Mach Learn (2016) 104:441–466
DOI 10.1007/s10994-015-5537-5

Sparse topical analysis of dyadic data using matrix tri-factorization

Ranganath Biligere Narayana Swamy[1]

Received: 14 December 2014 / Accepted: 30 September 2015 / Published online: 17 December 2015

Abstract Many applications involve dyadic data, where associations between one pair of domain entities, such as documents, words and associations between another pair, such as documents, users are completely observed. We motivate the analysis of such dyadic data introducing an additional discrete dimension, which we call topics, and explore sparse relationships between the domain entities and the topic, such as user-topic and document-topic relationships. For this problem of sparse topical analysis of dyadic data, we propose a formulation using sparse matrix tri-factorization. This formulation requires sparsity constraints, not only on the individual factor matrices, but also on the product of two of the factors. To the best of our knowledge, this problem of sparse matrix tri-factorization has not been studied before. We propose a solution that introduces a surrogate for the product of factors and enforce sparsity on this surrogate as well as on the individual factors through L1-regularization. The resulting optimization problem is efficiently solvable in an alternating minimization framework over sub-problems involving individual factors using the well known FISTA algorithm. For the sub-problems that are constrained, we use a projected variant of the FISTA algorithm. We also show that our formulation leads to independent sub-problems towards solving a factor matrix, thereby supporting parallel implementation leading to scalable solution. We perform experiments over bibliographic and product review data to show that the proposed framework based on sparse tri-factorization formulation results in better generalization ability and factorization accuracy compared to baselines that use sparse bi-factorization.

Keywords Non-negative matrix factorization · Matrix tri-factorization · Sparsity regularization

Editors: Concha Bielza, Joao Gama, Alipio Jorge, and Indrė Žliobaitė.

✉ Ranganath Biligere Narayana Swamy
ranga.nitk@gmail.com; ranganath@csa.iisc.ernet.in

[1] CSA Department, Indian Institute of Science, Bangalore, India

1 Introduction

Analysis of tagged data, where each data item is labeled with tags from a finite but very large set of tags, is emerging as a problem of immense importance in various industries. A common and familiar example is tagging of data items with users associated with them. An illustration of the significance of such analysis is the Netflix competition,[1] where the goal is to predict viewers ratings for the movies given the previous ratings by the viewers without any additional information about the viewers or the movies. In addition to explicit consumer reviews that a company has at its disposal, user comments posted on social media sites are turning out to be valuable resources for analyzing consumer preferences.

It is common in enterprise scenarios to associate tags (such as product names, business units, strategies, relevant industries, etc.) with documents, web and wiki pages, which are used as aids for enterprise search and various other analytics. Another similar example is hashtags for Twitter assigned by users.

For such dyadic data, we will assume that we have three central entities, which we will generically refer to as users (e.g., consumers, viewers, researchers), documents (e.g., movie scripts, product literature, research papers) and a vocabulary of words. One of the dyadic associations is between documents and words and the other, tagging of users with documents.

Although we focus on documents and words, the formulations and algorithms that we propose are equally applicable for discrete-valued dyadic associations between other kinds of entities, such as products and their features, movies and users etc., which we demonstrate experimentally.

Given such dyadic data, the task of interest is topical analysis over words, documents and users with an unobserved dimension, generically called topics. For example, genres correspond to topics in movies, while research and technology areas are examples of topics for enterprise documents. Knowledge of such topics can then be used for various purposes, such as recommendations, designing new products, etc. While topical analysis has been investigated extensively for dyadic data (Rosen-Zvi et al. 2004; Zheng et al. 2011), one important property of topical associations that has largely been overlooked is that of sparsity. While the actual number of topics may be large, the association between topics and documents, and similarly that between topics and users, is typically sparse. For example, most individual consumers prefer a small set of genres or product categories, and individual movies or products correspond to very few genres or categories. Similarly in the enterprise setting, most documents are associated with small sets of research areas, and any specific researcher usually has expertise on a limited set of research areas.

There is not much prior work that performs sparse topical analysis over dyadic data. Learning the probabilistic author-topic model (Rosen-Zvi et al. 2004) can be interpreted as estimating the three associations of interest to us i.e, the distribution over words for each topic, distribution over topics for each document and distribution over topics for each author. While this model was proposed to accommodate the knowledge of actual authors of documents, it does not directly address any notion of sparsity.

In this paper, we show that this problem can be formulated as a sparse matrix trifactorization problem. To handle user-document associations, this formulation imposes a support constraint on one of the factors. Additionally, it imposes sparsity constraints on the individual factor matrices, and also on the product of two of the factors. Instead of directly incorporating a sparsity constraint on the product, we introduce a surrogate matrix on which we enforce the sparsity constraint, while making this term as close as possible to the original

[1] http://www.netflixprize.com/.

product of factors. We show that in this formulation the sub-problems for the individual factor matrices are constrained least squares problems. The least squares problem is efficiently solvable using the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck and Teboulle 2009). We use a projected variant of FISTA to handle the non-negativity constraints in all of the sub-problems. The overall non-convex problem is solved efficiently using the alternating minimization framework, using the projected variant of FISTA for the sub-problems in each iteration.

There has been significant interest in the matrix factorization problem over the last decade or so (Lee and Seung 1999; Paatero and Tapper 1994). Non-negativity and sparseness have been studied in the context of bi-factorization (Lee and Seung 1999; Hoyer 2004; Kim and Park 2007). In the context of tri-factorization, Block Value Decomposition has been proposed for non-negativity constraints (Long et al. 2005). Other constraints studied in the context of tri-factorization include orthonormality (Ding et al. 2006). Sachan and Srivastava (2013) address the problem of sparse topical analysis of dyadic data using a coupled sparse bi-factorization approach but do not estimate the strength of user-document associations. To the best of our knowledge, the sparsity and support constraints that we consider in the context of tri-factorization have not been investigated before.

We perform experiments over bibliographic and product review data to demonstrate that the proposed sparse matrix tri-factorization formulation results in better generalization ability and factorization accuracy compared to baselines that use sparse bi-factorization.

Our main contributions are as follows.

(A) We formulate the problem of sparse topical analysis of dyadic data, as that of matrix tri-factorization with sparsity and support constraints. This has not been studied previously. (B) We propose an efficient solution for this problem using the alternating minimization framework, where the individual non-negatively constrained sparse least-squares sub-problems in each iteration are solved using the projected variant of FISTA algorithm. (C) We show that our formulation supports parallel implementation since the individual sub-problems for solving a factor matrix are independent. (D) We demonstrate that the sparse matrix tri-factorization formulation yields topical associations that are more accurate and generalize better than those resulting from state-of-the-art baselines for the problem.

The rest of the paper is organized as follows: Sect. 2 discusses the related work. In Sect. 3, we propose the sparse matrix tri-factorization formulation for sparse topical analysis of dyadic data. In Sect. 4, we propose an efficient solution for this formulation using projected variant of FISTA in an alternating minimization framework. Section 5 deals with experimental work. We conclude our work in Sect. 6.

2 Related work

In this section, we first review existing literature from the point-of-view of sparse topical analysis on dyadic data, and then discuss related work on matrix factorization.

2.1 Sparse topical analysis

Collective non-negative matrix factorization (NMF) with constraints (Sachan and Srivastava 2013) considers a problem similar to ours, where authorship information is provided in addition to the documents. Here, low rank approximations are sought for both the matrices, and in addition to the sparsity requirements on the individual factors, two factor products are also required to be sparse. However, this is approximated by minimizing the Frobenius

norms of the three factors. The optimization is performed using a dual ascent approach, where primal variables are exactly optimized in each iteration using gradient descent, and partial updates are made on the dual variables in the direction of the positive gradients. However, this model factorizes a binary author-document association matrix and may not result in accurate estimation of factor matrices and also does not estimate the strength of author-document associations.

In the probabilistic setting, a problem similar to non-negative matrix tri-factorization is addressed by the Author-Topic Model (Rosen-Zvi et al. 2004) in the context of bibliographic data. Here, the distribution over words for each topic, distribution over topics for each document and distribution over topics for each author explicitly capture three of the factors of interest. The fourth factor, which is the distribution over authors for each document, is not directly accounted for, but can be incorporated without much difficulty. The notion of sparsity, on the other hand, is not addressed by this model. Sparsity has however been extensively studied in recent years for probabilistic admixture models, mainly by replacing the Dirichlet distributions with those that promote sparsity, such as the Indian Buffet Process (Griffiths and Ghahramani 2005; Williamsonm et al. 2010) in the context of bi-factorization problems, such as LDA (Blei et al. 2003).

In the context of tri-factorization, sparsity has been introduced via hierarchical beta process in contextual focussed topic model (Chen et al. 2012), but words are required to belong to either a topic distribution corresponding to a publication venue, or an author, or else a document and the inference procedure employed in the model is complicated, hence intractable on large datasets. Further, unlike our model, it does not support parallelism.

There have been few other works on topic modeling in the context of bi-factorization considering the geometrical structure of the data (Cai et al. 2009) by locally consistent topic model (LTM) as well as double-latent-layered LDA (D-LDA) (Zhuang et al. 2010) for semi-defined classification of documents. LTM uses the local manifold structure of the data to regularize the learning of probability distributions so that two sufficiently close documents have similar conditional probability distributions $P(z/d)$ where z is the topic vector for the words in document d, by making use of Kullback–Leibler divergence to measure the distance between two conditional probability distributions.

Compared to LDA, D-LDA uses another latent variable y for documents in addition to topic variable z for words. It makes use of both supervised classification and unsupervised clustering to classify an unlabeled document into one of the known or unknown classes and finally cluster the documents belonging to the unknown class to form meaningful groups. The topic mixtures are tied to the class variable y so that topics for words z help in inferring the right label for the documents. However, both LTM and D-LDA do not take into account, the association between authors and documents.

2.2 Matrix factorization

The area of matrix bi-factorization with constraints has seen a lot of research over the years. Non-negative matrix factorization (NMF) (Lee and Seung 1999) imposes non-negative constraints on the two factors. Various methods have been proposed to solve NMF including projected gradient and projected quasi-Newton techniques (Lin 2007; Kim et al. 2007), and the active set method (Kim and Park 2008a). Though NMF is often found to generate sparse factors, approaches have been proposed to directly control the sparsity of the two factors (Hoyer 2004; Kim and Park 2007). The Factorization machine (Rendle 2010) is similar to a polynomial SVM (polynomial kernel) except that the parameter matrix for the interaction between variables in the model is factorized here and therefore the parameters learnt

under this model are not independent resulting in the high quality parameter estimates under sparsity. This model can be used for regression, binary classification and ranking of the vectors. But this is a supervised model since it needs training examples with labels to learn the model parameters. But our problem needs an unsupervised approach wherein we evaluate the clustering quality of document-topic and user-topic associations during the learning phase itself.

