarchical models for \(X\) and \(Y\) in which

\[
X_{\gamma,\tau} = \sum_{i=1}^{r} x_{\gamma,i,\tau} \sim \mathcal{P}(A_{\gamma,i} S_{i,\tau}),
\]

\[
Y_{\eta,\tau} = \sum_{i=1}^{r} y_{\eta,i,\tau} \sim \mathcal{N} \left( \sum_{i=1}^{r} B_{\eta,i} S_{i,\tau}, \sigma_2 \right).
\]

Note then that

\[
X_{\gamma,\tau} \sim \mathcal{P} \left( \sum_{i=1}^{r} A_{\gamma,i} S_{i,\tau} \right),
\]

and

\[
Y_{\eta,\tau} \sim \mathcal{N} \left( \sum_{i=1}^{r} B_{\eta,i} S_{i,\tau}, r\sigma_2 \right).
\]

due to the summable property of Gaussian and Poisson random variables. We note this assumes a Poisson and Gaussian model of uncertainty on the two collections of rows of the NMF [3].

Then proceeding as in (11) and (10) and assuming that the set of \(X_{\gamma,\tau}\) and \(Y_{\eta,\tau}\) are statistically independent conditional on \(A\), \(B\), and \(S\), we have that
the log likelihood (excluding additive terms which are constant with respect to \(A, B,\) and \(S\)) is

\[
\ln p(X, Y | A, B, S) = ^+ \left[ D(X \| AS) + \frac{1}{2r\sigma_2} \| Y - BS \|^2_F \right].
\]

Thus, the maximum likelihood estimators for \(A, B,\) and \(S\) are given by

\[
\arg \min_{A, B, S \geq 0} D(X \| AS) + \frac{1}{2r\sigma_2} \| Y - BS \|^2_F.
\]

We see that the MLE in the case of Poisson uncertainty on the observations in \(X\) and Gaussian uncertainty on the observations in \(Y\) is a specific instance of the \((D(\cdot \| \cdot), \| \cdot \|_F)\)-SSNMF objective where the regularization parameter \(\lambda\) is the inverse of a multiple of the variance of the Gaussian distribution.

Finally, we demonstrate that the MLE, in the case that the uncertainty on \(X\) and \(Y\) are Poisson distributed, is a specific instance of the \((D(\cdot \| \cdot), D(\cdot \| \cdot))\)-SSNMF model. This result follows from \([5, 12, 41]\); we sketch the derivation to be thorough.

Again, we assume that the distributions of the observed \(X\) and \(Y\) have means given by an exact factorization, \(E[X] = AS\) and \(E[Y] = BS\), with the uncertainty in both governed by a Poisson distribution. That is, we consider the hierarchical models for \(X\) and \(Y\) in which

\[
X_{\gamma,\tau} = \sum_{i=1}^r x_{\gamma,i,\tau} \quad \text{and} \quad x_{\gamma,i,\tau} \sim \mathcal{P}(x_{\gamma,i,\tau} | \eta_{\gamma,i,\tau} \), \]

\[
Y_{\eta,\tau} = \sum_{i=1}^r y_{\eta,i,\tau} \quad \text{and} \quad y_{\eta,i,\tau} \sim \mathcal{P}(y_{\eta,i,\tau} | \gamma_{\eta,i,\tau} \).
\]

Note then that

\[
X_{\gamma,\tau} \sim \mathcal{P}(X_{\gamma,\tau} | \sum_{i=1}^r A_{\gamma,i} S_{i,\tau}), \quad \text{and} \quad \]

\[
Y_{\eta,\tau} \sim \mathcal{P}(Y_{\eta,\tau} | \sum_{i=1}^r B_{\eta,i} S_{i,\tau})
\]

due to the summable property of Poisson random variables. We note that assumes different Poisson models of uncertainty on the two collections of rows of the NMF \([4]\).

