JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 1609–1624: Applications and Case Studies
https://doi.org/10.1080/01621459.2020.1862667

Topic Modeling on Triage Notes With Semiorthogonal Nonnegative Matrix Factorization

Yutong Li[a], Ruoqing Zhu[a], Annie Qu[b], Han Ye[c], and Zhankun Sun[d]

[a]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL; [b]Department of Statistics, University of California, Irvine, Irvine, CA; [c]Gies College of Business, University of Illinois at Urbana-Champaign, Champaign, IL; [d]College of Business, City University of Hong Kong, Kowloon, Hong Kong

ABSTRACT

Emergency department (ED) crowding is a universal health issue that affects the efficiency of hospital management and patient care quality. ED crowding frequently occurs when a request for a ward-bed for a patient is delayed until a doctor makes an admission decision. In this case study, we build a classifier to predict the disposition of patients using manually typed nurse notes collected during triage as provided by the Alberta Medical Center. These predictions can potentially be incorporated to early bed coordination and fast track streaming strategies to alleviate overcrowding and waiting times in the ED. However, these triage notes involve high dimensional, noisy, and sparse text data, which make model-fitting and interpretation difficult. To address this issue, we propose a novel semiorthogonal nonnegative matrix factorization for both continuous and binary predictors to reduce the dimensionality and derive word topics. The triage notes can then be interpreted as a non-subtractive linear combination of orthogonal basis topic vectors. Our real data analysis shows that the triage notes contain strong predictive information toward classifying the disposition of patients for certain medical complaints, such as altered consciousness or stroke. Additionally, we show that the document-topic vectors generated by our method can be used as features to further improve classification accuracy by up to 1% across different medical complaints, for example, 74.3%–75.3% accuracy for patients with stroke symptoms. This improvement could be clinically impactful for certain patients, especially when the scale of hospital patients is large. Furthermore, the generated word-topic vectors provide a bi-clustering interpretation under each topic due to the orthogonal formulation, which can be beneficial for hospitals in better understanding the symptoms and reasons behind patients' visits. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2019
Accepted December 2020

KEYWORDS
Dimension reduction; Emergency department crowding; Matrix factorization; Text mining; Topic modeling

1. Introduction

Emergency department (ED) crowding is an international phenomenon frequently faced by emergency physicians, nurses, and patients. Typically, a request for an admit ward-bed and preparations to receive the patient may be delayed until a doctor makes an admission decision (Morley et al. 2018). Studies have shown that ED crowding is associated with an increased risk of mortality, longer wait time and length of stay, and patient dissatisfaction (Chalfin et al. 2007; Sun et al. 2013). Existing literature suggests that if the hospital admissions of ED patients can be predicted early, or even before triage, then necessary steps can be taken to reduce the overcrowding and wait time of ED (Peck et al. 2012; Qiao 2015; Morley et al. 2018). The predicted information can be passed on to the target inpatient ward departments, where staff can begin their preparations early on and consequently reduce patient transfer delays and boarding.

Studies have been done to investigate potential solutions to reduce ED crowding (Morley et al. 2018), including modifying existing hospital administrative policies (Anantharaman 2008), declaring nurse-initiated protocols (Douma et al. 2016), and proposing alternative bed-management strategies (Barrett, Ford, and Ward-Smith 2012). Alternatively, machine learning approaches have begun to attract the interest of researchers within this field (Sun et al. 2011; Zhang et al. 2017; Hong, Haimovich, and Taylor 2018). In particular, Zhang et al. (2017) predicted ED admission outcomes by considering the effects of triage notes. These triage notes are manually typed into a computer by nurses, according to the patient's description during an ED visit. Zhang et al. (2017) applied a bag-of-words approach and showed the potential of using triage notes in predicting disposition. However, this simplistic approach failed to capture the underlying patterns and interactions between words. Additionally, the bag-of-words representation of text data is usually noisy, high dimensional, and sparse. This is due to the small number of words contained in each document compared to the large number of possible unique words (Salton, Wong, and Yang 1975). Hence, we consider a topic modeling approach for the text data collected from patients to improve the interpretation of the notes and the prediction performance of subsequent supervised learning methods (Aggarwal and Reddy 2013; Yaram 2016). Furthermore, our work is applicable for the