In the context of tri-factorization, the Block Value Decomposition (Long et al. 2005) approach extends NMF by minimizing decomposition error using Frobenius norm, while enforcing non-negative constraints on the three factors. The resulting optimization problem can be solved in the alternating non-negative least squares framework using multiplicative update rules.

In the orthogonal tri-factorization formulation (Ding et al. 2006), orthogonality constraints are introduced among the left and right factors, in addition to the non-negativity constraints, so that the low-dimensional embedding has a natural clustering interpretation. The optimization is again done using multiplicative update rules. To the best of our knowledge, we are not aware of any tri-factorization formulations considering sparsity and support constraints.

In summary, matrix factorization techniques for dyadic data generally approximate the data matrix as a product of the factor matrices, with finding unnormalized representation of the topical associations for entities, and without incorporation of any prior knowledge on the topical associations. But, dictionary learning approaches for signal and image processing problems enforce unit l_2-norm constraints on the columns of the left factor matrix (basis) (Mairal et al. 2010) so that one of the matrices do not become too large and the other too small. This basis is analogous to the word-topic matrix in our problem. However, recent papers (Kim and Park 2008b; Kasiviswanathan et al. 2011) employing NMF for topic modeling do not enforce unit l_2-norm on any of the word-topic or topic-document matrices.

On the other hand, probabilistic approaches (Rosen-Zvi et al. 2004; Chen et al. 2012) find the normalized representation of topical associations for the entities as distributions. These distributions are learnt through different criteria for maximizing the likelihood of the model. Additionally, they have the flexibility of incorporating prior knowledge on the topical associations for the entities.

3 Sparse matrix tri-factorization: formulation

We first formalize the notion of dyadic data. Data on m users, n documents and v terms can be captured using two dyadic matrices. The first is a user-document dyadic matrix $A \in \{0, 1\}^{m \times n}$, where $A_{ui} = 1$ indicates that user u *is* associated with document i, while $A_{ui} = 0$ indicates with certainty that user u *cannot be* associated with document i. The second is a term-document dyadic matrix $D \in \mathbb{R}_+^{v \times n}$, where D_{ij} denotes the number of times term i appears in document j.

Given such data, and assuming k topics, we need to approximate D as a product of three factor matrices: $D \approx \Phi\Theta\mathbb{A}$. The first factor here is $\Phi \in \mathbb{R}_+^{v \times k}$ with Φ_{it} denoting the association between word i and topic t. The second factor is $\Theta \in \mathbb{R}_+^{k \times m}$ with Θ_{tu} denoting the preference for topic t of user u. The third factor is $\mathbb{A} \in \mathbb{R}_+^{m \times n}$ with $\mathbb{A}_{ui}$ denoting the extent of association between user u and document i.

Let us now come to the constraints on the factors required for our problem. The first set is of the natural non-negativity constraints: $\Phi \geq 0, \Theta \geq 0, \mathbb{A} \geq 0$. To understand the second constraint, it is important to appreciate the difference between the two user-document

matrices A and $\mathbb{A}$. A_{ij} is binary-valued, indicating whether or not a specific user is associated with a document. In contrast, $\mathbb{A}_{ij}$ is non-negative real-valued, denoting the strength of the association. The first indicates whether or not a user is associated with a document, while the second indicates the strength of association between the user and the document. Clearly, this leads to the constraint that $\mathbb{A}_{ij}$ can be non-zero only when A_{ij} is 1. We denote this as $supp(\mathbb{A}) \subseteq supp(A)$.

Let us now examine the sparsity requirements on the factor matrices. First, the columns of the topic-user matrix Θ are required to be sparse. The second sparsity requirement is on the topic-document associations. This is captured by the product $\Theta\mathbb{A}$ of the individual factors Θ and $\mathbb{A}$. Therefore, we require the columns of the product $\Theta\mathbb{A}$ to be sparse. We enforce sparsity on the column vectors using the vector l_1-norm (Tibshirani 1996).

Finally, using the Frobenius norm to capture the approximation error between D and $\Phi\Theta\mathbb{A}$ and l_1 regularizers for the sparsity constraints, results in the following tri-factorization problem:

$$\arg\min_{\Phi,\Theta,\mathbb{A}} \frac{1}{2}\|D - \Phi\Theta\mathbb{A}\|_F^2 + \lambda_1 \sum_{j=1}^{n} \|(\Theta\mathbb{A})_j\|_1 + \lambda_2 \sum_{j=1}^{m} \|\Theta_j\|_1 + \lambda_3 \sum_{j=1}^{t} \|\Phi_j\|_1$$
$$\text{s.t. } supp(\mathbb{A}) \subseteq supp(A), \quad \Phi \geq 0, \quad \Theta \geq 0, \quad \mathbb{A} \geq 0, \tag{1}$$

where we have used the notation M_i to denote the ith column of matrix M and this applies to the rest of the paper. Note that we have additionally enforced sparsity on Φ to prevent overfitting. λ_1, λ_2 and λ_3 denote the regularization constants for the three sparsity constraints.

Typically, the alternating minimization framework is used to solve matrix factorization problems (Lin 2007; Kim and Park 2007, 2008a), where the sub-problems in the individual factor variables are convex with tractable solvers available. However, the sub-problem of (1) involving Θ looks as follows:

$$\arg\min_{\Theta} \frac{1}{2}\|D - \Phi\Theta\mathbb{A}\|_F^2 + \lambda_1 \sum_{j=1}^{n} \|(\Theta\mathbb{A})_j\|_1 + \lambda_2 \sum_{j=1}^{m} \|\Theta_j\|_1,$$
$$\text{s.t. } \Theta \geq 0, \tag{2}$$

Though the optimization problem in (2) is convex, it does not admit an efficient iterative solution leading to huge increase in time as discussed in experimental section. Regarding convergence, the local convergence of (1) is not guaranteed as discussed in the "Appendix".

To get around this, we reformulate our problem, with the motivation of making the individual sub-problems efficiently solvable using (variants of) algorithms such as FISTA (Beck and Teboulle 2009). We introduce a surrogate variable Q to capture the topic-document associations. We then enforce sparsity on the columns of Q using the l_1-norm, and simultaneously enforce Q to be close to the product $\Theta\mathbb{A}$ by minimizing the Frobenius norm of $Q - \Theta\mathbb{A}$. This results in the following modified formulation of the sparse matrix tri-factorization (SMTF) problem:

$$\arg\min_{\Phi,\Theta,Q,\mathbb{A}} \frac{1}{2}\|D - \Phi Q\|_F^2 + \frac{1}{2}\|Q - \Theta\mathbb{A}\|_F^2 + \lambda_Q \sum_{j=1}^{n} \|Q_j\|_1 + \lambda_\Theta \sum_{j=1}^{m} \|\Theta_j\|_1$$
$$+ \lambda_\Phi \sum_{j=1}^{t} \|\Phi_j\|_1$$
$$s.t. supp(\mathbb{A}) \subseteq supp(A), \quad \Phi \geq 0, \quad \Theta \geq 0, \quad Q \geq 0, \quad \mathbb{A} \geq 0, \tag{3}$$

Each of the four sub-problems of (3) now admit an efficient solution using (a projected version of) FISTA where λ_Q, λ_Θ and λ_Φ denote the regularization constants for the three sparsity constraints. Note that using a surrogate for the product of the factors $\Phi\Theta$ results in enforcing l_1-norm on $\mathbb{A}$ in the sub-problem involving $\mathbb{A}$ which is not desirable, the details of which are given in "Appendix" and also this can happen in (1). We elaborate more in the next section, where we propose an alternating minimization algorithm for the SMTF problem.

4 Sparse matrix tri-factorization: optimization

We now propose an alternating minimization algorithm for solving the sparse matrix tri-factorization formulation in (3). Recall that the optimization variables are the word-topic matrix Φ, the surrogate topic-document matrix Q, the topic-user matrix Θ and the user-document matrix $\mathbb{A}$. In this framework, starting with a suitable initialization of the four variables, we optimize over each of them in an alternating fashion, holding the others fixed as their current values. Local convergence for SMTF formulation (3) using alternating minimization framework is discussed in Sect. 4.2.

As we will show, the most general form for the individual sub-problems is that of l_1-regularized non-negative least squares (NNLS) (Kim and Park 2008b). Before moving on to the individual sub-problems, we first discuss some efficient algorithms for this general problem. As mentioned before, efficiency is critical, since a large number of instances of this problem are required to be solved in each iteration.

4.1 Efficiently solving sparse non-negative least squares

The constrained non-negative least squares problem comes up commonly in the context of matrix factorization and alternating minimization. The commonly used techniques for this are projected gradient (Lin 2007), projected subgradient (Sachan and Srivastava 2013), active set (Kim and Park 2008a) and Alternating Directions Method (ADM) (Yang and Zhang 2011; Kasiviswanathan et al. 2011). Projected gradient for non-negative constraints operates on the class of objective functions which are differentiable and therefore also applicable to the non-negative least squares problem. It projects the solution computed by gradient descent method onto the non-negative orthant at each iteration. The projected subgradient method is an extension of the projected gradient approach for non-differentiable functions. Hence, projected subgradient method can also be used to solve l_1-regularized non-negative least squares problems. In theory, the projected subgradient algorithm converges asymptotically to the minima for a convex problem.

The block principal pivoting (BPP) algorithm (Kim and Park 2008b) is an extension of the active set method, and can be used in conjunction with L_1 regularizer on the right factor matrix, but not on the left factor matrix, which is the case for two of our sub-problems. Additionally, the BPP algorithm requires the coefficient matrix in the least squares problem to be full column rank. This condition is not guaranteed to be satisfied in the various sub-problems, and empirically we have found it to be violated frequently. Empirically, we found that ADM converges very slowly or may not converge to the true solution if the columns of the coefficient matrix are not normalized to 1 and this column normalization does not suit the alternating minimization framework.

FISTA has recently been shown to have a fast convergence rate for non-smooth convex problems (Beck and Teboulle 2009). FISTA minimizes cost functions of the form $f(x)+g(x)$, where f is convex, smooth and its gradient is Lipschitz, and g is convex and continuous.

FISTA finds the quadratic approximation of $F(x) = f(x) + g(x)$ at a given point y through the form

$$Q_L(x, y) := f(y) + \langle x - y, \nabla f(y) \rangle + \frac{L}{2}\|x - y\|^2 + g(x).$$

It is shown that the minimizer of $Q_L(x, y)$ over x is unique, and takes the following form:

$$p_L(y) := \arg\min_x \left\{ g(x) + \frac{L}{2}\|x - \left(y - \frac{1}{L} \nabla f(y)\right)\|^2 \right\}.$$

FISTA can be used to solve the unconstrained least squares problem $\arg\min_h \frac{1}{2}\|c - Wh\|_F^2$ for column matrix h. This is because $f(h) = \arg\min_h \frac{1}{2}\|c - Wh\|_F^2$ is convex, smooth, and its gradient is Lipschitz. Further, $g(h) = 0$ is a convex, continuous function. FISTA is also effective when g is the l_1-norm on h i.e., $g(h) = \lambda\|h\|_1$. It has been proved that sequence of objective function values $F(x_k)$ where x_k is the solution computed by FISTA at kth iteration, converges to optimal function value $F(x^*)$ at a rate no worse than $O(\frac{1}{k^2})$.

