The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Word Co-Occurrence Regularized Non-Negative Matrix Tri-Factorization for Text Data Co-Clustering

Aghiles Salah
School of Information Systems
Singapore Management University,
Singapore
asalah@smu.edu.sg

Melissa Ailem
University of Southern California, USA
& INRIA, France
melissa.ailem@inria.fr

Mohamed Nadif
LIPADE
University of Paris Descartes, France
mohamed.nadif@parisdescartes.fr

Abstract

Text data co-clustering is the process of partitioning the documents and words simultaneously. This approach has proven to be more useful than traditional one-sided clustering when dealing with sparsity. Among the wide range of co-clustering approaches, Non-Negative Matrix Tri-Factorization (NMTF) is recognized for its high performance, flexibility and theoretical foundations. One important aspect when dealing with text data, is to capture the semantic relationships between words since documents that are about the same topic may not necessarily use exactly the same vocabulary. However, this aspect has been overlooked by previous co-clustering models, including NMTF. To address this issue, we rely on the distributional hypothesis stating that words which co-occur frequently within the same context, e.g., a document or sentence, are likely to have similar meanings. We then propose a new NMTF model that maps frequently co-occurring words roughly to the same direction in the latent space to reflect the relationships between them. To infer the factor matrices, we derive a scalable alternating optimization algorithm, whose convergence is guaranteed. Extensive experiments, on several real-world datasets, provide strong evidence for the effectiveness of the proposed approach, in terms of co-clustering.

Introduction

Co-clustering is an important extension of traditional one-sided clustering that aims to partition the rows and columns of a data matrix simultaneously (Hartigan 1972; Banerjee et al. 2007; Govaert and Nadif 2013). This approach has been shown to be useful in many application domains such as text mining (Dhillon, Mallela, and Modha 2003; Salah and Nadif 2017; Ailem, Role, and Nadif 2017a) to group words and documents simultaneously, bioinformatics (Cheng and Church 2000; Madeira and Oliveira 2004; Cho et al. 2004) to group genes and experimental conditions simultaneously or to identify heterogeneous groups of two related entities (Pio et al. 2015), collaborative filtering (Hofmann and Puzicha 1999; Shan and Banerjee 2008; Deodhar and Ghosh 2010) to group users and items.

It turns out that, when dealing with high dimensional sparse data, such as text, co-clustering is more useful than traditional one-sided clustering even if we are interested in a clustering along one dimension only (Salah, Rogovschi, and Nadif 2016; Ailem, Role, and Nadif 2017b). In fact, co-clustering exhibits several advantages, over one-sided clustering: it leverages the inherent duality between the rows and columns of a data matrix, which makes it possible to improve clustering along both dimensions. It performs an implicit adaptive dimensionality reduction at each stage, which lends itself to scalable and effective algorithms for sparse data. It produces meaningful and interpretable results; in the case of document-word matrices, for example, co-clustering annotates sets of documents by clusters of words.

Among existing co-clustering approaches, Non-Negative Matrix Tri-Factorization (NTMF) has proven to be useful for this task (Long, Zhang, and Yu 2005; Ding et al. 2006). NMTF turns the co-clustering problem into a matrix approximation problem. In the context of text data, this approach seeks a decomposition of the document-word matrix $\mathbf{X}$ into three non-negative latent factor matrices, i.e., $\mathbf{X} \approx \mathbf{ZSW}^\top$, where $\mathbf{Z}$ and $\mathbf{W}$ play the role of the document and column cluster membership matrices, respectively, while $\mathbf{S}$ can be viewed as a "summary" of $\mathbf{X}$, due to co-clustering. Key advantages to consider NMTF for co-clustering are: high performance, efficiency, easy to implement inference and flexibility: NMTF can be easily composed to build more complex models and incorporate side information.

One important aspect when dealing with text data, is to preserve the semantic relationships between words. However, previous co-clustering algorithms, including NMTF, have overlooked this aspect. This is an important issue that may induce a significant loss of semantics, i.e., words with similar meanings are not guaranteed to have similar latent representations. Consequently, documents which are about the same topic, using similar (but not identical) words, are not guaranteed to be mapped to the same direction in the latent space. Clearly, this may severely impede the co-clustering performance of NMTF.

To address this issue, we propose to extend NMTF to account for the semantic relationships between words. The research question is how to capture and leverage such relationships. In this work, we rely on the distributional hypothesis (Harris 1954) stating that, words which co-occur frequently within the same "context" (e.g., a sentence, document, etc.) have similar meanings. Following this hypothesis, we propose to regularize the word factors in NMTF

based on the word co-occurrence matrix, in such a way to encourage frequently co-occurring words to have similar representations in the latent space. The word co-occurrence matrix encodes the number of times each pair of words occurred within the same context. The context is a design choice. Without loss of generality, in this work we use the documents as the context in which words co-occur. The objective of leveraging the relationships among words has been investigated recently in (Ailem, Salah, and Nadif 2017; Salah, Ailem, and Nadif 2017), but these works focused on one-sided clustering only.

To learn the factor matrices, we derive a scalable alternating optimization algorithm, whose convergence is guaranteed. Through extensive experiments, on several real-world datasets, we evaluate the performance of our model in terms of document clustering as well as word clustering. This is an important contribution over previous studies, in which co-clustering models are usually evaluated in terms of document partitioning, only. Empirical results show that, thanks to leveraging the word co-occurrences, our model preserves more semantics and, thereby, substantially outperforms previous NMTF models, on the co-clustering task.

Related Work

The literature on co-clustering is rich. Here we provide a brief overview of works that are most closely related to ours. For detailed surveys of existing co-clustering algorithms please refer to (Madeira and Oliveira 2004; Banerjee et al. 2007; Govaert and Nadif 2013).

NMTF-based co-clustering is pioneered by Long, Zhang, and Yu (2005) under the name of Block Value Decomposition. Ding et al. (2006) introduced orthogonality constraints on the document and word factors and emphasized their importance for the clustering task. More precisely, they shown theoretically that under these constraints, NMTF is equivalent to the simultaneous k-means co-clustering algorithm. To enforce these orthogonality constraints, Ding et al. (2006) rely on the Lagrangian method. Later on, Yoo and Choi (2010) proposed a more direct way to introduce such orthogonality constraints, by exploiting the true gradient information on Stiefel manifolds. Since these contributions, NMTF has been widely considered for co-clustering.