CONTACT Annie Qu aqu2@uci.edu Department of Statistics, University of California, Irvine, Irvine, CA 92627.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

current COVID-19 pandemic as ED crowding could be fatal for patients who were delayed from proper treatments (Mareiniss 2020).

The nonnegative matrix factorization (NMF) has drawn much attention due to its simplicity and interpretability. Its applications include image analysis (Lee and Seung 1999), cluster analysis (Kim and Park 2008), and text mining (Shahnaz et al. 2006). The purpose of the NMF is to uncover nonnegative latent factors and relationships to provide meaningful interpretations for practical applications, such as this triage data. The NMF was first studied by Paatero and Tapper (1994) as positive matrix factorization, and was widely adopted due to Lee and Seung's (1999, 2001) work. Specifically, the NMF seeks to approximate a matrix $\mathbf{X}$ as a product of two lower-rank nonnegative matrices, $\mathbf{F}$ and $\mathbf{G}$. For the text mining application in this article, $\mathbf{X}$ refers to the word-by-document matrix (bag-of-words), $\mathbf{F}$ refers to the word-topic matrix, and $\mathbf{G}$ refers to the document-topic matrix.

In this article, we aim to classify patients' disposition using a triage notes dataset provided by the Alberta Health Services, the largest integrated provincial health care system in Canada. Additionally, we aim to understand the main reasons behind patients' visits and eventual dispositions, which is also essential for the hospital. This dataset contains around 500,000 anonymous patient records collected from September 2014 to August 2016, each with a medical complaint, their disposition at the ED, and text information regarding the reason for the visit. Each record was input by the nurse according to the description by the patient. Additional information, including the demographic and vital signs of the patients, were also provided. However, these features are not within the scope of this study, and would not be considered in this analysis.

To analyze the triage dataset via a topic modeling approach, we propose a novel semiorthogonal nonnegative matrix factorization (SONMF) under the framework of both continuous and binary matrices. Our model factorizes a target matrix into the product of an orthogonal matrix $\mathbf{F}$ and a nonnegative matrix $\mathbf{G}$. As opposed to the existing orthogonal NMF methods (Ding et al. 2006; Yoo and Choi 2008), our model does not enforce nonnegativity on $\mathbf{F}$. This formulation provides an alternative and meaningful interpretation of the word-topic vectors while retaining the by-parts interpretation of the document-topic vectors. We show that this formulation yields basis topic vectors with uncorrelated loadings, which subsequently generates interpretable topics with distinct meanings. The mixed signs within the word-topic matrix introduce further sub-clusters within each word-topic vector, which are negatively correlated and have opposite meanings.

Numerically, our model can achieve strict orthogonality due to the removal of nonnegativity on $\mathbf{F}$, as opposed to the approximate solutions in existing literature (Ding et al. 2006; Yoo and Choi 2008; Kimura, Tanaka, and Kudo 2015). The enforcement of exact orthogonality also serves as a regularization by shrinking the model space. It eliminates multicollinearity between the basis vectors, and subsequently reduces the risk of overfitting (Abdi and Williams 2010; Wang et al. 2019). This has advantages for both increasing the classification performance of subsequent supervised learning approaches by using the generated topic vectors as new features, and the interpretation of these topic vectors themselves.

The article is organized as follows. Section 2 discusses the motivation of this study while Section 3 briefly reviews the NMF. The proposed method for both the continuous and binary cases are presented in Section 4. Section 5 provides a set of numerical experiments on simulated datasets, and Section 6 focuses on the in-depth analysis and discussion of the triage dataset. The conclusions of this study are presented in Section 7.