FISTA with constant step size for the unconstrained l_1-regularized least squares problem $\arg\min_h \frac{1}{2}\|c - Wh\|_F^2 + \lambda\|h\|_1$ is described in Algorithm 1 where $eigs(W^T \times W, 1)$ is the largest eigen-value of $W^T \times W$, $soft(a, b) = sign(a) \times max(|a| - b, 0)$ (Kasiviswanathan et al. 2011) is the soft operator and x_c is the converged value of x_k. The disadvantage of FISTA with constant step size is that for large scale problems, the Lipschitz constant (largest eigen-value) may not be efficiently computable.

Algorithm 1 FISTA with constant step size

1: **Input:L=L(f)-A Lipschitz constant of** $\nabla f = eigs(W^T \times W, 1)$;
2: **Step 0.**Take $x_0 = h$, $y_1 = x_0$, $t_1 = 1$.
3: **Step k.**($k \geq 1$) Compute until convergence
4: $p_L(y_k) = \arg\min_x \{\lambda\|x\|_1 + \frac{L}{2}\|x - (y_k - \frac{1}{L} \times W^T(W \times y_k - c))\|^2\}$
5: $p_L(y_k) = soft(y_k - \frac{1}{L} \times W^T(W \times y_k - c), \frac{\lambda}{L})$
6: $x_k = p_L(y_k)$
7: $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
8: $y_{k+1} = x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1})$
9:
10:
11: **Output:**$h = x_c$

FISTA with backtracking overcomes the largest eigen-value computation step in Algorithm 1 by computing the Lipschitz constant (L_k) for every kth step and has the same rate of convergence as that of FISTA with constant step size. FISTA with backtracking is described in Algorithm 2.

Note that the computational effort required in step 2 of Algorithm 2 depends on the values of L_0 and η. Smaller values of L_0 and η increase the computation time of step 2 and large values of η decrease the computation time in step 2 but increase the number of iterations needed to converge.

Projected FISTA for non-negativity constraints: Our sub-problems additionally involve a non-negativity constraint in the l_1-regularized least squares problem:

$$\arg\min_{h \geq 0} \frac{1}{2}\|c - Wh\|_F^2 + \lambda\|h\|_1. \tag{4}$$

Algorithm 2 FISTA with backtracking

Step 0. Take $L_0 > 0$, $\eta > 1$ and $x_0 \in R^n$. Set $y_1 = x_0$, $t_1 = 1$.
Step k.($k \geq 1$) Find the smallest non-negative integers i_k such that with $\bar{L} = \eta^{i_k} L_{k-1}$,
$F(p_{\bar{L}}(y_k)) \leq Q_{\bar{L}}(p_{\bar{L}}(y_k), y_k)$
Set $L_k = \eta^{i_k} L_{k-1}$ and Compute until convergence
$p_{L_k}(y_k) = \arg\min_x \{\lambda \|x\|_1 + \frac{L_k}{2} \|x - (y_k - \frac{1}{L_k} \times W^T (W \times y_k - c))\|^2\}$
$p_{L_k}(y_k) = soft(y_k - \frac{1}{L_k} \times W^T (W \times y_k - c), \frac{\lambda}{L_k})$
$x_k = p_{L_k}(y_k)$
$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
$y_{k+1} = x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1})$

Output:$h = x_c$

We use projected FISTA to solve (4) where $W \in R^{m \times n}$, $h \in R^{n \times 1}$, $c \in R^{m \times 1}$. For non-negative orthant projection, we define projected FISTA to contain an additional projection step after the computation of $p_L(y_k)$ and $p_{L_k}(y_k)$ in Algorithms 1 and 2 respectively. The projection step is $(x_k)_m = 0$, $\forall m$ **s.t.** $(x_k)_m < 0$. So this gives two versions of projected FISTA, with constant step size and with backtracking.

We now state a proof of correctness for the convergence of the projected FISTA with non-negative orthant constraint for a broader class of problems.

Theorem 1 *If $g(x)$ is fully separable in terms of x, then projected FISTA converges for the problem* $\arg\min_{x \geq 0} f(x) + g(x)$, *when f is convex, smooth and its gradient is Lipschitz, and g is convex and continuous.*

Proof At iteration k, for any of our sub-problems the $p_L(y_k)$ will look as below:
$p_L(y_k) = \arg\min_{x \geq 0} \{g(x) + \frac{L}{2} \|x - (y_k - \frac{1}{L} \times \nabla f(y_k))\|_F^2\}$,
where $x \in R^n$. It is easy to see that $p_L(y_k)$ can be split into n independent sub-problems as follows:

$$p_L(y_k) = \arg\min_{x \geq 0} \sum_{i=1}^{n} \left\{ g(x_i) + \frac{L}{2} \|x_i - (y_k - \frac{1}{L} \times \nabla f(y_k))_i\|^2 \right\}$$
$$= \left\{ \arg\min_{x_i \geq 0} \left\{ g(x_i) + \frac{L}{2} \|x_i - (y_k - \frac{1}{L} \times \nabla f(y_k))_i\|^2 \right\} \right\}_{i=1:n},$$

since $g(x)$ and the squared Frobenius norm are separable. After the correct computation of $p_L(y_k)$ from independent sub-problems, the convergence proof for FISTA (Beck and Teboulle 2009) applies here as well. □

It is easy to see that since (4) is of the form in Theorem 1 where $f(h) = \frac{1}{2} \|c - W \times h\|_F^2$ and $g(h) = \lambda \|h\|_1$, the projected FISTA converges for l_1-regularized non-negative least squares problem.

Sindhwani and Ghoting (2012) study the problem of l_1-projection using FISTA, but have also suggested that the same strategy works for non-negative orthant projection. They don't specifically give a correctness proof for general objective functions or show experiments with non-negative orthant projections using FISTA.

We now have two versions of projected FISTA for non-negative orthant constraint, one with constant step size and the other with backtracking. We use both of these versions for different sub-problems, as we discuss in the next subsection.

4.2 Proposed algorithm

Having looked at projected FISTA for solving the l_1-regularized non-negative least squares problem, we now investigate the individual sub-problems for the four factor matrices.

Solving Φ: The sub-problem for Φ is as follows:

$$\begin{aligned}
&\arg\min_{\Phi\geq 0} \frac{1}{2}\|D-\Phi Q\|_F^2 + \lambda_\Phi \sum_{j=1}^{t} \|\Phi_j\|_1 \\
&= \arg\min_{\Phi\geq 0} \frac{1}{2}\left\|D^T - Q^T\Phi^T\right\|_F^2 + \lambda_\Phi \sum_{j=1}^{v} \|(\Phi^T)_j\|_1 \\
&= \left\{\arg\min_{(\Phi^T)_i\geq 0} \frac{1}{2}\|(D^T)_i - Q^T(\Phi^T)_i\|_F^2 + \lambda_\Phi\|(\Phi^T)_i\|_1\right\}_{i=1:v}, \qquad (5)
\end{aligned}$$

where we have split the sub-problem into further independent sub-problems involving individual $(\Phi^T)_i$ variables.

Let us set $f((\Phi^T)_i) = \frac{1}{2}\|(D^T)_i - Q^T(\Phi^T)_i\|_F^2$ and $g((\Phi^T)_i) = \lambda_\Phi\|(\Phi^T)_i\|_1$. Observe that f is convex, smooth and its gradient is Lipschitz. Further, g is convex and continuous. Therefore, we can directly use the projected version of Algorithm 1 for solving each individual sub-problem in $(\Phi^T)_i$. We use the version with constant step size because the largest eigenvalue computation step is required only once in each iteration of alternating minimization for solving all the sub-problems (in the order of thousands) in Φ whereas projected FISTA with backtracking requires computation of Lipschitz constant L_k for each step in an individual sub-problem as we discussed in Sect. 4.1.

Solving Θ: The sub-problem for Θ looks very similar to that for Φ:

$$\begin{aligned}
&\arg\min_{\Theta\geq 0} \frac{1}{2}\|Q-\Theta\mathbb{A}\|_F^2 + \lambda_\Theta \sum_{j=1}^{m} \|\Theta_j\|_1 \\
&= \arg\min_{\Theta\geq 0} \frac{1}{2}\left\|Q^T - \mathbb{A}^T\Theta^T\right\|_F^2 + \lambda_\Theta \sum_{j=1}^{t} \|(\Theta^T)_j\|_1 \\
&= \left\{\arg\min_{(\Theta^T)_i\geq 0} \frac{1}{2}\|(Q^T)_i - \mathbb{A}^T(\Theta^T)_i\|_F^2 + \lambda_\Theta\|(\Theta^T)_i\|_1\right\}_{i=1:t}, \qquad (6)
\end{aligned}$$

where we have split the last step into t independent sub-problems involving individual $(\Theta^T)_i$ variables. Observe that this problem is identical to that in (5), with Q substituted for D and $\mathbb{A}$ for Q. Therefore, for the same reasons, we can use projected version of Algorithm 1 for solving the t sub-problems independently. As for Φ, we use the version with constant step size to solve sub-problems in Θ also.

Solving Q: The sub-problem for Q is as follows:

$$\begin{aligned}
&\arg\min_{Q\geq 0} \frac{1}{2}\|D-\Phi Q\|_F^2 + \frac{1}{2}\|Q-\Theta\mathbb{A}\|_F^2 + \lambda_Q \sum_{i=1}^{n} \|Q_i\|_1 \\
&= \arg\min_{Q\geq 0} \frac{1}{2}\|[D;\Theta\mathbb{A}] - [\Phi; eye(t,t)]Q\|_F^2 + \lambda_Q \sum_{i=1}^{n} \|Q_i\|_1 \\
&= \left\{\arg\min_{Q_i\geq 0} \frac{1}{2}\|[D;\Theta\mathbb{A}]_i - [\Phi; eye(t,t)]Q_i\|_F^2 + \lambda_Q\|Q_i\|_1\right\}_{i=1:n}, \qquad (7)
\end{aligned}$$

where we have split it into independent sub-problems involving individual Q_i variables. We have used the notation $eye(n, n)$ to denote the identity matrix with n rows and n columns, and [;] for vertical concatenation of matrices.

As before, we can set $f(Q_i) = \frac{1}{2}\|[D; \Theta\mathbb{A}]_i - [\Phi; eye(t,t)]Q_i\|_F^2$ and $g(Q_i) = \lambda_Q\|Q_i\|_1$. Again, observe that these satisfy the conditions for FISTA. Therefore, we can apply projected version of Algorithm 1 to solve the n sub-problems independently. Again, we use the version with constant step size for similar reasons as before.