An important stream of efforts has focused on preserving the intrinsic geometry of the data (Gu and Zhou 2009; Shang, Jiao, and Wang 2012; Du and Shen 2013; Allab, Labiod, and Nadif 2017). These approaches construct two nearest neighbor graphs to encode the document manifold and the word manifold. The objective is then to seek document and word factors which are smooth with respect to these manifolds. Modeling document and word manifolds is an orthogonal direction to the present work; here we focus on preserving the semantic relationships between words. On the other hand, there have been efforts spent on developing scalable NMTF-based co-clustering algorithms (Wang et al. 2011) and, more recently, on incorporating side information, e.g., social network, into NMTF (Pei, Chakraborty, and Sycara 2015).

Our model is a novel contribution to existing work on NMTF, and as far as we know, this is the first that focuses on preserving the semantic relationships among words.

Notation. Matrices are denoted with boldface uppercase letters and vectors with boldface lowercase letters. The Frobenius norm is denoted by $\|.\|$ and the Hadamard multiplication by $\odot$. The document-word matrix is represented by $\mathbf{X} = (x_{ij}) \in \mathbb{R}_+^{n \times d}$, its i^{th} row represents the weighted term frequency vector of document $i \in \mathcal{I}$, i.e., $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top$ where $\top$ denotes the transpose. The word co-occurrence matrix is represented by $\mathbf{C} = (c_{jj'}) \in \mathbb{R}_+^{d \times d}$, where row $j \in \mathcal{J}$ corresponds to word w_j, column $j' \in \mathcal{J}$ denotes context word $w_{j'}$, and each entry $c_{jj'}$ encodes the number of times the pair of words $(w_j, w_{j'})$ occurred within the same context.

Word Co-occurrence Regularized NMTF

In this section, we describe Word Co-occurrence regularized NMTF (WC-NMTF), a new model that leverages the relationships among words to preserve more semantics and, consequently, improve co-clustering on text data.

In addition to the document-word matrix $\mathbf{X} \in \mathbb{R}_+^{n \times d}$, we consider the word co-occurrence matrix $\mathbf{C} \in \mathbb{R}_+^{d \times d}$ which encodes the number of times each pair of words occurred within the same context. To better quantify the associations between words, we further rely on a non-linear transformation of the word co-occurrences, based on the Point-wise Mutual Information (PMI). The PMI is an information theoretic measure widely used to quantify the association between pairs of outcomes arising from discrete random variables. Our choice for the PMI is motivated by previous studies (Newman, Karimi, and Cavedon 2009), which shown that this measure is highly correlated with human judgments, in assessing word relatedness. Formally, the PMI between words w_j and $w_{j'}$ is given by

$$\text{PMI}(w_j, w_{j'}) = \log \frac{p(w_j, w_{j'})}{p(w_j)p(w_{j'})}. \tag{1}$$

Given the word co-occurrence matrix $\mathbf{C}$, the PMI between w_j and $w_{j'}$ can be estimated empirically as follows

$$\text{PMI}(w_j, w_{j'}) = \log \frac{c_{jj'} \times c_{..}}{c_{j.} \times c_{.j'}} \tag{2}$$

where $c_{..} = \sum_{jj'} c_{jj'}$, $c_{j.} = \sum_{j'} c_{jj'}$ and $c_{.j'} = \sum_{j} c_{jj'}$.

Since it is intractable to work directly with the PMI matrix, which is dense and high-dimensional, we propose to approximate it with the Sparse Shifted Positive PMI matrix (SPPMI). In the rest of the paper, we use the notation $\mathbf{M} = (m_{jj'}) \in \mathbb{R}_+^{d \times d}$ to refer to the SPPMI matrix, where

$$m_{jj'} = \max\{\text{PMI}(w_j, w_{j'}) - \log(N), 0\}. \tag{3}$$

Where N is a hyperparameter that controls the sparsity of $\mathbf{M}$. With the SPPMI matrix $\mathbf{M}$ in place, we now give the formulation of our model.

WC-NMTF seeks a decomposition of $\mathbf{X} \in \mathbb{R}_+^{n \times d}$ into three low-dimensional non-negative latent factor matrices $\mathbf{Z} = (z_{ik}) \in \mathbb{R}_+^{n \times g}$, $\mathbf{W} = (w_{j\ell}) \in \mathbb{R}_+^{d \times m}$ and $\mathbf{S} = (s_{k\ell}) \in \mathbb{R}_+^{g \times m}$, such that $\mathbf{X} \approx \mathbf{ZSW}^T$. The factor matrices $\mathbf{Z}$ and $\mathbf{W}$ play the role of the row and column cluster membership

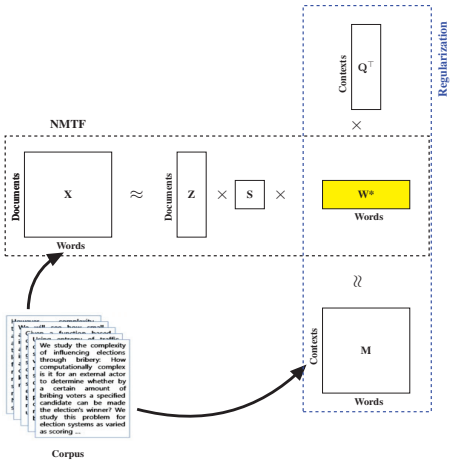


Figure 1: Illustrative scheme of WC-NMTF. $\mathbf{X} \approx \mathbf{Z}\mathbf{S}\mathbf{W}^*$ with $\mathbf{W}^* = \mathbf{W}^\top$ and $\mathbf{M} \approx \mathbf{W}^*\mathbf{Q}^\top$ with $\mathbf{W}^* = \mathbf{W}$.

matrices, whereas $\mathbf{S}$ plays the role of a summary of $\mathbf{X}$ due to co-clustering. In order to reflect the semantic relationships between words, WC-NMTF further assumes that, the word factors $\mathbf{W}$ are involved in factorizing the SPPMI matrix $\mathbf{M}$. Putting all this together, results in the objective function of WC-NMTF, $F = F(\mathbf{Z}, \mathbf{W}, \mathbf{S}, \mathbf{Q})$, given by

$$F = \underbrace{\frac{1}{2}||\mathbf{X} - \mathbf{Z}\mathbf{S}\mathbf{W}^\top||^2}_{\text{NMTF}} + \underbrace{\frac{\lambda}{2}||\mathbf{M} - \mathbf{W}\mathbf{Q}^\top||^2}_{\text{Regularization term}}, \tag{4}$$

where λ is the regularization parameter, and $\mathbf{Q} \in \mathbb{R}_+^{d \times m}$ is an extra factor (context factor) due to the decomposition of $\mathbf{M}$. We can interpret the above objective function as jointly decomposing the document-word and SPPMI matrices. The intuition behind the factorization of the SPPMI matrix $\mathbf{M}$, is to encourage words with similar meanings, those with high PMI, to have closer representations in the latent space. In doing so, WC-NMTF effectively preserves semantics, and can capture that documents which are about the same topic are similar even if they do not use exactly the same words. Figure 1 provides a graphical illustration of WC-NMTF.