2. Motivation

To improve the management of ED patients, existing studies (Powell et al. 2012; Qiu et al. 2015) have considered initiating early bed requests at the triage stage or soon after based on predicted admissions. The goal is to start the bed coordination of the intensive unit (IU) when a patient is still waiting or being treated in the ED. This can significantly reduce the overall waiting and length of stay (LoS) of admitted patients. Such early bed coordination strategies rely on the admission predictions with data collected from the triage stage. Some studies (Ieraci et al. 2008; Saghafian et al. 2012) explore the advantages of streaming patients based on the acuity or predicted disposition of patients. A fast track may be created as a separate stream from the ED main waiting queue, with designated doctors treating the less acute patients. This can reduce the waiting time before treatment for all ED patients, and thus mitigate ED crowding. Predicting patient discharges at the triage stage can contribute to improve the designs of ED streaming strategies.

Early bed coordination and fast track streaming strategies are illustrated in Figure 1. The two flow charts illustrate the impacts of the aforementioned strategies to improve ED operations on patient waiting and LoS. In particular, the LoS of discharged patients may be reduced from T_2 to T_1 with a fast track. The LoS of admitted patients may be reduced from T_4 to T_3 with early bed coordination. Reductions of the LoS can also shorten the waiting time before treatment. For these operational strategies to successfully improve ED patient flow, accurately predicting patient admission/discharge at the triage stage is an essential step. This motivates our research to implement new methods to utilize the rich text data in the triage notes to achieve our goal.

Besides building a classifier for predicting the disposition, we also want to understand the main complaints and symptoms of the patients visiting the ED. Solely using the words from the triage notes as features can achieve the first goal, but is insufficient to address the second objective, as important information may be overlooked if the underlying interactions between words are ignored. Thus, it is natural for us to consider word clustering/topic modeling as a feature extraction method to uncover latent semantic information.

The NMF has been shown to be effective in document clustering and topic modeling applications (Shahnaz et al. 2006; Aggarwal and Reddy 2013). The nonnegative enforcement of the NMF naturally captures the structure of a word-document matrix (Salton, Wong, and Yang 1975) with a by-parts interpretation. Given a nonnegative word-document matrix $\mathbf{X}_{p\times n}$, the NMF factorizes $\mathbf{X}_{p\times n}$ into two lower rank-k nonnegative matrices $\mathbf{F}_{p\times k}$ and $\mathbf{G}_{n\times k}$ (i.e., $\mathbf{X}_{p\times n} \approx \mathbf{F}_{p\times k}\mathbf{G}_{n\times k}^T$), where k is the rank and serves as the number of topics/clusters. $\mathbf{F}$ can be interpreted as the word-topic matrix, where the words with

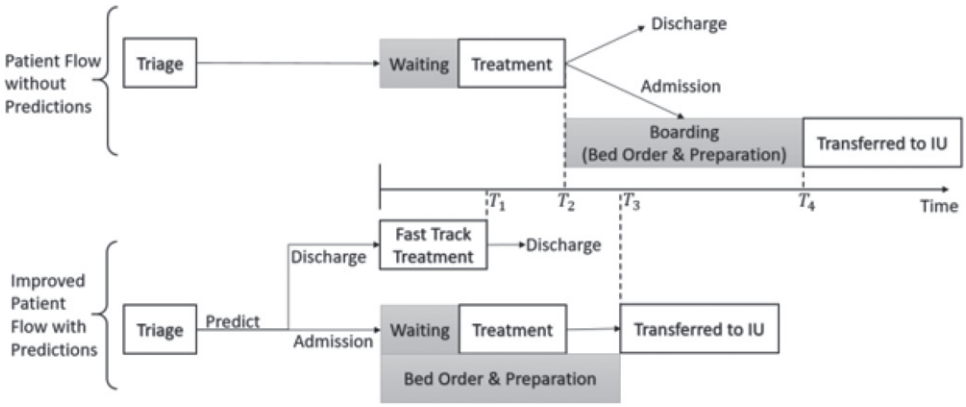

Figure 1. ED patient flow with and without streaming using predictions at triage.

the largest weights within each topic approximately define the topic's meaning. **G** is regarded as the document-topic matrix, where each document points in the direction of the topics with various magnitudes.