Then proceeding as in \([11]\) and \([6]\) and assuming that the set of \(X_{\gamma,\tau}\) and \(Y_{\eta,\tau}\) are statistically independent conditional on \(A, B,\) and \(S\), we have that the log likelihood (excluding additive terms which are constant with respect to \(A, B,\) and \(S\)) is

\[
\ln p(X, Y | A, B, S) = ^+ \left[ D(X \| AS) + D(Y \| BS) \right].
\]

Thus, the maximum likelihood estimators for \(A, B,\) and \(S\) are given by

\[
\arg \min_{A, B, S \geq 0} D(X \| AS) + D(Y \| BS).
\]

We see that the MLE in the case of Poisson uncertainty on the observations in \(X\) and \(Y\) is a specific instance of the \((D(\cdot \| \cdot), D(\cdot \| \cdot))\)-SSNMF objective where the regularization parameter is \(\lambda = 1\).

**B Multiplicative Updates Derivations**

Here we provide intuition for the derivation of multiplicative updates for \((D(\cdot \| \cdot), \| \cdot \|_F)\)-SSNMF and \((D(\cdot \| \cdot), D(\cdot \| \cdot))\)-SSNMF (Algorithms \([2, 3]\)).

First, note that the multiplicative updates for \((D(\cdot \| \cdot), \| \cdot \|_F)\)-SSNMF (Algorithm \([2]\)) follow from those for \((\| \cdot \|_F, D(\cdot \| \cdot))\)-SSNMF (Algorithm \([1]\)) by swapping the roles of \(X\) and \(Y\), and \(A\) and \(B\).

Next, the multiplicative updates for \((D(\cdot \| \cdot), D(\cdot \| \cdot))\)-SSNMF (Algorithm \([3]\)) are derived as follows. The gradients of \(F_4(A, B, S; X, Y)\) with respect to \(A, B,\) and \(S\) are, respectively

\[
\nabla_A F_4 = WS^T - \left[ \frac{W \circ X}{W \circ AS} \right] S^T,
\]

\[
\nabla_B F_4 = LS^T - \left[ \frac{L \circ Y}{L \circ BS} \right] L^T, \quad \text{and}
\]

\[
\nabla_S F_4 = -A^T \left[ \frac{W \circ X}{W \circ AS} \right] \circ W + A^T W
\]

\[
- \lambda B^T \left[ \frac{L \circ Y}{L \circ BS} \right] \circ L + \lambda B^T L.
\]

The multiplicative updates of Algorithm \([3]\) are given by

\[
A \leftarrow A - \Gamma \circ \nabla_A F_4 \quad \text{when} \quad \Gamma = \frac{A}{WS^T},
\]

\[
B \leftarrow B - \Gamma \circ \nabla_B F_4 \quad \text{when} \quad \Gamma = \frac{B}{LS^T}, \quad \text{and}
\]

\[
S \leftarrow S - \Gamma \circ \nabla_S F_4 \quad \text{when} \quad \Gamma = \frac{S}{A^T W + \lambda B^T L}.
\]

**C Additional 20 Newsgroups Results**

In this section, we include additional analysis and results for the 20 Newsgroups data set. First, we summarize in Table \([6]\) the hyperparameters used for the methods described in Section \([3.2]\). We select the hyperparameters that result in the highest average classification accuracy of the validation set. For the SSNMF models, we search over
tol \in \{10^{-4}, 10^{-3}, 10^{-2}\}, and \lambda \in \{10, 10^2, 10^3\}, and for the NMF model, we search over tol \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}.

| Model          | Hyperparameters |
|----------------|-----------------|
| (|| \cdot ||_F, || \cdot ||_F) | tol = 10^{-4}, \lambda = 10^2 |
| (|| \cdot ||_F, D(|| \cdot ||)) | tol = 10^{-4}, \lambda = 10^2 |
| (D(|| \cdot ||), || \cdot ||_F) | tol = 10^{-3}, \lambda = 10^3 |
| (D(|| \cdot ||), D(|| \cdot ||)) | tol = 10^{-3}, \lambda = 10^3 |
| || \cdot ||_{F-NMF} | tol = 10^{-4} |

Table 6: Hyperparameter selection for NMF and SSNMF models by selecting the hyperparameters that result with the highest average classification accuracy of the validation set (over 10 trials).