Inference

In this section, we shall derive an alternating optimization algorithm to infer the latent factor matrices from the data. To this end, we rewrite (4), using the Trace operator, as follows

$$\begin{aligned} F &= \frac{1}{2}\operatorname{Tr}\left((\mathbf{X} - \mathbf{Z}\mathbf{S}\mathbf{W}^\top)(\mathbf{X} - \mathbf{Z}\mathbf{S}\mathbf{W}^\top)^\top\right) \\ &+ \frac{\lambda}{2}\operatorname{Tr}\left((\mathbf{M} - \mathbf{W}\mathbf{Q}^\top)(\mathbf{M} - \mathbf{W}\mathbf{Q}^\top)^\top\right) \\ &= \frac{1}{2}\operatorname{Tr}\left(\mathbf{X}\mathbf{X}^\top - 2\mathbf{X}\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top + \mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top\right) \\ &+ \frac{\lambda}{2}\operatorname{Tr}\left(\mathbf{M}\mathbf{M}^\top - 2\mathbf{M}\mathbf{Q}\mathbf{W}^\top + \mathbf{W}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^\top\right). \end{aligned}$$

In the following, we derive a set of multiplicative update rules in order to minimize F under the constraints of positivity of $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{S}$ and $\mathbf{Q}$. Let $\boldsymbol{\alpha} \in \mathbb{R}^{n \times g}$, $\boldsymbol{\beta} \in \mathbb{R}^{d \times m}$, $\boldsymbol{\mu} \in \mathbb{R}^{g \times m}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{d \times m}$ be the Lagrange multipliers for the constraints, the Lagrange function $L(\mathbf{Z}, \mathbf{W}, \mathbf{S}, \mathbf{Q}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\gamma}) = L$ is given by

$$F + \operatorname{Tr}(\boldsymbol{\alpha}\mathbf{Z}^\top) + \operatorname{Tr}(\boldsymbol{\beta}\mathbf{W}^\top) + \operatorname{Tr}(\boldsymbol{\mu}\mathbf{S}^\top) + \operatorname{Tr}(\boldsymbol{\gamma}\mathbf{Q}^\top).$$

The derivatives of L with respect to $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{S}$ and $\mathbf{Q}$ are

$$\frac{\partial L}{\partial \mathbf{Z}} = \mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top - \mathbf{X}\mathbf{W}\mathbf{S}^\top + \boldsymbol{\alpha} \tag{5a}$$

$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{W}(\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}} + \lambda\mathbf{Q}^\top\mathbf{Q}) - \mathbf{X}^\top\widetilde{\mathbf{Z}} - \lambda\mathbf{M}\mathbf{Q} + \boldsymbol{\beta} \tag{5b}$$

$$\frac{\partial L}{\partial \mathbf{S}} = \mathbf{Z}^\top\mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W} - \mathbf{Z}^\top\mathbf{X}\mathbf{W} + \boldsymbol{\mu} \tag{5c}$$

$$\frac{\partial L}{\partial \mathbf{Q}} = \lambda\mathbf{Q}\mathbf{W}^\top\mathbf{W} - \lambda\mathbf{M}^\top\mathbf{W} + \boldsymbol{\gamma}. \tag{5d}$$

where we introduced $\widetilde{\mathbf{Z}} = \mathbf{Z}\mathbf{S}$, for presentation purposes. Making use of the Kuhn-Tucker conditions $\boldsymbol{\alpha} \odot \mathbf{Z} = 0$, $\boldsymbol{\beta} \odot \mathbf{W} = 0$, $\boldsymbol{\mu} \odot \mathbf{S} = 0$ and $\boldsymbol{\gamma} \odot \mathbf{Q} = 0$ we obtain the following stationary equations:

$$(\mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top) \odot \mathbf{Z} - (\mathbf{X}\mathbf{W}\mathbf{S}^\top) \odot \mathbf{Z} = 0 \tag{6a}$$

$$\Big(\mathbf{W}(\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}} + \lambda\mathbf{Q}^\top\mathbf{Q}) - \mathbf{X}^\top\widetilde{\mathbf{Z}} - \lambda\mathbf{M}\mathbf{Q}\Big) \odot \mathbf{W} = 0 \tag{6b}$$

$$(\mathbf{Z}^\top\mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}) \odot \mathbf{S} - (\mathbf{Z}^\top\mathbf{X}\mathbf{W}) \odot \mathbf{S} = 0 \tag{6c}$$

$$(\mathbf{Q}\mathbf{W}^\top\mathbf{W}) \odot \mathbf{Q} - (\mathbf{M}^\top\mathbf{W}) \odot \mathbf{Q} = 0. \tag{6d}$$

Based on the above equations we derive the following multiplicative update rules

$$\mathbf{Z} \leftarrow \mathbf{Z} \odot \frac{\mathbf{X}\mathbf{W}\mathbf{S}^\top}{\mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top} \tag{7a}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{(\mathbf{X}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{M}\mathbf{Q})}{\mathbf{W}(\mathbf{S}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{Q}^\top\mathbf{Q})} \tag{7b}$$

$$\mathbf{S} \leftarrow \mathbf{S} \odot \frac{\mathbf{Z}^\top\mathbf{X}\mathbf{W}}{\mathbf{Z}^\top\mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}} \tag{7c}$$

$$\mathbf{Q} \leftarrow \mathbf{Q} \odot \frac{\mathbf{M}^\top\mathbf{W}}{\mathbf{Q}\mathbf{W}^\top\mathbf{W}}. \tag{7d}$$

Theorem 1. *(I) Fixing $\mathbf{S}$ and $\mathbf{W}$, the objective function of WC-NMTF (4) is non-increasing under the update formula (7a). (II) Fixing $\mathbf{Z}$ and $\mathbf{W}$, (4) is non increasing under the update formula (7c). (III) Fixing $\mathbf{W}$, (4) is non increasing under the update formula (7d). (IV) Fixing $\mathbf{Z}$, $\mathbf{S}$ and $\mathbf{Q}$, (4) is monotonically decreasing under the update rule (7b).*

Proof.