However, most NMF methods do not consider the redundancy of features due to the absence of orthogonal constraints. This leads to correlated topic vectors, which could impair the classification performances of subsequent model fitting and make the topic vectors less interpretable. This problem is more prevalent in the semi-NMF (Ding, Li, and Jordan 2010), where the nonnegativity constraint on the **F** matrix is no longer imposed. This results in the potential cancellation of terms across different word-topics due to different signs, and could increase the risk of overfitting. Therefore, in this article, we enforce the **F** matrix to be strictly orthogonal to address this issue, which subsequently yields orthogonal word-topic vectors. Generally, any constraint added to the model would result in a smaller model space. In our case, the solution set that satisfies $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ is smaller than that of the unconstrained formulation, which naturally serves as an regularization to prevent overfitting (Hastie, Tibshirani, and Friedman 2009).

Orthogonal topic vectors have been previously considered by Ding et al. (2006). They enforce both nonnegativity and orthogonality on the word-topic matrix **F**, implying that each word can only belong to a single topic as each row can only have 1 nonzero element. However, this assumption could be too restrictive, as most words naturally have multiple or ambiguous meanings, and possibly belong to multiple topics. This is also supported by checking the distribution of the number of topics each word belongs to (see Section A.5 of the Appendix). Thus, removing nonnegativity from the orthogonal matrix can preserve the interpretation of uncorrelated topic vectors, while retaining the flexibility of words falling into multiple topics.

In addition, representing the topics as a linear combination of positive and negative terms (words) also leads to a different interpretation compared to the conventional by-parts representation of the NMF. The positive loading of a word indicates that the word belongs to a topic with positive strength, while a negative loading represents the deviation of the word from the topic. The words with large positive weights under a topic not only indicate that they are the most representative of this topic but also imply that these words tend to appear together and are highly correlated with each other. On the other hand, the words with negative weights indicate that these words are negatively correlated with this topic and can be viewed as separate clusters, effectively serving as the acronyms of the positive words within the same topic. This naturally creates a bi-clustering structure within a topic, in contrast to the zero representation of words in the nonnegative case. This enables us to identify the main reasons for patients' visits, and potentially understand the factors that the patient is discharged upon by looking at the words with the largest negative loadings under a topic. This can be beneficial for providing additional insights to the management plans of a hospital for patient admission.

From a computational perspective, existing algorithms are either based on multiplicative updates (Lee and Seung 1999; Lin 2007; Ding, Li, and Jordan 2010), least squares update (Cichocki, Zdunek, and Amari 2007; Kimura, Tanaka, and Kudo 2015), or through approximate separability (Recht et al. 2012; Gillis and Vavasis 2013). Multiplicative updates typically run into the zero-lock problem (Cichocki, Zdunek, and Amari 2007) when forcing nonnegativity. This occurs when zero elements in either **F** and **G** remain zero throughout the entire optimization process, which causes these elements to be locked at zero and thus prevents updating to an optimal solution. On the other hand, the assumption of separability may not be suitable under our orthogonal formulation. The hierarchical based alternating least squares update scheme (Cichocki, Zdunek, and Amari 2007; Cichocki and Phan 2009) is not affected by the above limitations and is thus incorporated into our model.