Solving $\mathbb{A}$: The sub-problem for $\mathbb{A}$ looks as follows:

$$\begin{aligned}
&\arg\min_{\mathbb{A}\geq 0} \frac{1}{2}\|Q - \Theta\mathbb{A}\|_F^2 \text{ s.t. } supp(\mathbb{A}) \subseteq supp(A) \\
&= \arg\min_{\mathbb{A}\geq 0} \frac{1}{2}\|Q - \Theta\mathbb{A}\|_F^2, \mathbb{A}_{ij} = 0\ \forall i, j \text{ s.t. } \mathbb{A}_{ij} \notin supp(A) \\
&= \left\{\arg\min_{\mathbb{A}_j\geq 0} \frac{1}{2}\|Q_j - \Theta\mathbb{A}_j\|_F^2, \mathbb{A}_{ij} = 0\ \forall i \text{ s.t. } \mathbb{A}_{ij} \notin supp(A)\right\}_{j=1:n}, \qquad (8)
\end{aligned}$$

where subscript ij denotes ith row and jth column and in the last line, we have split the problem into n independent sub-problems involving $\mathbb{A}_j$ variables.

Because of the constraint, the problem in (8) does not fit into the FISTA framework. However, we can reformulate the problem by incorporating the constraints in the definition of the variable as given below.

Let $G(j) = \{i : \mathbb{A}_{ij} \in supp(A)\}$ denote the indices of potential authors for document j. Let $\Theta_{G(j)}$ be the truncated matrix containing the columns from Θ indexed by $G(j)$, denoting the topical associations of these potential authors. Let $\{\mathbb{A}_j\}_{G(j)}$ be the truncated column vector containing the elements indexed by $G(j)$, denoting the document associations of the potential authors. Then it is easy to see that the individual sub-problems in (8) can be rewritten as

$$\arg\min_{\{\mathbb{A}_j\}_{G(j)}\geq 0} \frac{1}{2}\|Q_j - \Theta_{G(j)}\{\mathbb{A}_j\}_{G(j)}\|_F^2. \qquad (9)$$

This is a non-negative least squares problem which can be solved by projected version of Algorithm 2. For solving sub-problems in $\mathbb{A}$, we use FISTA with backtracking version because Lipschitz constant(largest eigen-value) needs to be calculated separately for each of the n sub-problems in $\mathbb{A}$ having different $\Theta_{G(j)}$ in FISTA with constant step size.

Note that all the sub-problems involved in solving each of the factor matrices Φ, Q, Θ and $\mathbb{A}$ are independent supporting parallelism. Solving each of these factor matrices Φ, Q, Θ and $\mathbb{A}$ alternatingly using projected versions of FISTA constitutes one main iteration of alternating minimization algorithm. Local convergence for this algorithm is guaranteed by the alternating minimization framework (Lin 2007) as discussed in the "Appendix".

5 Experiments

In this section, we experimentally evaluate various aspects of our proposed approach over synthetic and real world datasets. (A) Evaluation of projected FISTA for sparse non-negative least squares; (B) Evaluation of SMTF for dyadic data: generalization ability, accuracy with respect to available gold-standard; (B1) Effect of the sparseness on the factor matrices; (B2) Comparison with existing baselines; (B3) Evaluation for non-document data; (B4) Evaluation of execution time for SMTF and the baselines w.r.t each dataset.

5.1 Evaluating projected FISTA

In our first experiment, we empirically compare the convergence rates of projected FISTA and projected subgradient for the l_1-regularized non-negative least squares problem in (4). We are not aware of any prior experimental evaluation of projected FISTA for this problem. In this section, we report experimental results for projected FISTA with constant step size. The convergence behaviour is similar for projected FISTA with backtracking for smaller values of L_0 and η. So, we plot only the objective function values of (4) using projected FISTA with constant step size.

For this experiment, we randomly generate instances of the l_1-regularized non-negative least squares problem (4) by generating instances of c, W and h. Given dimensions m and n, we generate a non-negative real $m \times n$ matrix W and a non-negative real $n \times 1$ column vector h by drawing each entry independently from $N(\mu, \sigma)$. We then use a sparsity parameter $s \in [0, 1]$ to set each element of h independently to 0 by drawing from a Bernoulli distribution with parameter s. We call this the synthetic Sparse Non-negative Least Squares (SNNLS_s) data.

Based on the data sizes in our real data experiments, we generate (W,h) matrices for ($m = 500, n = 30$) and ($m = 5000, n = 10$) to evaluate convergence for both small and large matrices. In each case, we experimented with multiple values of the sparseness parameter $\lambda = 0, 0.1, 1, 10$, and we run the algorithm until convergence (with $\epsilon = 10^{-4}$). In Fig. 1a, we only plot the objective function values over iterations for one ($m = 5000, n = 10$) sample for $\lambda = 0, 0.1, 1, 10$. We have experimented with ($m = 500, n = 30$) sized matrices and for different samples also and the convergence trends look very similar. We can see that objective function stabilizes in about 60 iterations to values depending on λ. The corresponding plots for the projected subgradient algorithm are shown in Fig. 1b.

The two algorithms start with the same objective function value, which is 70,177. The plots record values after the first iteration. Observe that the values are so different for the two algorithms from the first iteration onwards that we needed to plot them separately. We can see that projected FISTA converges to minima in about 60 iterations. In contrast, projected subgradient converges very slowly conforming to the theoretical guarantees. The objective function values for projected FISTA and projected subgradient after 60 iterations for $\lambda = (0, 0.1, 1, 10)$ are (7, 7, 10, 36) and around 9.0e13 respectively. This is an empirical

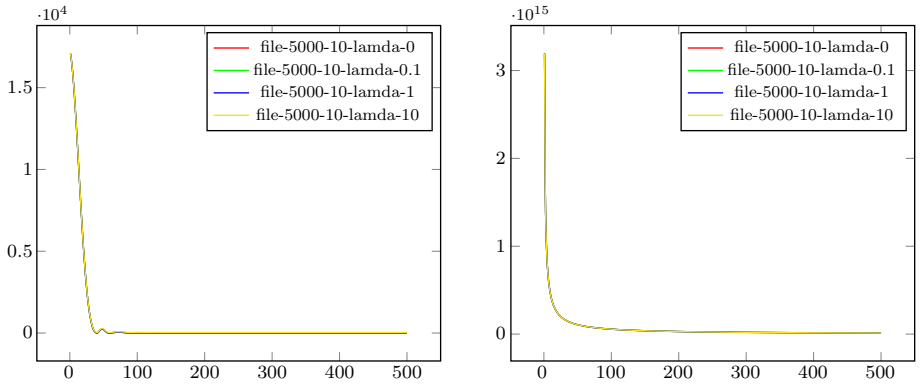

Fig. 1 Convergence plots for **a** projected FISTA and **b** projected subgradient

validation of our problem formulation that every sub-problem is solvable by the projected FISTA approach.

5.2 Evaluating SMTF on dyadic data

We first perform extensive experiments on user-document data, where we investigate the effect of sparsity parameters and compare performance with available baselines, and then briefly explore non-document data.

5.2.1 Evaluations on user-document data

Though our formulation supports arbitrary items and features, in our evaluation for this paper, we focus largely on document-word data, with a brief exploration of other kinds of items and features. Recall that our formulation takes as input a word-document matrix D and a user-document matrix A, and produces as output a word-topic matrix Φ, a topic-user matrix Θ, a topic-document matrix Q and a user-document matrix $\mathbb{A}$ using which we evaluate the generalization ability and accuracy of topic-document and user-topic associations.

Datasets: Our first dataset is the **DBLP** abstracts dataset[2] (DBLP), from which we use a subset of 6320 documents involving 3377 authors covering 8 conferences. Of these, 5533 documents are used for training, and the remaining 787 for testing, thereby ensuring that each author in test dataset is also present in at least one training document. This is done by letting each author vote with probability 0.7 and taking a majority vote whether to send a document to training set or test set. We consider each conference as a topic, so that each document is labeled with exactly one of the 8 topics. After eliminating rare words and stop words, the vocabulary is of size 3989 with a total of 0.57 million word occurrences.

As a second example of author-research paper associations, we use the **NIPS** dataset[3] (NIPS), which is a publication dataset from the Neural Information Processing Systems (NIPS) conference proceedings (volume 0-12). This collection contains 1,740 documents (complete papers) written by a total of 2,037 authors. Of these, 1,514 documents were used for training and 226 documents for testing, in a similar manner to that of the DBLP dataset. There are no gold standard topics available for this dataset. The vocabulary has 7717 words with a total of 2.2 million word occurrences.

In our third dataset, we consider associations between users and reviews. We use a subset of the **Product Review** (REV) dataset[4] from amazon.com, containing 9651 reviews written by 5675 reviewers in 10 different product categories (apparel, books, camera, computers, jewelry, kitchen items, magazines, music, sports, video). We create multi-author documents by concatenating all reviews written by different reviewers for one specific product. This results in 5998 documents, one for each product. Of these, 5040 documents were used for training and 958 documents for testing. We treat each product category as one topic, so that each document is labeled with exactly one of ten topics. The vocabulary has 8587 words with a total of 0.53 million word occurrences.

Setting parameters for SMTF: Recall that the SMTF formulation (3) involves three parameters for inducing sparsity, namely, λ_Θ, λ_Q and λ_Φ respectively. We perform a grid search over the space of these parameters and evaluate performance for each grid point.

[2] http://www.cs.uiuc.edu/hbdeng/data/kdd2011.htm.

[3] http://www.arbylon.net/resources.html.

[4] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

Evaluation measures: We evaluate accuracy of the individual factors against gold-standards where available, and additionally measure error on a test dataset using factors 'learnt' from training data. We discuss these in more detail below.

Document topic assignment accuracy: This corresponds to the evaluation of the topic-document matrix Q for SMTF and the corresponding equivalent for Collective Matrix Factorization Model (CMF) and Author-Topic Model (ATM) against an available gold-standard on the training dataset. We denote this as *DTa*. For DBLP and REV, we have gold-standard document-topic associations available, where each document is associated with exactly one topic. This can be interpreted as a hard clustering of the documents. To evaluate clustering accuracy, we use the F1 measure and the Adjusted Rand Index (ARI) over the pairwise clustering decisions. The Adjusted Rand Index for pairwise decisions is defined as $\frac{2(ab-cd)}{((a+d)(d+b)+(a+c)(c+b))}$, where a is the number of true positive pairs, b is the number of true negative pairs, c is the number of false positive pairs and d is the number of false negative pairs, respectively. The $F1$ measure is the harmonic mean of precision (P) and recall (R) over pairwise decisions: $F1 = \frac{2PR}{(P+R)}$, where $P = \frac{a}{(a+c)}$ and $R = \frac{a}{(a+d)}$. Higher values of $F1$ and ARI indicate higher clustering accuracy.