- (I) Absorbing $\mathbf{S}$ into $\mathbf{W}$, i.e., $\mathbf{W} \leftarrow \mathbf{W}\mathbf{S}^\top$, equation (7a) is similar to that of original NMF, therefore based on (Lee and Seung 2001) the objective function of WC-NMTF is non-increasing under this update rule.
- (II) Equation (7c) is identical to that of 3-factor NMF, thereby based on Theorem 7 in (Ding et al. 2006), the objective function (4) is non-increasing under equation (7c) for any fixed matrices $\mathbf{Z}$ and $\mathbf{W}$.
- (III) The update formula of $\mathbf{Q}$ (7d) is similar to that of NMF. Therefore, based on Theorem 1 in (Lee and Seung 2001), (III) also holds.

- (IV) Hence, it remains to demonstrate that (4) is non-increasing under the update rule of $\mathbf{W}$, for any fixed matrices $\mathbf{Z}$, $\mathbf{S}$ and $\mathbf{Q}$.

To prove (IV), we follow a similar approach to that described in (Lee and Seung 2001), which is inspired by the Expectation-Maximization (EM) algorithm and consists in using an auxiliary function.

Definition. *$G(w, w')$ is an auxiliary function for $F(w)$ if the following conditions are satisfied $G(w, w') \geq F(w)$ and $G(w, w) = F(w)$.*

A key point to the auxiliary function is described by the following lemma.

Lemma 1. If G is an auxiliary function for F, then F is non-increasing under the update

$$w^{(t+1)} = \underset{w}{\operatorname{argmin}}\, G(w, w^{(t)}). \tag{8}$$

Proof. $F(w^{(t+1)}) \leq G(w^{(t+1)}, w^{(t)}) \leq G(w^{(t)}, w^{(t)}) = F(w^{(t)})$.□

Next we make use of an appropriate auxiliary function to demonstrate that our objective function F is non-increasing under the update rule (7b). The following lemma yields an auxiliary function for F.

Lemma 2. The function G defined as follows

$$\begin{aligned} G(\mathbf{W}, \mathbf{W}^{(t)}) = &\frac{1}{2}\operatorname{Tr}\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{M}\mathbf{M}^\top\right) \\ &- \operatorname{Tr}\left(\mathbf{X}\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top + \lambda\mathbf{M}\mathbf{Q}\mathbf{W}^\top\right) \\ &+ \sum_{j,\ell} \frac{\left(\mathbf{W}^{(t)}\left(\mathbf{S}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{Q}^\top\mathbf{Q}\right)\right)_{j\ell}}{2w_{j\ell}^{(t)}} w_{j\ell}^2 \end{aligned} \tag{9}$$

is an auxiliary function for $\tilde{F}$.

Proof. It is obvious that $G(\mathbf{W}, \mathbf{W}) = F(\mathbf{W})$. Now we show that $G(\mathbf{W}, \mathbf{W}^{(t)}) \geq F(\mathbf{W})$, by making use of the following proposition (Ding et al. 2006).

Proposition 1. *Let $\mathbf{A} \in \mathbb{R}_+^{d\times d}$ and $\mathbf{B} \in \mathbb{R}_+^{m\times m}$ denote any symmetric matrices. For any matrices $\mathbf{H},\ \mathbf{H}' \in \mathbb{R}_+^{d\times m}$ the following inequality holds*

$$\sum_{j,\ell} \frac{(\mathbf{A}\mathbf{H}'\mathbf{B})_{j\ell} h_{j\ell}^2}{h'_{j\ell}} \geq \operatorname{Tr}\left(\mathbf{H}^\top\mathbf{A}\mathbf{H}\mathbf{B}\right). \tag{10}$$

Proof. The proof of the above proposition is available in (Ding et al. 2006), please refer to Proposition 6.

The first two terms in (9) are the same as in F. Based on Proposition 1, the following inequality holds

$$\begin{aligned} &\sum_{j,\ell} \frac{\left(\mathbf{W}^{(t)}\left(\mathbf{S}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{Q}^\top\mathbf{Q}\right)\right)_{j\ell}}{2w_{j\ell}^{(t)}} w_{j\ell}^2 \geq \\ &\operatorname{Tr}\left(\mathbf{W}^\top\mathbf{W}\left(\mathbf{S}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{Q}^\top\mathbf{Q}\right)\right). \end{aligned}$$

From the above inequality we have $G(\mathbf{W}, \mathbf{W}^{(t)}) \geq F(\mathbf{W})$, thereby $G(\mathbf{W}, \mathbf{W}^{(t)})$ is an auxiliary function of $F(\mathbf{W})$.□

Thus, to prove (IV) of Theorem 1 it is sufficient to show that equation (7b) for all $w_{j\ell}$ satisfies *Lemma 1*, where the auxiliary function G is given by *Lemma 2*. Substituting equation (9) to $G(\mathbf{W}, \mathbf{W}^{(t)})$ in *Lemma 1* leads to solve $\frac{\partial G(\mathbf{W}, \mathbf{W}^{(t)})}{\partial w_{j\ell}} = 0$, which gives us

$$(\mathbf{X}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{M}\mathbf{Q})_{j\ell} = \frac{\left(\mathbf{W}^{(t)}(\mathbf{S}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{Q}^\top\mathbf{Q})\right)_{j\ell}}{w_{j\ell}^{(t)}} w_{j\ell}.$$

Thus, $w_{j\ell}^{(t+1)} = \operatorname{argmin}_w G(w, w_{j\ell}^{(t)})$, $\forall j, \ell$, yields

$$w_{j\ell}^{(t+1)} = w_{j\ell}^{(t)} \frac{(\mathbf{X}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{M}\mathbf{Q})_{j\ell}}{\left(\mathbf{W}^{(t)}(\mathbf{S}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{Q}^\top\mathbf{Q})\right)_{j\ell}}.$$

It follows from the latter result and *Lemma 1* that F is non-increasing under the update of $w_{j\ell}$ in equation (7b), $\forall j, \ell$. Given that (7b) is element-wise, the objective function of WC-NMTF is non-increasing under update rule (7b). ■

Thereby, based on Theorem 1, the fact that (7a), (7b), (7d) and (7c) satisfy the KKT conditions at convergence, and F is bounded from below by 0, alternating the application of (7a), (7b), (7d) and (7c) will monotonically decrease criterion (4) and converge to a locally optimal solution.