The current orthogonal NMF algorithms do not yield strict orthogonal solutions due to the nonnegative constraint as algorithms need to sacrifice orthogonality to conserve the nonnegativity (Ding et al. 2006; Yoo and Choi 2008; Kimura, Tanaka, and Kudo 2015). By removing the nonnegative constraint on the orthogonal matrix, our model can achieve a strictly orthogonal solution ($\mathbf{F}^T\mathbf{F} = \mathbf{I}$) for **F** rather than an approximated solution ($\mathbf{F}^T\mathbf{F} \approx \mathbf{I}$). This can be achieved with an SVD-based initialization and a novel implementation of an orthogonal preserving update scheme (Wen and Yin 2013) through the Stiefel manifold. The orthogonal preserving update scheme (Wen and Yin 2013) was effective in other applications, but so far has not been implemented in solving NMF-related problems. On the other hand, Yoo and Choi (2008) derived their multiplicative

update rules for the orthogonal-NMF with consideration of the Stiefel manifold, but their solution still deviates from exact orthogonality due to the additional constraint of nonnegativity.

We then implement the second-order least squares update (Cichocki and Phan 2009) for $\mathbf{G}$ with nonnegative projection to avoid zero-locking. Interestingly, this update can be reduced to a simple form due to the strict orthogonality of $\mathbf{F}$. This also ensures our algorithm converges quickly and leads to numeric stability consistently.

3. Notations and Background of the NMF

In this section, we provide the notations and background of the NMF. Let $\mathbf{X}$ be an $p \times n$ real matrix, and $\mathbf{x}_j$ be the jth column, that is, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_j]$. Nonnegative matrix factorization (Lee and Seung 1999) aims to factorize a nonnegative matrix $\mathbf{X}$ into the product of two nonnegative matrices, $\mathbf{F}$ and $\mathbf{G}$:

$$\begin{aligned} &\text{argmin}_{\mathbf{F},\mathbf{G}} \ ||\mathbf{X} - \mathbf{F}\mathbf{G}^T||_F^2, \\ &\text{subject to} \quad \mathbf{F} \geq 0, \mathbf{G} \geq 0, \end{aligned} \tag{1}$$

where $||\cdot||_F$ is the Frobenius norm. Typically, $\mathbf{F}$ and $\mathbf{G}$ are lower ranks, for example, $\mathbf{F} \in \mathbb{R}^{p\times k}$ and $\mathbf{G} \in \mathbb{R}^{n\times k}$, where $k \ll \min(n, p)$. More specifically, the columns of $\mathbf{X}$ can be rewritten as $\mathbf{x}_{p\times 1} \approx \mathbf{F}_{p\times k}\mathbf{g}_{k\times 1}^T$, where $\mathbf{x}$ and $\mathbf{g}$ are the corresponding columns for $\mathbf{X}$ and $\mathbf{G}$. Thus, each column vector $\mathbf{x}$ is approximated as a linear combination of $\mathbf{F}$, weighted by the rows of $\mathbf{G}$, or equivalently, $\mathbf{F}$ can be regarded as the matrix that consists of the basis vectors for the linear approximation of $\mathbf{X}$.

The above problem in (1) can be solved by alternating the updates between $\mathbf{F}$ and $\mathbf{G}$ while fixing the other via an matrix-wise alternating block coordinate descent scheme (Lee and Seung 1999, 2001; Ding et al. 2006; Yoo and Choi 2008; Ding, Li, and Jordan 2010). In Lee and Seung (1999, 2001), $\mathbf{F}$ and $\mathbf{G}$ are updated by multiplying the current value with an adaptive factor that depends on the rescaling of the gradient of (1):

$$\mathbf{F}_{ik} \leftarrow \mathbf{F}_{ik}\frac{(\mathbf{X}\mathbf{G})_{ik}}{(\mathbf{F}\mathbf{G}^T\mathbf{G})_{ik}} \quad \text{and} \quad \mathbf{G}_{ik} \leftarrow \mathbf{G}_{ik}\frac{(\mathbf{X}^T\mathbf{F})_{ik}}{(\mathbf{G}\mathbf{F}^T\mathbf{F})_{ik}}. \tag{2}$$

The NMF can easily be extended by incorporating additional constraints on the factor matrices, such as the sparse NMF (Hoyer 2004), orthogonal-NMF (Ding et al. 2006), and semi-NMF (Ding, Li, and Jordan 2010). On the other hand, computational efficient factorization algorithms also play a major role in current NMF research (Cichocki, Zdunek, and Amari 2007; Cichocki and Phan 2009; Hsieh and Dhillon 2011; Gillis and Vavasis 2013).