Author topic assignment accuracy: This is the evaluation of the topic-user matrix Θ for SMTF and the corresponding equivalent for CMF and ATM against an available gold-standard on the training dataset. We denote this as *ATa*. Since DBLP and REV have gold-standard topic labels associated with documents, we consider as 'gold-standard' $\overline{\Theta_{ij}}$, the fraction of documents associated with user j that are labeled with topic i. Since for each user j, $\Theta_{.j}$ (and $\overline{\Theta_{.j}}$) potentially assigns the user to multiple topics i with non-zero weights, we measure the soft clustering accuracy of the authors using the categorical clustering distance (CCD) (Zhou et al. 2005). CCD is obtained by solving the following optimization problem: $CCD(\Theta, \overline{\Theta}) = \min_{w_{k,j}} \sum_{k=1}^{K} \sum_{j=1}^{J} w_{k,j} \sum_{i=1}^{m} \|\overline{\Theta}_{k,i} - \Theta_{j,i}\|$, subject to $w_{k,j} \geq 0$, $\sum_{k=1}^{K} w_{k,j} = \frac{1}{J}$, $\sum_{j=1}^{J} w_{k,j} = \frac{1}{K}$ for all k, j, where K and J are the number of topics in $\overline{\Theta}$ and Θ, respectively, and m is the number of users. Smaller values of CCD indicate better agreement with the gold-standard.

Test error: Assuming a fixed set of authors and topics across corpora, the word-topic (Φ) and topic-user (Θ) matrices can be learnt from a training corpus and used for analyzing a held-out test corpus. The quality of the learnt matrices $\hat{\Phi}$ and $\hat{\Theta}$ can be measured using the Frobenius error in fitting the test corpus D_t. This can be measured by solving the following optimization problem for SMTF: $\text{TE}(D_t; \hat{\Phi}, \hat{\Theta}) = \min_{Q \geq 0, \mathbb{A} \geq 0} \frac{1}{2}\|D_t - \hat{\Phi} Q\|_F^2 + \frac{1}{2}\|Q - \hat{\Theta}\mathbb{A}\|_F^2$ while imposing sparsity on Q. For CMF, we use the corresponding formulation: $\min_{Q \geq 0, \sum_1 \geq 0, \sum_2 \geq 0} \frac{1}{2}\|D_t - Q \sum_1 \hat{\Phi}\|_F^2 + \frac{1}{2}\|A - \hat{\Theta} \sum_2 Q^T\|_F^2$ while imposing sparsity on Q. Note that lower values of test error indicate better generalization performance. We denote this as *TE*. Since ATM is trained to maximize a very different objective function (loglikelihood), we do not evaluate such error for it.

Before showing the results of experiments which quantitatively evaluate the accuracy of the factor matrices, we present an example list of topics for DBLP dataset and the associations of words and authors to the topics obtained from the SMTF formulation (3) for DBLP dataset in Table 1.

The topics presented in Table 1 correspond to the 8 conferences (AAAI, CIKM, IJCAI, SIGIR, SIGMOD, VLDB, ICDM, KDD) in DBLP dataset. Table 1 contains the list of ten topmost words and authors for the corresponding topics.

(B) In our first experiment with dyadic data, we study the effect of sparseness on the factor matrices in terms of the different evaluation measures outlined in Sect. 5.2.1. For experiments, our main focus is on $SMTF_1$ based on (3) using surrogate topic-document

Table 1 Illustration of topics for DBLP with top most list of words and authors associated with topics

TOPIC1 WORD	**TOPIC2 WORD**	**TOPIC3 WORD**	**TOPIC4 WORD**
Database	Mining	Based	Web
Systems	Databases	Learning	Search
System	Rules	Knowledge	Results
Object	Association	Model	Pages
Oriented	Large	Approach	User
Performance	Algorithm	Case	Engines
Design	Patterns	Reasoning	Page
Distributed	Algorithms	Problem	Engine
Parallel	Frequent	Logic	Services
Memory	Discovery	Planning	Content
AUTHOR	**AUTHOR**	**AUTHOR**	**AUTHOR**
Mary Hunter Utt	Tanzeem Choudhury	Jennifer Seberry	Tokuro Matsuo
Aditya Krishna Menon	ric Grgoire	Takuji Takahashi	Jongeun Jun
Imran R. Mansuri	Francisco Azuaje	Bou-Ho Yang	Alex Lopez
Rahul Balakrishnan	Lei Yu	Ellen L. Drascher	Ya Zhang
Arvind Hulgeri	Michael A. Gennert	Bruce Gordon	Takushi Sogo
Anne Descour	Dniel Fogaras	Srinivasan Jagannathan	Daniela Pucci de Farias
Shang-Yoon Hahn	Eric Horvitz	Chengyu Sun	Mustafa Uysal
Hongbo Liu	Stphane Galland	Bing Liu 0003	Jrmie Mary
Susan P. Ennis	Raya Fidel	Jihoon Yang	Cristian Riveros
Jos Borges	Shaozhi Ye	Masahiro Terabe	Hao Huang
TOPIC5 WORD	**TOPIC6 WORD**	**TOPIC7 WORD**	**TOPIC8 WORD**
Data	Xml	Query	Retrieval
Streams	Queries	Queries	Information
Management	Index	Processing	Language
Model	Relational	Optimization	Document
Clustering	Documents	Join	Text
Applications	Efficient	Time	Models
High	Tree	Execution	Relevance
Dimensional	Xquery	Expansion	Probabilistic
Analysis	Structure	Evaluation	Feedback
Base	Twig	Approximate	Cross
AUTHOR	**AUTHOR**	**AUTHOR**	**AUTHOR**
Oliver Schulte	Claudio Petrone	Dejing Dou	Antonio Gentile
Martin Choquette	Pong Chi Yuen	Manish Gupta 0003	Aleix Gimnez-Va
Anupam Mediratta	Christophe Lcluse	Hang Yu	Daniel Olmedilla

Table 1 continued

AUTHOR	AUTHOR	AUTHOR	AUTHOR
Ralph Michaelis	Carl Weir	Tokuji Okada	M. Mineev
Pablo Hernndez	Paul Schmidt	X. Zhao	Michael Schwartz
Ron Obermarck	Fatih Kahraman	Oded Netzer	Badrish Chandramouli
V. Puig	Bin Zhang	Colin Bell	David A. Casta
Hua Shi	Thomas Ohler	Carter Collins	Katherine A. Morris
John A. Moyne	Johan de Kleer	Rajendra T. Dodhiawala	Brian Grom
Thomas J. Gambino	Thomas C. Wolf	Arun C. Surendran	Keiko Horiguchi

Table 2 $SMTF_1$ Test Error for $\lambda_Q = 0$ and $\lambda_Q > 0$

DBLP				NIPS				REV			
λ_Φ	λ_Θ	λ_Q	*TE*	λ_Φ	λ_Θ	λ_Q	*TE*	λ_Φ	λ_Θ	λ_Q	*TE*
0	0	0	**1.3219e5**	0	0.1	0	**7.4316e5**	0	0	0	4.4373e4
0	0.1	0	1.3224e5	1	0.1	0	7.4345e5	0	0.1	0	**4.4368e4**
3	0.1	0	1.3241e5	5	0.1	0	7.4485e5	2	0.1	0	4.4577e4
7	0.1	0	1.3262e5	10	0.1	0	7.4685e5	7	0.1	0	4.5024e4
10	0.1	0	1.3277e5					10	0.1	0	4.5240e4
0	0.1	2	1.3147e5	0	0.1	50	**7.3630e5**	0	0.1	2	4.3133e4
0	0.1	10	1.3127e5	0	0.1	100	7.4288e5	0	0.1	4	**4.3129e4**
0	0.1	20	1.3651e5	0	0.1	150	7.6880e5	0	0.1	10	4.6168e4
2	0.1	2	1.3151e5	0.1	0.1	100	7.4345e5	2	0.1	2	4.4174e4
2	0.1	5	**1.3124e5**	0.1	0.1	250	8.2356e5	2	0.1	5	4.3545e4
2	0.1	10	1.3241e5	0.5	0.1	150	7.6722e5	2	0.1	7	4.4624e4
5	0.1	2	1.3159e5	0.5	0.1	250	8.2424e5	5	0.1	2	4.3270e4
5	0.1	5	1.3142e5	0.5	0.1	350	9.2636e5	5	0.1	5	4.4307e4
5	0.1	10	1.3259e5	1	0.1	250	8.2653e5	5	0.1	7	4.5219e4
7	0.1	2	1.3164e5					7	0.1	2	4.3396e4
7	0.1	5	1.3155e5					7	0.1	5	4.4697e4
7	0.1	7	1.3155e5					7	0.1	7	4.5530e4

The bold optimal values indicate best metrics results for different settings of $\lambda_\phi, \lambda_\theta, \lambda_Q$ sparsity parameters

matrix Q. Additionally, we evaluate $SMTF_2$ based on (1) for comparisons with $SMTF_1$. Because of huge time taken by $SMTF_2$ as reported in Table 10 and since local convergence of $SMTF_2$ is not guaranteed as discussed in the "Appendix", we evaluate $SMTF_2$ only for few parameter combinations including those that perform best for $SMTF_1$.

For $SMTF_1$, we have seen that sparse Φ does not help. Therefore, our main focus for the rest of the experiment is on the other two sparsity factors Q and Θ.

(B1) First, for $SMTF_1$, we check how sparsity on Q using λ_Q affects test error. We wish to verify if imposing sparsity on Q fits the held out data better. Formally we check if: $\min_{\lambda_\Phi,\lambda_\Theta}\{TE : \lambda_Q = 0\} > \min_{\lambda_\Phi,\lambda_\Theta}\{TE : \lambda_Q > 0\}$.