Computational Complexity Analysis.

As stated in the following Proposition, the computational complexity of the `WC-NMTF` algorithm scales linearly with the number of non-zero entries in the document-word and SPPMI matrices. In practice $\mathbf{X}$ and $\mathbf{M}$ are very sparse, i.e., $nz_X \ll n \times d$ and $nz_M \ll d \times d$, and `WC-NMTF` converges within 100 iterations. Furthermore, multiplicative update rules (7a), (7b), (7c) and (7d) are trivially parallelizable, thereby `WC-NMTF` can easily scale to large datasets.

Proposition 2. *Let nz_X and nz_M denote respectively the number of non-zero entries in $\mathbf{X}$ and $\mathbf{M}$, and let it be the number of iterations. The computational complexity of* `WC-NMTF` *is given in $O(it{\cdot}g{\cdot}(nz_X{+}nz_M){+}it{\cdot}g^2{\cdot}(n{+}d))$.*

Proof. The computational bottleneck of `WC-NMTF` is with the multiplicative update formulas (7a), (7b), (7c) and (7d). The number of operation in (7b), including multiplications, additions and divisions, is $g(2nz_X + 2dm + g(2n + 2m)) + m(3nz_M + 3d + m(2d + 1 + 2g))$. The complexity of (7b) is thereby given in $O(g \cdot (nz_X + nz_M) + g^2(n + d))$, where have assumed that g and m are of the same order. Similarly, we can derive the complexities of (7a), (7c) and (7d), which are of order $O(g{\cdot}nz_X{+}g^2{\cdot}(n{+}d))$, $O(g{\cdot}nz_X{+}g^2{\cdot}(n{+}d))$ and $O(m \cdot nz_M + m \cdot d)$, respectively. Therefore, the total computational complexity of `WC-NMTF` is $O(it \cdot g \cdot (nz_X + nz_M) + it \cdot g^2 \cdot (n + d))$.■

Experiments

In this section, we study the impact of the word co-occurrences on the co-clustering task. For that purpose, we benchmark our model, WC-NMTF, against popular NMF and NMTF models on several real-world datasets.

Datasets. We use five popular benchmark datasets, described in Table 1, namely **CSTR** (Li 2005), **CLASSIC4**[1], the 20-newsgroups dataset **NG20**[2], and two datasets from the TREC collection[3], namely **TREC** and **LA Times**. These datasets are carefully selected to represent various particular challenging situations: different numbers of clusters, different sizes, different degrees of cluster overlap and different degrees of cluster balance. The Balance coefficient is the ratio of the minimum cluster size to the maximum cluster size.

Table 1: Description of Datasets.

Datasets	Characteristics					
	#Documents	#Words	#Clusters	$nz_X(\%)$	Balance	$nz_M(\%)$
CLASSIC4	7095	5896	4	0.59	0.323	2.41
CSTR	6387	16921	4	0.25	0.40	0.47
NG20	18846	26214	20	0.59	0.628	1.80
TREC	878	7453	10	2.61	0.037	15.84
LA Times	6279	31472	6	1.17	0.28	14.73

Competing methods. When $\lambda = 0$ in equation (4), WC-NMTF degenerates to the original NMTF (Long, Zhang, and Yu 2005). Thus, the best way to evaluate the effect of the word co-occurrences is to compare `WC-NMTF` to `NMTF` (Long, Zhang, and Yu 2005). Moreover, in order to show that leveraging the relationships among words is beneficial for text document clustering, we also consider several strong NMF and NMTF variants, namely the original `NMF` (Xu, Liu, and Gong 2003), orthogonal NMF (`ONMF`) (Yoo and Choi 2010), projective NMF (`PNMF`) (Yuan, Yang, and Oja 2009), graph regularized NMF (`GNMF`) (Cai et al. 2011), orthogonal NMTF (Ding et al. 2006) and graph regularized NMTF (`GNMTF`) (Shang, Jiao, and Wang 2012). All the above algorithms have been found to perform very well and better than several other approaches in terms of text document clustering. For the sake of completeness, we also integrate the results of spherical k-means algorithm `Skmeans` (Dhillon and Modha 2001), which is widely used for document clustering.

Settings. We use the TF-IDF weighting scheme for all datasets, and we normalize each document to unit L_2 norm so as to remove the biases induced by the length of documents. For each dataset, we set g to the real number of clusters and $m = g$ for the NMTF models. We perform 50 runs for all methods, using 50 different starting points obtained with `Skmeans`. The setting of the regularization parameter λ is trivial due to the high stability of `WC-NMTF` vis-a-vis this parameter. Hence, we set $\lambda = 1$ and discuss this choice in details later. The hyperparameter N in (3) controls the sparsity of the SPPMI matrix. A large value of N may cause to much loss of information, we therefore set $N = 2$, which is a good trade-off between keeping much information and increasing the sparsity of the PPMI matrix.

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

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

[3] http://trec.nist.gov

Evaluation metrics. We retain two widely used measures to assess the quality of clustering, namely the Normalized Mutual Information (NMI) (Strehl and Ghosh 2002) and the Adjusted Rand Index (ARI). Intuitively, NMI quantifies how much the estimated clustering is informative about the true clustering, while the ARI measures the degree of agreement between an estimated clustering and a reference clustering; both NMI and ARI are equal to 1 if the resulting clustering is identical to the true one.

Document clustering. In Table 2, we report the performance of the different models in terms of NMI and ARI, over all datasets. All the results are averaged over fifty different starting points. Between brackets, we report the result corresponding to the trial with the lowest criterion.

From this Table, we can see that our algorithm `WC-NMTF` provides substantially better results than the other competing methods, in terms of both metrics. It is worth recalling that `WC-NMTF` corresponds to `NMTF` with an extra word co-occurrences regularization term, which captures the semantic relationships between words. We can therefore grant the performance improvement of `WC-NMTF`, over `NMTF`, to the word co-occurrences regularization term.

To characterize the circumstances in which `WC-NMTF` provides the most significant improvement, we report in Figure 2 the distribution of the vocabulary sizes of the documents, over all datasets. From this Figure and Table 2 we observe that the improvement reached by `WC-NMTF`, relative to `NMTF`, tends to be more important on datasets containing many documents with a limited vocabulary, namely CSTR, CLASSIC4 and NG20. This make sense since when there is a lack of document-word information, it may be tough for `NMTF` to infer good document factors. In this case, the word co-occurrences in `WC-NMTF` play a preponderant role in supplementing the lack of document-word information.