As in Section 3.2, we consider the “typical” (achieving median accuracy within trials) decomposition for NMF, and remaining SSNMF models. We display in Figure 3 the B_{train} matrices for each of the median accuracy SSNMF decompositions, and in Figure 4, the coefficients matrix of the SVM classifier for the median accuracy NMF decomposition. Further, we report in Tables 7, 9, 11 and 12, the top 10 keywords representing each topic for each of the models.

For the (|| \cdot ||_F, D(|| \cdot ||))-SSNMF model, we (qualitatively) observe from Table 7 that topics 4, 5, 7 and 10 are overlapping topics associated to the class Computers; see Figure 3a. Similarly, topics 1 and 9 that capture the subjects of cryptography, electronics, and space, have various overlapping keywords (“space”, “key”, “chip”), and are associated with the class Sciences. On the other hand, topics associated to class Politics and Religion are less overlapping. Lastly, topics 3 and 8 are recreation topics (“game”, “team”, “car”) relating to autos where in addition topic 3 (“hockey”, “player”, “nhl”) is specific to hockey and topic 8 (“baseball”) is specific to baseball.

For the (D(|| \cdot ||), D(|| \cdot ||))-SSNMF model, we (qualitatively) observe from Table 9 that topic 5 (“space”, “moon”, “time”), topic 7 (“key”, “chip”, “clipper”), and topic 11 (“data”, “government”, “buy”) are associated with class Sciences. Further, we (qualitatively) observe from Table 11 that topic 4 (“church”, “belief”, “bible”), and topic 6 (“atheists”, “paul”) are both related to religion (“god”, “jesus”) and are associated to class Religion; see Figure 3c. Lastly, topic 12 (“car”, “hockey”, “players”, “baseball”) is a recreation topic that captures autos, hockey, and baseball subjects, whereas topics 2 and 9 are broad and not as informative.

For the (D(|| \cdot ||), D(|| \cdot ||))-SSNMF model, we observe from Figure 3c that topic 5 (“space”, “moon”, “time”), topic 7 (“key”, “chip”, “clipper”), and topic 11 (“data”, “government”, “buy”) are associated with class Sciences. Further, we (qualitatively) observe from Table 11 that topic 4 (“church”, “belief”, “bible”), and topic 6 (“atheists”, “paul”) are both related to religion (“god”, “jesus”) and are associated to class Religion; see Figure 3c. Lastly, topic 12 (“car”, “hockey”, “players”, “baseball”) is a recreation topic that captures autos, hockey, and baseball subjects, whereas topics 2 and 9 are broad and not as informative.
topic 11 (“god”, “jesus”, “christ”, “faith”) relates to religion and specifically Christianity. Lastly, we observe in Figure 4 that topic 13 is shared across 3 classes (Computers, Sciences, and Religion), and is not as informative as the other topics.

Figure 4: The normalized coefficients matrix of the SVM classifier (with NMF-$\|\cdot\|_F$) corresponding to the median test classification accuracy equal to 71.67. Here, all negative coefficients are thresholded to 0, and then each column is normalized to showcase the distribution of the topic over classes.

Clustering Analysis

In this section, we measure the performance of the NMF and SSNMF topic models in a clustering-motivated score. In these experiments, we measure the similarity of ground-truth clusters, encoded by a given label matrix $M$, to NMF/SSNMF computed clusters, encoded by the SSNMF/NMF representation matrix $S$. We denote by $M$ the (column-wise) one-hot encoded label matrix which maps documents to the subgroups to which they belong $\left( M \in \{0, 1\}^{13 \times 8980} \right)$ (subgroups by documents). We denote by $S$ be the representation matrix computed by NMF/SSNMF, in which the $i$th row provides the association of each document with the $i$th topic.