In Table 2, we evaluate $SMTF_1$ for different settings of λ_Φ and λ_Θ for $\lambda_Q = 0$ and $\lambda_Q > 0$ for the three datasets. Our evaluations over different sparsity parameter settings showed that

Table 3 $SMTF_2$ Test Error

DBLP				NIPS				REV			
λ_Φ	λ_Θ	$\lambda_{\Theta\mathbb{A}}$	*TE*	λ_Φ	λ_Θ	$\lambda_{\Theta\mathbb{A}}$	*TE*	λ_Φ	λ_Θ	$\lambda_{\Theta\mathbb{A}}$	*TE*
0	0	0	1.5324e5	0	0	0	1.3493e6	0	0	0	5.2067e4
0	0.1	0	1.5194e5	0	0.1	0	1.3295e6	0	0.1	0	5.0951e4
0	0.6	0	**1.5087e5**	0	2	0	1.2196e6	0	1.5	0	4.8508e4
0	2	0	1.5176e5	0	5	0	**1.2134e6**	0	5	0	**4.8376e4**
2	0	0	1.5314e5	0.5	0	0	1.3492e6	2	0	0	5.2047e4
100	0	0	1.5131e5	1	0	0	1.3490e6	7	0	0	5.2030e4
250	0	0	1.5233e5	5	0	0	1.3477e6	15	0	0	5.1997e4
2	2	0	1.5177e5	1	2	0	1.2194e6	2	1	0	4.8784e4
0	0	7	6.3735e5	0	0.1	50	2.785e7	0	0	0.03	5.2051e4
2	0.1	5	4.1271e5	–	–	–	–	0	0.1	0.03	5.1061e4
	–	–	–	–	–	–	–	0	0.1	4	1.3582e5

The bold optimal values indicate best metrics results for different settings of $\lambda_\phi, \lambda_\theta, \lambda_Q$ sparsity parameters

for a fixed value of $(\lambda_\Phi, \lambda_\Theta)$, as λ_Q increases from zero, the TE curve is convex. TE decreases to a minimum and then increases again. The location of the convex curve depends on the $(\lambda_\Phi, \lambda_\Theta)$ value. We also noticed that sparse Θ does not affect TE since we minimize the term $\|Q - \Theta\mathbb{A}\|_F^2$ so that the surrogate Q and product $\Theta\mathbb{A}$ are close. In Table 2, we record the value of TE for $SMTF_1$ over all curves when $\lambda_Q = 0$ and when $\lambda_Q > 0$ for the three different datasets. We see that in all the three datasets, for $SMTF_1$, lower TE can be achieved when $\lambda_Q > 0$, showing the importance of sparsity of Q for generalization.

We evaluate TE for $SMTF_2$ by solving the following optimization problem: $\text{TE}(D_t; \hat{\Phi}, \hat{\Theta}) = \min_{\mathbb{A}\geq 0} \frac{1}{2}\|D_t - \hat{\Phi}\hat{\Theta}\mathbb{A}\|_F^2$ while imposing sparsity on $\hat{\Theta}\mathbb{A}$, analogous to Q in $SMTF_1$. Table 3 records the value of TE for $SMTF_2$ over parameter combinations that perform best for $SMTF_1$ and with $(\lambda_\Phi = 0; \lambda_{\Theta\mathbb{A}} = 0), (\lambda_\Theta = 0; \lambda_{\Theta\mathbb{A}} = 0), (\lambda_{\Theta\mathbb{A}} = 0), (\lambda_{\Theta\mathbb{A}} > 0)$ for the three different datasets. Unlike $SMTF_1$, sparsity of $\Theta\mathbb{A}$ in $SMTF_2$ increases the TE. Sparsity individually for Φ and Θ decreases the TE. Additional sparsity in Θ along with sparsity in Φ further decreases TE than just the sparsity in Φ. But sparsity in Θ alone performs the best for all the three datasets. From Tables 2 and 3, $SMTF_1$ performs better than $SMTF_2$ in terms of generalization ability.

Next, we check how sparsity on Q using λ_Q in $SMTF_1$ affects document-topic assignment accuracy (DTa). Similar to experiments for *TE*, we hypothesize that sparsity on Q using λ_Q improves document-topic assignment accuracy (DTa) which we formally state as: $\max_{\lambda_\Phi,\lambda_\Theta}\{DTa - F1, DTa - ARI : \lambda_Q = 0\} < \max_{\lambda_\Phi,\lambda_\Theta}\{DTa - F1, DTa - ARI : \lambda_Q > 0\}$.

In Table 4, we evaluate $SMTF_1$ for different settings of λ_Φ and λ_Θ for $\lambda_Q = 0$ and $\lambda_Q > 0$ for DBLP and REV, which have gold-standard topics. As for TE, for DTa we see a similar trend, with the difference that the curve is concave. For a fixed value of $(\lambda_\Phi, \lambda_\Theta)$, as λ_Q increases from zero, DTa first increases to a maximum and then falls off. Here also DTa is not affected by sparsity on Θ. In Table 4, we record the value of DTa for $\lambda_Q = 0$ and for $\lambda_Q > 0$ across curves. For DBLP, we see that better DTa (in terms of both F1 and

Table 4 $SMTF_1$ Document-topic accuracy for $\lambda_Q = 0$ and $\lambda_Q > 0$

DBLP					REV				
λ_Φ	λ_Θ	λ_Q	DTa-F1	DTa-ARI	λ_Φ	λ_Θ	λ_Q	DTa-F1	DTa-ARI
0	0	0	0.2538	0.1286	0	0	0	0.5208	0.4509
0	0.1	0	0.2547	0.1293	0	0.1	0	**0.5235**	**0.4541**
3	0.1	0	0.2548	0.1292	2	0.1	0	0.5078	0.4349
7	0.1	0	0.2568	0.1308	7	0.1	0	0.5169	0.4461
100	0.1	0	0.2757	**0.1424**	10	0.1	0	0.5102	0.4441
150	0.1	0	0.2785	0.1254					
250	0.1	0	**0.2806**	0.1211					
350	0.1	0	0.2701	0.0973					
0	0.1	5	0.2626	0.1344	0	0.1	0.03	**0.5265**	**0.4577**
0	0.1	10	0.2875	0.1578	0	0.1	2	0.4686	0.3905
0	0.1	20	0.2771	0.1269	0	0.1	4	0.4638	0.3849
2	0.1	2	0.2623	0.1333	0	0.1	10	0.1819	1.01e−5
2	0.1	5	0.2803	0.1562	2	0.1	2	0.4338	0.3445
2	0.1	10	0.3225	**0.1899**	2	0.1	5	0.3897	0.2883
2	0.1	20	0.2891	0.1320	2	0.1	7	0.3734	0.2669
5	0.1	5	0.2729	0.1486	5	0.1	2	0.4213	0.3268
5	0.1	7	0.2870	0.1593	5	0.1	5	0.4176	0.3241
5	0.1	10	**0.3260**	0.1834	5	0.1	7	0.3642	0.2569
5	0.1	20	0.2524	0.0145	7	0.1	2	0.4169	0.3217
7	0.1	2	0.2668	0.1340	7	0.1	5	0.3318	0.2104
7	0.1	5	0.2588	0.1320	7	0.1	7	0.2354	0.0825
7	0.1	7	0.2798	0.1535	–	–	–	–	–

The bold optimal values indicate best metrics results for different settings of λ_ϕ, λ_θ, λ_Q sparsity parameters

Table 5 $SMTF_2$ Document-topic accuracy for $\lambda_{\Theta\mathbb{A}} = 0$ and $\lambda_{\Theta\mathbb{A}} > 0$

DBLP					REV				
λ_Φ	λ_Θ	$\lambda_{\Theta\mathbb{A}}$	DTa-F1	DTa-ARI	λ_Φ	λ_Θ	$\lambda_{\Theta\mathbb{A}}$	DTa-F1	DTa-ARI
100	0	0	**0.2578**	**0.1330**	0	0	0	**0.5024**	**0.4319**
100	0.1	0	0.2560	0.1299	0	0.1	0	0.5014	0.4307
250	0	0	0.2448	0.1086	–	–	–	–	–
250	0.1	0	0.2424	0.1036	–	–	–	–	–
2	0	10	**0.2790**	**0.0526**	0	0	0.03	**0.5084**	**0.4403**
2	0.1	10	0.2592	1.2437e−6	0	0.1	0.03	0.4376	0.3547
5	0	10	0.2590	−1.9708e−4	–	–	–	–	–
5	0.1	10	0.2590	−2.4231e−4	–	–	–	–	–

The bold optimal values indicate best metrics results for different settings of λ_ϕ, λ_θ, λ_Q sparsity parameters

Table 6 $SMTF_1$ Author-topic accuracy for $\lambda_Q = 0$ versus $\lambda_Q > 0$ and $\lambda_\Theta = 0$ versus $\lambda_\Theta > 0$

DBLP				REV			
λ_Φ	λ_Q	λ_Θ	ATa-CCD	λ_Φ	λ_Q	λ_Θ	ATa-CCD
0	0	0	568.2	0	0	0	**515.2**
2	0	0	**568**	2	0	0	526.1
5	0	0	568.1	5	0	0	532.7
7	0	0	568.5	7	0	0	535.2
0	2	0	561.4	0	0.03	0	**508.2**
0	5	0	562.8	0	2	0	527.9
0	7	0	**551.1**	0	4	0	531
–	–	–	–	0	5	0	573.1
0	0	0.1	569.6	0	0	0.1	548
0	0	0.6	**544.9**	0	0	1.5	**495.9273**
3	0	0.1	569.7	2	0	0.1	553.5
7	0	0.1	570	7	0	0.1	553
10	0	0.1	569.7	10	0	0.1	567.3
0	2	0.1	555	0	2	0.1	481.4
0	5	0.1	545.7	0	4	0.1	**454.34**
0	5	0.15	538.9	0	7	0.1	775.4
0	10	0.1	531.2	2	2	0.1	523.9
0	15	0.6	**423**	2	5	0.1	536.5
2	2	0.1	557.1	2	7	0.1	518.3
2	5	0.1	532.7	2	7	1.5	559.7
2	7	0.1	531.3	5	2	0.1	527.7
5	2	0.1	556.3	5	5	0.1	497.61
5	5	0.1	537.3	5	7	0.1	539.4
5	7	0.1	533	7	2	0.1	542.9
7	2	0.1	557	7	2	1.5	538.5
7	5	0.1	542	7	5	0.1	564.3
7	7	0.1	537	7	7	0.1	570.3

The bold optimal values indicate best metrics results for different settings of $\lambda_\phi, \lambda_\theta, \lambda_Q$ sparsity parameters

ARI) is achieved with $\lambda_Q > 0$, suggesting that sparsity in Q leads to better recovery of document-topic associations.

For the REV dataset, sparse Q is not as helpful for DTa. DTa improves very marginally for small epsilon>0, and then falls off.

Table 5 records the value of DTa for $SMTF_2$ over parameter combinations that perform best with $SMTF_1$ when $\lambda_Q = 0$ and $\lambda_Q > 0$ for the two datasets. Additional sparsity on Θ along with sparsity on $\Theta\mathbb{A}$, analogous to Q in $SMTF_1$ does not help, thereby decreasing DTa, see Table 5. So, with $\lambda_\Theta = 0$, sparsity on $\Theta\mathbb{A}$ has an effect similar to that in $SMTF_1$ for REV dataset. For DBLP, sparsity on $\Theta\mathbb{A}$ improves DTa-F1 but not DTa-ARI. Tables 4 and 5 show that $SMTF_1$ outperforms $SMTF_2$ in terms of DTa-F1 and DTa-ARI for both the datasets.