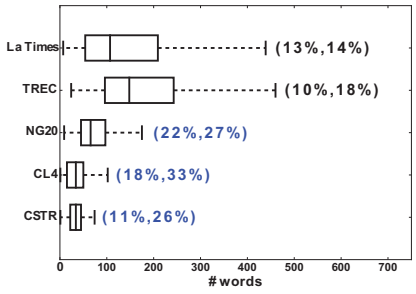


Figure 2: Distribution of the vocabulary sizes of the documents. For each dataset, we report the difference in performance (NMI%,ARI%) between `WC-NMTF` and `NMTF`.

Word clustering. To further characterize the effect of the word co-occurrences, we compare the quality of word clusters inferred by `WC-NMTF` and `NMTF`. This is an important improvement over most previous studies where co-clustering algorithms are often evaluated in terms of object (document) partitioning, only.

Evaluating the quality of word clusters is, however, a challenging task due to the lack of benchmark datasets provid-

Table 2: Average NMI and ARI over different datasets. The best result of each method is indicated in parentheses.

Datasets	Metrics	Skmeans	NMF	ONMF	PNMF	GNMF	NMTF	ONMTF	GNMTF	WC−NMTF
CSTR	NMI	0.65±0.02 (0.66)	0.65±0.01 (0.65)	0.65±0.05 (0.66)	0.66±0.01 (0.67)	0.57±0.08 (0.42)	0.67±0.03 (0.66)	0.65±0.03 (0.70)	0.64±0.04 (0.68)	**0.76±0.01 (0.77)**
	ARI	0.68±0.05 (0.72)	0.54±0.01 (0.55)	0.56±0.04 (0.58)	0.56±0.01 (0.59)	0.53±0.11 (0.39)	0.64±0.09 (0.56)	0.62±0.07 (0.74)	0.57±0.05 (0.67)	**0.81±0.01 (0.82)**
CLASSIC4	NMI	0.60±0.02 (0.59)	0.51±0.09 (0.52)	0.55±0.09 (0.50)	0.59±0.05 (0.51)	0.65±0.04 (0.57)	0.55±0.03 (0.58)	0.54±0.03 (0.59)	0.58±0.03 (0.62)	**0.72±0.06 (0.76)**
	ARI	0.47±0.01 (0.47)	0.36±0.10 (0.43)	0.39±0.09 (0.40)	0.44±0.01 (0.43)	0.49±0.05 (0.45)	0.44±0.01 (0.43)	0.43±0.01 (0.43)	0.42±0.03 (0.44)	**0.71±0.08 (0.76)**
NG20	NMI	0.48±0.02 (0.53)	0.43±0.01 (0.43)	0.44±0.02 (0.43)	0.45±0.02 (0.45)	0.52±0.01 (0.52)	0.40±0.02 (0.42)	0.41±0.01 (0.42)	0.43±0.02 (0.46)	**0.63±0.01 (0.64)**
	ARI	0.30±0.03 (0.34)	0.24±0.01 (0.25)	0.22±0.02 (0.21)	0.24±0.02 (0.28)	0.35±0.05 (0.37)	0.23±0.02 (0.23)	0.25±0.02 (0.27)	0.28±0.02 (0.32)	**0.47±0.02 (0.50)**
TREC	NMI	0.59±0.04 (0.62)	0.59±0.02 (0.60)	0.61±0.03 (0.60)	0.63±0.03 (0.62)	0.63±0.02 (0.62)	0.59±0.02 (0.60)	0.58±0.03 (0.60)	0.56±0.03 (0.61)	**0.67±0.03 (0.70)**
	ARI	0.42±0.06 (0.47)	0.43±0.04 (0.45)	0.45±0.04 0.44	0.46±0.05 (0.46)	0.50±0.04 (0.50)	0.43±0.03 (0.45)	0.42±0.04 (0.45)	0.36±0.04 (0.44)	**0.53±0.05 (0.63)**
LA Times	NMI	0.52±0.05 (0.54)	0.42±0.02 (0.44)	0.42±0.02 (0.45)	0.43±0.03 (0.46)	0.47±0.02 (0.50)	0.42±0.02 (0.44)	0.41±0.02 (0.42)	0.41±0.01 (0.42)	**0.53±0.03 (0.56)**
	ARI	0.49±0.05 (0.51)	0.36±0.04 (0.41)	0.35±0.06 (0.44)	0.37±0.06 (0.35)	0.43±0.03 (0.46)	0.35±0.04 (0.41)	0.34±0.04 (0.37)	0.34±0.03 (0.38)	**0.50±0.06 (0.55)**

ing the ground truth labels for words. Herein, we propose to evaluate the word clusters in terms of interpretability. To human subjects, interpretability is closely related to coherence (Newman et al. 2010), i.e., how much the top words of each cluster are "associated" with each other. To do so, for each word cluster ℓ, we select its top 30 words based on the ℓth column of $\mathbf{W}$. Then, to measure the degree of association between top word pairs, we use the PMI, which is highly correlated with human judgments (Newman, Karimi, and Cavedon 2009). Because WC−NMTF already exploits the PMI estimated from the word co-occurrences in each dataset, we propose to use an external corpus to estimate the PMI in this experiment. Following (Newman, Karimi, and Cavedon 2009), we use the whole English WIKIPEDIA corpus, that consists of approximately 4 millions of documents and 2 billions of words. Hence, $p(w_j)$ is the probability that word w_j occurs in WIKIPEDIA, and $p(w_j, w_{j'})$ is the probability that words w_j and $w_{j'}$ co-occur in a 5-word window in any WIKIPEDIA document. For each cluster we average the PMI's among its top word pairs, and for a model we average PMI across clusters.

Figure 3 shows the average PMI obtained by WC−NMTF and NMTF, over the different datasets. It is clear that WC−NMTF successes in capturing more semantics and inferring more interpretable word clusters than NMTF. For illustrative purposes and to support the results of Figure 3, we present in Table 3 the top words[4], characterizing the word clusters, inferred by WC−NMTF on the NG20 dataset; for each cluster we average the PMI's among its top word pairs.

Word representations. The key intuition behind our model, WC-NMTF, is to associate words having similar meanings (or which are about the same topic) with closer latent representations. Hereafter, we propose to verify this intuition empirically, by comparing the word representations inferred by WC−NMTF to those inferred by NMTF.