We employ two approaches to clustering or mixture assignment. The first is hard clustering in which the documents are assigned to a single cluster corresponding to computed topics. In this approach, we apply a mask to the representation matrix, $\hat{S} = \text{label}(S)$, where label($\cdot$) assigns the largest entry of each column to 1 and all other entries to 0. The second approach is soft clustering in which the documents are assigned to a distribution of clusters corresponding to the topics. In this approach, we normalize each of the columns of the representation matrix to produce $\hat{S}$.

Now, in either approach, we apply a metric $P$ which measures the association between the $\ell$th topic-documents association $\hat{S}_\ell$ and the best ground truth subgroup-documents association, $M_I$ (here $M_I$ is the $I$th row of $M$); that is, for topic $\ell$, we define $I$ as

$$I = \arg\max_i \frac{\|\hat{S}_\ell \odot M_i\|_1}{\|M_i\|_1},$$

and define score $P$ for the $\ell$th topic as

$$P(\hat{S}_\ell) = \frac{\|\hat{S}_\ell \odot M_I\|_1}{\|M_I\|_1},$$

where $\| \cdot \|_1$ denotes the $\ell_1$-norm. We note that this metric is similar to that of [44]; we use score $P$ instead as it allows us to measure clustering performance topic-wise. We also note that the learned topics of NMF and SSNMF methods need not be in one-to-one correspondence with the subgroups in Table 3 as topics are also learnt for the classification task at hand.

Now, we present in Table 8 the average (averaged over topics) score $P$ for the representation matrices computed by each of the NMF/SSNMF models in both the hard-clustering and soft-clustering settings. The scores $P$ for each topic (for both hard-clustering and soft-clustering) and the maximizing subgroup (indicated by $I$) are listed in the bottom four rows of the keyword table associated to each model; see the last four rows of Tables 7, 9, 10, 11, and 12.

---

3In the 20 Newsgroups data set, each document belongs to only one subgroup.
Table 7: Top keywords representing each topic of the \( (\parallel \cdot \parallel_F, \parallel \cdot \parallel_F) \)-SSNMF model referred to in Figure 3a.

| Keywords | 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 |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|-----------|-----------|-----------|-----------|
| world    | people  | game    | use     | x       | would   | x       | game    | would   | please  | god       | people    | one       | god       |
| like     | israel  | team    | thanks  | c       | Jews    | get     | like    | one     | church  | people    | would     | would     | would     |
| space    | gun     | car     | x       | know    | time    | need    | games   | space   | anyone  | would     | would     | one       | people    |
| one      | one     | year    | using   | r       | israel  | image   | games   | anyone  | would   | killed    | jesus     | one       | people    |
| chip     | hockey  | window  | window  | r       | image   | windows | baseball | like    | window  | balls     | people    | southerns | believe   |
| key      | would   | espn    | know    | like    | window  | law     | problem | windows | think   | know      | people    | christ    | christ    |
| use      | armenian| graphics| widget  | law     | problem | windows | think   | know    | week    | like       | jews      | religion  | religion  |
| good     | government| players| system  | pe      | program| government| right   | mac     | version | like      | like       | well      | bible     |
| phone    | said    | nhl     | program| please  | motif   | system  | pe       | right   | get     | help      | apple      | christian |
| edu      | turkish | games   | anyone  | verseion | get    | system  | right   | government| get    | help      | say        | faith    | bible     |
| Keywords | 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 |
| hard     | electronics | mideast | hockey | windows | guns | windows | baseball | crypto | windows | guns | christian | mideast | christian |
| score    | 0.2060 | 0.6594 | 0.3411 | 0.1471 | 0.1466 | 0.5698 | 0.7500 | 0.1441 | 0.5226 | 0.1625 | 0.4574 |
| soft     | space   | nudeast | hockey | graphs | windows | guns | windows | baseball | crypto | graphs | christian | mideast | christian |
| score    | 0.3135 | 0.4560 | 0.4222 | 0.2389 | 0.1914 | 0.2278 | 0.3064 | 0.5933 | 0.2128 | 0.4983 | 0.2490 | 0.4525 |

Table 8: Listed scores are average over 11 trials; in each trial, we average score \( P \) across all topics.