(B1) For $SMTF_1$, we evaluate the impact of sparsity of Q and Θ, incorporated through λ_Q and λ_Θ on author-topic assignment accuracy (ATa). Here we check if sparsity on both Q and Θ increases author-topic assignment accuracy (ATa) i.e, decreases the ATa-CCD. Formally,

Table 7 $SMTF_2$ Author-topic accuracy for $\lambda_{\Theta\mathbb{A}} = 0$ versus $\lambda_{\Theta\mathbb{A}} > 0$ and $\lambda_{\Theta} = 0$ versus $\lambda_{\Theta} > 0$

DBLP				REV			
λ_{Φ}	$\lambda_{\Theta\mathbb{A}}$	λ_{Θ}	ATa-CCD	λ_{Φ}	$\lambda_{\Theta\mathbb{A}}$	λ_{Θ}	ATa-CCD
2	0	0	571.25	0	0	0	568.14
0	7	0	548.26	0	0.03	0	547.43
0	0	0.6	548.22	0	0	1.5	**508**
–	–	–	–	0	0	5	511.4
0	15	0.6	**423**	0	0.03	0.1	551
–	–	–	–	0	4	0.1	650

The bold optimal values indicate best metrics results for different settings of $\lambda_{\phi}, \lambda_{\theta}, \lambda_{Q}$ sparsity parameters

we check if: $\min_{\lambda_{\Phi}}\{ATa\text{-}CCD : \lambda_Q = 0, \lambda_{\Theta} = 0\} > \min_{\lambda_{\Phi}}\{ATa\text{-}CCD : \lambda_Q > 0, \lambda_{\Theta} = 0\} > \min_{\lambda_{\Phi}}\{ATa\text{-}CCD : \lambda_Q = 0, \lambda_{\Theta} > 0\} > \min_{\lambda_{\Phi}}\{ATa\text{-}CCD : \lambda_Q > 0, \lambda_{\Theta} > 0\}$.

In Table 6, we record the values of ATa-CCD for DBLP and REV by evaluating $SMTF_1$ for the four settings $(\lambda_Q = 0, \lambda_{\Theta} = 0)$, $(\lambda_Q > 0, \lambda_{\Theta} = 0)$, $(\lambda_Q = 0, \lambda_{\Theta} > 0)$ and $(\lambda_Q > 0, \lambda_{\Theta} > 0)$ over specific values of λ_{Φ}. We see that for both datasets, sparsity individually for Q and Θ helps. Sparsity of Θ helps more significantly than sparsity of Q. Sparsity of Q has an effect similar to that of DTa for both datasets. Simultaneous sparsity in both Θ and Q brings even bigger benefits for both DBLP and REV datasets. The overall pattern is that sparsity in Q and sparsity in Θ are both beneficial in different extents for author-topic assignment.

For similar settings, we evaluate $SMTF_2$ for best parameter combinations in Table 6 to obtain ATa-CCD for DBLP, REV and record the corresponding results in Table 7. Compared to $SMTF_1$ for ATa-CCD, sparsity individually for $\Theta\mathbb{A}$, Θ helps and sparsity of Θ helps more than sparsity of $\Theta\mathbb{A}$ here as well for both datasets. Imposing sparsity on both Θ and $\Theta\mathbb{A}$ brings even bigger benefits for DBLP, but does not help for REV dataset. Sparsity on Θ alone helps the most for REV dataset. From Tables 6 and 7, we can observe that $SMTF_1$ performs better than $SMTF_2$ for ATa-CCD metric.

(B2) For dyadic data, the *Collective Matrix Factorization Model* (**CMF**) (Sachan and Srivastava 2013) and the probabilistic *Author-Topic Model* (**ATM**) (Rosen-Zvi et al. 2004) can perform sparse topical analysis and we compare our proposed SMTF approaches, $SMTF_1$ and $SMTF_2$, against both of these baselines. For ATM, we use available code [5], and search over the hyper-parameters α and β to identify the best performing configurations. For CMF, since no source code is available, we use our own implementation in matlab, following specifications mentioned in the paper. (We impose l_1-sparsity on the factor matrices Q, Φ and Θ and penalize their Frobenius norms.) We use projected subgradient descent to make an update for the unconstrained optimization and then project the updates to constrained space after each iteration. Again, we search over the sparsity parameters of the Q, Φ and Θ to identify the best configuration.

In Table 8, we report the best performance for each of the compared models across parameter configurations in terms of test error, document-topic accuracy and author-topic accuracy for the three datasets. While author-topic model performs the best in getting the document-topic assignment correct, our model $SMTF_1$ performs best in one more important metrics i.e, author-topic assignment which can be used in some important applications such as

[5] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

Table 8 Performance comparison between SMTF, CMF and ATM

Model	DBLP				NIPS	REV			
	TE	DTa-F1	DTa-ARI	ATa-CCD	TE	TE	DTa-F1	DTa-ARI	ATa-CCD
$SMTF_1$	**131,240**	0.33	0.19	**423**	**736,300**	**43,129**	0.53	0.46	**454**
$SMTF_2$	150,870	0.28	0.13	**423**	1,213,400	48,376	0.51	0.44	508
CMF	154,580	0.25	0.01	428	1,294,200	48,588	0.16	0.01	884
ATM	–	**0.35**	**0.24**	518	–	–	**0.60**	**0.55**	612

The bold optimal values indicate best metrics results for different settings of $\lambda_\phi, \lambda_\theta, \lambda_Q$ sparsity parameters

Table 9 Performance comparison between SMTF, CMF and ATM for synthetic product dataset

Model	TE	DTa-F1	DTa-ARI	ATa-CCD
$SMTF_1$	**887**	0.28	0.09	420
$SMTF_2$	2124	**0.29**	0.13	**417**
CMF	3528	0.25	0.01	424
ATM	–	**0.29**	**0.18**	555

The bold optimal values indicate best metrics results for different settings of $\lambda_\phi, \lambda_\theta, \lambda_Q$ sparsity parameters

automated reviewer recommendation (Rosen-Zvi et al. 2004), in recommending researchers with similar interests for academic collaboration, for friends recommendation in social media like facebook and twitter etc. and also has the best generalization error.

We do not report TE for ATM since it is a probabilistic model and has a different criterion of optimizing the likelihood rather than minimizing the Frobenius error.

5.2.2 Evaluation on non-document data

(B3) We also performed an experiment to compare the models for non-document items that have a small number of binary features. To simulate non-document items, such as products, movies, etc, which are associated with a small number of binary features, we modify the DBLP dataset as follows. We restrict the vocabulary by taking only the top 3 words from each of the 8 topics (associated with conferences). We select only those documents which contain at least one of these words. For any selected document and a word, the corresponding entry in D is set to 1 if the document contains the word, and to 0 otherwise. This results in a dataset containing 5018 training documents and 1145 documents for test dataset with a total of 3366 authors. The vocabulary has 24 words with a total of 27 thousand word occurrences. We call this the synthetic Product dataset (PROD_s).

The results are recorded in Table 9. Again, we see the same trend here as well that $SMTF_1$ performs the best in terms of TE, $SMTF_2$ in terms of ATa-CCD, ATM and $SMTF_2$ over DTa-F1 and ATM in DTa-ARI metrics.

5.2.3 Execution time

(B4) Finally, we record the sequential run-time for our model SMTF ($SMTF_1$ and $SMTF_2$) and the baselines CMF, ATM w.r.t each dataset in Table 10 until convergence is achieved on an Intel machine with 6-core processors and 24GB RAM.

Table 10 Run-time comparison between SMTF, CMF and ATM for three datasets in minutes

Model	NIPS	DBLP	REV
$SMTF_1$	252	535	863
$SMTF_2$	624	1342	2935
CMF	49	96	175
ATM	–	**6**	**7**

The bold optimal values indicate best metrics results for different settings of $\lambda_\phi, \lambda_\theta, \lambda_Q$ sparsity parameters

We see that ATM has very little sequential execution time compared to SMTF and CMF in Table 10. The time required for execution of $SMTF_2$ is approximately 3 times higher than that required for $SMTF_1$ as seen from Table 10. This is because, for solving the sub-problem involving Θ in (1), since Lipschitz constant of the gradient of the $\arg\min_{\Theta \geq 0} \|D - \Phi\Theta\mathbb{A}\|_F^2$ is not known, we choose projected FISTA with backtracking to solve for Θ. Note that this problem cannot be decomposed into a series of l1-constrained non-negative least squares problems because of its middle position. Hence, the computation of step 2 in Algorithm 2 and $p_{L_k}(y_k)$ requires matrix multiplications leading to enormous increase in time. Additionally, the sub-problem involving $\mathbb{A}$ takes twice the amount of time than that required for solving $\mathbb{A}$ in (3) because the coefficient matrix ($\Phi \times \Theta$) in (1) has higher dimension ($v \times n$) than the corresponding coefficient matrix (Θ) in (3) to solve $\mathbb{A}$.

Regarding the parallel implementation of SMTF and the baselines, to the best of our knowledge, the parallel version of ATM is not known. In the other baseline CMF, even though the sub-problems for some of the factor matrices cannot be decomposed into a series of independent sparse non-negative least squares problems, they can be parallelized provided the matrix multiplications in the update step for factor matrices are executed parallely.

For the parallel execution of $SMTF_1$, the following factors may be considered for high scalability:

(A1) Parallel computation of independent sub-problems (sparse non-negative least squares problem) for each of the factor matrices Φ, Θ, Q and $\mathbb{A}$; (A2) Projected FISTA with backtracking version for evaluation of Θ matrix because the computation of Lipschitz constant i.e., the largest eigen-value of PP^T will take a large time in projected FISTA with constant step size; (A3) For the sub-problem in Θ matrix, the computation of $p_{L_k}(y_k)$ involving matrix to vector multiplications and also computation of Lipschitz constant in step 2 of Algorithm 2 containing independent terms involving matrix to vector multiplications are parallelized.

For $SMTF_2$, we consider similar factors as for $SMTF_1$ such as:

(C1) Parallel computation of independent sub-problems (sparse non-negative least squares problem) for the factor matrices Φ and $\mathbb{A}$; (C2) Projected FISTA with backtracking version for evaluation of Θ matrix as explained above; (C3) For the sub-problem in Θ matrix, the computation of $p_{L_k}(y_k)$ involving matrix multiplications and also computation of Lipschitz constant in step 2 of Algorithm 2 containing independent terms involving matrix multiplications are parallelized.

To summarize the experiments, we have seen that using the proposed sparse tri-factorization approach, the role that sparsity of Q and Θ plays in the performance of the model—sparsity leads to improvements in test error, and recovery of both document-topic and user-topic associations. Additionally, we have experimented with original sparse tri-factorization approach without surrogate topic-document matrix for comparison purposes. Further, for the task of topical analysis, for dyadic data, it outperforms existing tri-factorization baselines and is competitive with respect to the probabilistic author topic model.