Figure 4 shows the distribution of pairwise cosine similarities between top words[5] characterizing the same "true" document class, using the word factors inferred by each method (over different values of m). We observe that the top words of each class (topic) tend to have closer latent representations under WC−NMTF, compared to NMTF, which validates the intuition behind WC-NMTF.

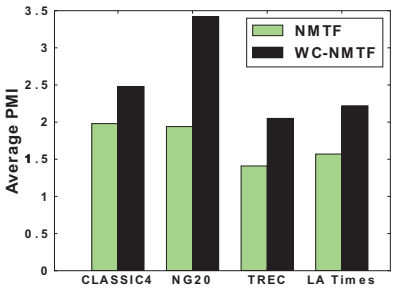


Figure 3: Cluster interpretability[6]: average PMI score. WC−NMTF leads more interpretable clusters than NMTF.

Impact of hyperparameters λ and m. Figure 5 shows the performance of WC−NMTF as a function of the regularization parameter λ. As it is clear from this Figure, WC−NMTF is highly stable relative to the variations of λ, and it seems to provide better performances when $\lambda \geq 0.01$, which facilitates the setting of this parameter.

In WC−NMTF, m plays the role of the number of column clusters and dimensions of the word latent space, simultaneously. Hence, setting m equal to the number of document clusters g—which is typically small—might seem not enough to infer good word representations reflecting various regularities between them. In other words, higher value of m may be expected to offer more capacity to encode the relationships among words and, thereby, improve document clustering even more.

Figure 6 depicts the performance of our model as a function of m. For each dataset, we vary m from the real number of document clusters g to 300. Surprisingly, smaller values of m yield slightly better performance in almost all cases. One possible explanation to this phenomena, is that when the value of m increases, we infer finer (or specialized) relationships among words that are not necessarily relevant for the clustering task. This suggests that, small values of m are enough to reflect topical/semantic relationships among

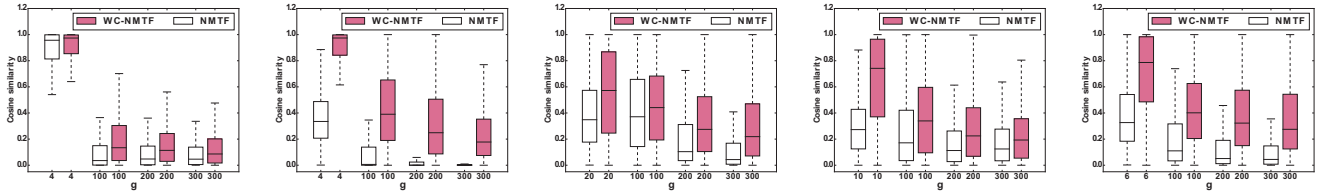
[4]The dataset we used contains stemmed words, so we have, for example, "allerg" instead of "allergy" in the table.

[5]The top 30 words of each true document class were obtained by keeping only words appearing in most documents of this class.

[6]The original corpus of CSTR is not available, this is the reason why we do not consider this dataset in this experiment.

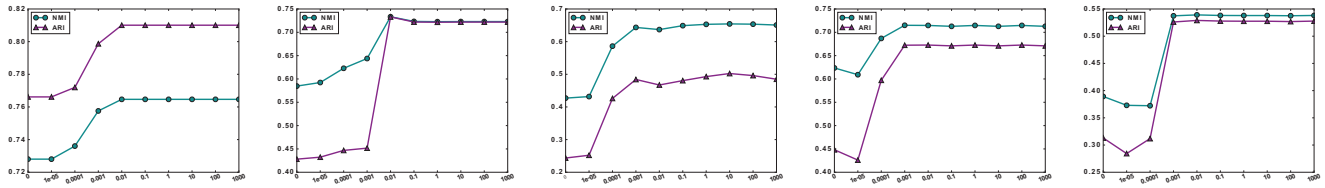
Table 3: Ten best word clusters (in terms of PMI) discovered by WC−NMTF on NG20, characterized by their top 20 words.

cluster (1)	cluster (2)	cluster (3)	cluster (4)	cluster (5)	cluster (6)	cluster (7)	cluster (8)	cluster (9)	cluster (10)
armenia ottoman	radio amplifi	treatment chronic	key crypto	hit obp	car wagon	window zip	team goal	god spirit	war jew
turk sahak	output pul	diseas nutrit	secur escrow	baseb clemen	ford compart	file download	hockei score	christian doctrin	peac isrel
extermin istanbul	input signal	patient yeast	clipper de	pitch winfield	auto callison	disk mswindow	play playoff	jesu holi	arab islam
argic kar	wire resistor	clinic diagnosi	encrypt privaci	bat 3b	toyota sedan	do instal	cup nhl	bibl passag	isr occupi
serdar vilayet	circuit shack	symptom ocom	secret classifi	pitcher pennent	rear camaro	util cica	detroit leaf	faith spiritu	territori palestinian
armenian arromdian	batteri transistor	diet candida	algorithm rsa	career outfield	mustang turbo	ms exe	montreal coach	christ profet	bosnia zionist
serazuma muratoff	frequenc capacitor	infect sinu	kei cryptograf	hitter bogg	tauru tbird	directori config	goali selann	teach testam	palestin policy
appressian ankara	audio khz	fysician vitamin	tap keyescrow	homer sandberg	torqu nissan	microsoft ini	puck lemieux	sin salvat	gaza iraq
melkonian caucasian	voltag diod	syndrom allerg	nsa decrypt	rbi shortstop	chevi munni	setup win3	potvin lafontain	lord unto	moham syria
ohanu hovannisian	amp volt	therapi fungal	wiretap cryptografi	catcher mjonesdonald	coup aas7po	app norton	defenseman penguin	scriptur psalm	arabia saudi
PMI: **4.43**	PMI: **3.83**	PMI: **3.78**	PMI: **3.71**	PMI: **3.70**	PMI: **3.69**	PMI: **3.21**	PMI: **3.09**	PMI: **3.03**	PMI: **3.01**



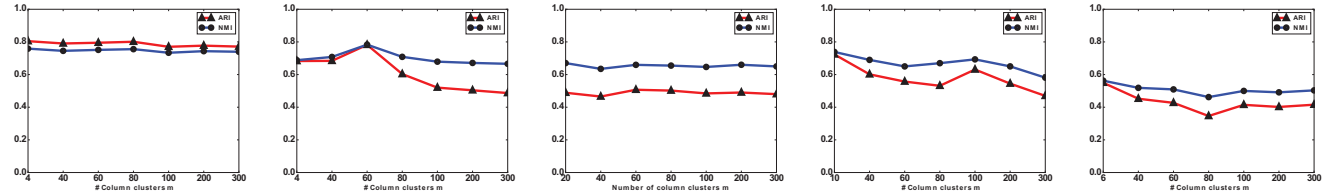
CSTR CLASSIC4 NG20 TREC LA Times

Figure 4: Distribution of pairwise cosine similarities between the top 30 words characterizing each document class, computed using the word factors obtained by NMTF and WC−NMTF.



CSTR CLASSIC4 NG20 TREC LA Times

Figure 5: Impact of the regularization parameter λ.



CSTR CLASSIC4 NG20 TREC LA Times

Figure 6: Impact of the number of column clusters m on NMI and ARI results.

words that are indeed relevant for the clustering task.

Conclusion

We propose to regularize the word factors in NMTF based on the word co-occurrences. This gives rise to a new co-clustering model, WC-NMTF, which aims to preserve the semantic relationships between words. The intuition behind our regularization scheme is to pull closer the latent representations of similar words. In doing so, WC-NMTF successfully preserves more semantics, which allows it to noticeably improve the performance of NMTF models in terms of both document and word clustering, as illustrated through extensive experiments.

Our work opens different avenues for future research. On one hand, the idea of leveraging the word co-occurrences to capture the semantic relationships between words can be extended to other co-clustering models. On the other hand, being flexible with solid theoretical foundations, WC-NMTF can be extended in several directions. For instance, it can be augmented to leverage the geometric structure of the document manifold. Such an extension is expected to yield further performance improvements, since it is well established that manifold regularization is beneficial for clustering. Furthermore, in this work we use the documents as the context

in which words occur. While this choice is effective, other type of contexts, e.g., sentences, deserve to be investigated.

References

Ailem, M.; Role, F.; and Nadif, M. 2017a. Model-based co-clustering for the effective handling of sparse data. *Pattern Recognition* 72:108–122.

Ailem, M.; Role, F.; and Nadif, M. 2017b. Sparse poisson latent block model for document clustering. *IEEE TKDE* 29(7):1563–1576.

Ailem, M.; Salah, A.; and Nadif, M. 2017. Non-negative matrix factorization meets word embedding. In *ACM SIGIR*, 1081–1084.

Allab, K.; Labiod, L.; and Nadif, M. 2017. Multi-manifold matrix decomposition for data co-clustering. *Pattern Recognition* 64:386–398.

Banerjee, A.; Dhillon, I. S.; Ghosh, J.; Merugu, S.; and Modha, D. S. 2007. A generalized maximum entropy approach to bregman co-clustering and matrix approximations. *J. Mach. Learn. Res.* 8:1919–1986.

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph regularized nonnegative matrix factorization for data representation. *IEEE TPAMI* 33(8):1548–1560.

Cheng, Y., and Church, G. M. 2000. Biclustering of expression data. In *ISMB*, volume 8, 93–103.

Cho, H.; Dhillon, I. S.; Guan, Y.; and Sra, S. 2004. Minimum sum-squared residue co-clustering of gene expression data. In *SDM*, 114–125.

Deodhar, M., and Ghosh, J. 2010. Scoal: a framework for simultaneous co-clustering and learning from complex data. *ACM TKDD* 4(3):11:1–11:31.

Dhillon, I. S., and Modha, D. S. 2001. Concept decompositions for large sparse text data using clustering. *Mach. Learn.* 42(1-2):143–175.

Dhillon, I. S.; Mallela, S.; and Modha, D. S. 2003. Information-theoretic co-clustering. In *SIGKDD*, 89–98.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *SIGKDD*, 126–135.

Du, L., and Shen, Y.-D. 2013. Towards robust co-clustering. In *IJCAI*, 1317–1322.

Govaert, G., and Nadif, M. 2013. *Co-clustering*. John Wiley & Sons.

Gu, Q., and Zhou, J. 2009. Co-clustering on manifolds. In *SIGKDD*, 359–368.

Harris, Z. S. 1954. Distributional structure. *Word* 10(2-3):146–162.

Hartigan, J. A. 1972. Direct clustering of a data matrix. *Journal of the american statistical association* 67(337):123–129.

Hofmann, T., and Puzicha, J. 1999. Latent class models for collaborative filtering. In *IJCAI*, volume 99, 688–693.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *NIPS*, 556–562.

Li, T. 2005. A general model for clustering binary data. In *SIGKDD*, 188–197.

Long, B.; Zhang, Z. M.; and Yu, P. S. 2005. Co-clustering by block value decomposition. In *SIGKDD*, 635–640.

Madeira, S. C., and Oliveira, A. L. 2004. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM TCBB* 1(1):24–45.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *NAACL HLT*, 100–108.

Newman, D.; Karimi, S.; and Cavedon, L. 2009. External evaluation of topic models. In *Australasian Doc. Comp. Symp.*

Pei, Y.; Chakraborty, N.; and Sycara, K. 2015. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *IJCAI*, 2083–2089.

Pio, G.; Ceci, M.; Malerba, D.; and D'Elia, D. 2015. Comirnet: a web-based system for the analysis of mirna-gene regulatory networks. *BMC bioinformatics* 16(9):S7.

Salah, A.; Ailem, M.; and Nadif, M. 2017. A way to boost semi-nmf for document clustering. In *CIKM*, 2275–2278.

Salah, A., and Nadif, M. 2017. Model-based von mises-fisher co-clustering with a conscience. In *SDM*, 246–254.

Salah, A.; Rogovschi, N.; and Nadif, M. 2016. Model-based co-clustering for high dimensional sparse data. In *AISTATS*, 866–874.

Shan, H., and Banerjee, A. 2008. Bayesian co-clustering. In *IEEE ICDM*, 530–539.

Shang, F.; Jiao, L.; and Wang, F. 2012. Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition* 45(6):2237–2250.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3(Dec):583–617.

Wang, H.; Nie, F.; Huang, H.; and Makedon, F. 2011. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *IJCAI*, 1553–1558.

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *SIGIR*, 267–273.

Yoo, J., and Choi, S. 2010. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information processing & management* 46(5):559–570.

Yuan, Z.; Yang, Z.; and Oja, E. 2009. Projective nonnegative matrix factorization: Sparseness, orthogonality, and clustering. *Neural Process. Lett*.