6 Conclusion

We explored the problem of sparse topical analysis of dyadic data using non-negative matrix tri-factorization framework. To make the matrix tri-factorization formulation tractable, we introduced a surrogate matrix for the product of topic-user Θ and user-document $\mathbb{A}$ matrices with l_1-sparsity constraints on individual factor matrices as well as on the product of Θ and $\mathbb{A}$ matrices with support constraints on $\mathbb{A}$. This has not been studied before. We used projected FISTA for solving each of the factor matrices Φ, Q, Θ and $\mathbb{A}$ in an alternating minimization framework supporting parallelism. Experimentally, we have demonstrated that the proposed approach outperforms existing baselines for the task of sparse topical analysis for dyadic data.

Compliance with ethical standards

Conflict of interest The author declares that there is no conflict of interest.

Appendix

Derivation for l_1-norm on $\mathbb{A}$

The sub-problem involving $\mathbb{A}$ with surrogate on $\Phi\Theta$ will involve the terms $\{\|(\Theta\mathbb{A})_j\|_1\}_{j=1}^n$ resulting in weighted l_1-norm on elements of columns of $\mathbb{A}$ given by the following derivation with notation X_{i*} denoting ith row of X. Note that l_1-norm is separable on the terms in the following derivation because of non-negative constraints on the optimization variables.

$$\begin{aligned}
\|(\Theta\mathbb{A})_j\|_1 &= \sum_{i=1}^{t} \|\Theta_{i*}\mathbb{A}_j\|_1 \\
&= \sum_{i=1}^{t} \| \sum_{k=1}^{m} \Theta_{ik}\mathbb{A}_{kj}\|_1 \\
&= \sum_{i=1}^{t} \sum_{k=1}^{m} \|\Theta_{ik}\mathbb{A}_{kj}\|_1 \\
&= \sum_{i=1}^{t} \sum_{k=1}^{m} \|\Theta_{ik}\|_1 \|\mathbb{A}_{kj}\|_1 \\
&= \sum_{k=1}^{m} \sum_{i=1}^{t} \|\Theta_{ik}\|_1 \|\mathbb{A}_{kj}\|_1 \\
&= \sum_{k=1}^{m} \|\mathbb{A}_{kj}\|_1 \sum_{i=1}^{t} \|\Theta_{ik}\|_1 \\
&= \sum_{k=1}^{m} \|\mathbb{A}_{kj}\|_1 \|\Theta_k\|_1.
\end{aligned}$$

Local convergence for SMTF

In Powell (1973), and Bertsekas (1999), it is shown that block coordinate descent methods require sub-problems to have unique solutions in order to guarantee local convergence. For the case of two blocks, Grippo and Sciandrone (2000) have shown that this uniqueness condition need not be satisfied. A classical example for the case of two blocks is NMF which employs the alternating minimization, a class of block coordinate descent method.

In SMTF, we have four independent convex sub-problems in Φ, Q, Θ, $\mathbb{A}$ respectively and hence the uniqueness condition need not be satisfied because strict convexity is not guaranteed for sub-problems in Φ, Θ, $\mathbb{A}$. So, we tweak the problem in (3) so that uniqueness condition required by block coordinate descent for local convergence is guaranteed as follows.

Consider a variant of (3) wherein we have added the weighted squared Frobenius norms of Φ, Θ, $\mathbb{A}$ (Kim and Park 2008b) as shown below.

$$\begin{aligned}
&\arg\min_{\Phi,\Theta,Q,\mathbb{A}} \frac{1}{2}\|D-\Phi Q\|_F^2+\frac{1}{2}\|Q-\Theta\mathbb{A}\|_F^2+\lambda_Q\sum_{j=1}^{n}\|Q_j\|_1+\lambda_\Theta\sum_{j=1}^{m}\|\Theta_j\|_1\\
&\quad+\lambda_\Phi\sum_{j=1}^{t}\|\Phi_j\|_1+c_\Phi\|\Phi\|_F^2+c_\Theta\|\Theta\|_F^2+c_{\mathbb{A}}\|\mathbb{A}\|_F^2\\
&s.t. supp(\mathbb{A})\subseteq supp(A),\quad \Phi\geq 0,\quad \Theta\geq 0,\quad Q\geq 0,\quad \mathbb{A}\geq 0
\end{aligned}\tag{10}$$

The sub-problem for Q in (10) remains the same as that in (3) and is strictly convex because the coefficient matrix $[\Phi; eye(t,t)]$ has a full column rank implying the Hessian of the convex sub-problem is positive definite. Hence a unique solution is guaranteed for the sub-problem in Q.

The addition of weighted squared Frobenius norms of $\Phi, \Theta, \mathbb{A}$ will make the corresponding sub-problems for Φ, Θ, $\mathbb{A}$ in (10) strictly convex by making the columns of the corresponding coefficient matrices independent, thereby ensuring local convergence for (10). The sub-problems for Φ, Θ, $\mathbb{A}$ in (10) are given below.

Solving Φ:

$$\begin{aligned}
&\arg\min_{\Phi\geq 0}\frac{1}{2}\|D-\Phi Q\|_F^2+\lambda_\Phi\sum_{j=1}^{t}\|\Phi_j\|_1+c_\Phi\|\Phi\|_F^2\\
&=\arg\min_{\Phi\geq 0}\|[\sqrt{0.5}D\cdot, \cdot zeros(v,t)]^T-[\sqrt{0.5}Q\cdot, \cdot\sqrt{c_\Phi}eye(t,t)]^T\Phi^T\|_F^2\\
&\quad+\lambda_\Phi\sum_{j=1}^{v}\|(\Phi^T)_j\|_1.
\end{aligned}$$

The symbol $[\cdot,\cdot]$ denotes the horizontal concatenation of matrices.

Solving Θ:

$$\begin{aligned}
&\arg\min_{\Theta\geq 0}\frac{1}{2}\|Q-\Theta\mathbb{A}\|_F^2+\lambda_\Theta\sum_{j=1}^{m}\|\Theta_j\|_1+c_\Theta\|\Theta\|_F^2\\
&=\arg\min_{\Theta\geq 0}\|[\sqrt{0.5}Q\cdot, \cdot zeros(t,m)]^T-[\sqrt{0.5}\mathbb{A}\cdot, \cdot\sqrt{c_\Theta}eye(m,m)]^T\Theta^T\|_F^2\\
&\quad+\lambda_\Theta\sum_{j=1}^{t}\|(\Theta^T)_j\|_1.
\end{aligned}$$

Solving $\mathbb{A}$:

$$\begin{aligned}&\arg\min_{\mathbb{A}\geq 0}\frac{1}{2}\|Q-\Theta\mathbb{A}\|_F^2+c_{\mathbb{A}}\|\mathbb{A}\|_F^2 \text{ s.t. } supp(\mathbb{A})\subseteq supp(A)\\&=\arg\min_{\mathbb{A}\geq 0}\|[\sqrt{0.5}Q;zeros(m,n)]-[\sqrt{0.5}\Theta;\sqrt{c_{\mathbb{A}}}eye(m,m)]\mathbb{A}\|_F^2,\\&\mathbb{A}_{ij}=0\forall i,j \text{ s.t. } \mathbb{A}_{ij}\notin supp(A).\end{aligned}$$

It is easy to see that coefficient matrices of the sub-problems for Φ, Q, Θ, $\mathbb{A}$ in (10) have full column rank ensuring local convergence.

Experimentally we observed that having very small values for $c_\Phi, c_\Theta, c_{\mathbb{A}}$ makes the coefficient matrices of the sub-problems to be of full column rank and gives approximately the same results as those obtained from (3).

But the same technique of adding weighted squared Frobenius norm of the matrices to (1) cannot be employed to Θ because of its middle position in (1) i.e., in $\|D-\Phi\Theta\mathbb{A}\|_F^2$. Hence local convergence is not guaranteed for (1).

References

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage–thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*(1), 183–202.

Bertsekas, D. P. (1999). *Nonlinear programming*. Belmont: Athena Scientific.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning r'esearch*, *3*, 993–1022.

Cai, D., Wang, X., & He, X. (2009). Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the 26th annual international conference on machine learning* (pp. 105–112). ACM.

Chen, X., Zhou, M., & Carin, L. (2012). The contextual focused topic model. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 96–104). ACM.

Ding, C., Li, T., Peng, W., & Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 126–135). ACM.

Griffiths, T. L., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. *NIPS*, *18*, 475–482.

Grippo, L., & Sciandrone, M. (2000). On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations Research Letters*, *26*(3), 127–136.

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

Kasiviswanathan, S. P., Melville, P., Banerjee, A., & Sindhwani, V. (2011). Emerging topic detection using dictionary learning. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 745–754). ACM.

Kim, D., Sra, S., & Dhillon, I. S. (2007). Fast Newton-type methods for the least squares nonnegative matrix approximation problem. In *SDM*. SIAM.

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

Kim, H., & Park, H. (2008a). Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, *30*(2), 713–730.

Kim, J., & Park, H. (2008b). Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Eighth IEEE international conference on data mining, 2008. ICDM'08* (pp. 353–362). IEEE.

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

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

Long, B., Zhang, Z. M., & Yu, P. S. (2005). Co-clustering by block value decomposition. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining* (pp. 635–640). ACM.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, *11*, 19–60.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111–126.

Powell, M. J. (1973). On search directions for minimization algorithms. *Mathematical Programming*, *4*(1), 193–201.

Rendle, S. (2010). Factorization machines. In *IEEE 10th international conference on data mining (ICDM)* (pp 995–1000). IEEE.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence* (pp. 487–494). AUAI Press.

Sachan, M., & Srivastava, S. (2013). Collective matrix factorization for co-clustering. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 93–94). International World Wide Web Conferences Steering Committee.

Sindhwani, V., & Ghoting, A. (2012). Large-scale distributed non-negative sparse coding and sparse dictionary learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 489–497). ACM.

Tibshirani, R. (1996). Regression selection and shrinkage via the LASSO. *Journal of the Royal Statistical Society Series B*, *58*(1), 267–288.

Williamsonm, S., Wang, C., Heller, K. A., & Blei, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 1151–1158).

Yang, J., & Zhang, Y. (2011). Alternating direction algorithms for L1-problems in compressive sensing. *SIAM Journal on Scientific Computing*, *33*(1), 250–278.

Zheng, G., Guo, J., Yang, L., Xu, S., Bao, S., Su, Z., et al. (2011). Mining topics on participations for community discovery. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 445–454). ACM.

Zhou, D., Li, J., & Zha, H. (2005). A new Mallows distance based metric for comparing clusterings. In *Proceedings of the 22nd international conference on machine learning* (pp. 1028–1035). ACM.

Zhuang, F., Luo, P., Shen, Z., He, Q., Xiong, Y., & Shi, Z. (2010) D-lda: a topic modeling approach without constraint generation for semi-defined classification. In *2010 IEEE 10th international conference on data mining (ICDM)* (pp. 709–718). IEEE.