4. Methodology

In this section, we present the derivation and implementation of the proposed method in both continuous and binary settings. Although both settings serve the same purpose in reducing a matrix into a lower rank representation, the inherent structure of a binary matrix requires a different optimization approach. We first present the continuous case in Section 4.1, and then the binary case in Section 4.2. The initialization, convergence criterion and proposed algorithms are presented in Section 4.3.

4.1. The SONMF for Continuous Matrix

Consider the following matrix factorization problem with a cost function denoted as $C(\mathbf{F}, \mathbf{G})$,

$$\begin{aligned} &\text{argmin}_{\mathbf{F},\mathbf{G}} \ C(\mathbf{F}, \mathbf{G}) = \text{argmin}_{\mathbf{F},\mathbf{G}} ||\mathbf{X} - \mathbf{F}\mathbf{G}^T||_F^2, \\ &\text{subject to } \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{G} \geq 0, \end{aligned}$$

where $\mathbf{X} \in \mathbb{R}^{p\times n}$, $\mathbf{F} \in \mathbb{R}^{p\times k}$, and $\mathbf{G} \in \mathbb{R}^{n\times k}$. We solve this problem by alternatively updating the matrices $\mathbf{F}$ and $\mathbf{G}$. However, the uniqueness of the proposed method is to take advantage of the Stiefel manifold $\mathcal{M}_n^p$, where $\mathcal{M}_n^p$ is the feasible set $\{\mathbf{F} \in \mathbb{R}^{p\times k} : \mathbf{F}^T\mathbf{F} = \mathbf{I}\}$. Following the formulation in Wen and Yin (2013), we initialize $\mathbf{F}$ as a column-wise orthonormal matrix, then enforces the solution path of $\mathbf{F}$ to be exactly on this manifold, thereby preserving strict orthogonality throughout the entire optimization process.

The update scheme is an adaptation of the gradient descent, but preserves the orthogonality at a reasonable computational cost. Under the matrix representation, the gradient of $\mathbf{F}$ is $\nabla\mathbf{F} = \frac{\partial C}{\partial \mathbf{F}} = 2\mathbf{F}\mathbf{G}^T\mathbf{G} - 2\mathbf{X}\mathbf{G}$, derived directly from the objective function without any constraints. However, the new update $\mathbf{F}_{n+1} = \mathbf{F}_n - \tau\nabla\mathbf{F}_n$ may not satisfy $\mathbf{F}_{n+1} \in \mathcal{M}_n^p$, where τ is a step size for the line search. Instead, we need to first project $(-\nabla\mathbf{F})$ onto the tangent space of $\mathcal{M}_n^p$ at $\mathbf{F}$. To do so, we first use $\mathbf{F}$ and $\nabla\mathbf{F}$ to define a skew-symmetric matrix $\mathbf{S} = (\nabla\mathbf{F})\mathbf{F}^T - \mathbf{F}(\nabla\mathbf{F})^T$. Next, we apply the Cayley transformation to yield an orthogonal matrix $\mathbf{Q} = (\mathbf{I} + \frac{\tau}{2}\mathbf{S})^{-1}(\mathbf{I} - \frac{\tau}{2}\mathbf{S})$. The $\mathbf{F}$ matrix can then be updated via $\mathbf{F}_{n+1} = \mathbf{Q}\mathbf{F}_n$. Since $\mathbf{Q}$ is an orthogonal matrix, we have $\mathbf{F}_{n+1}^T\mathbf{F}_{n+1} = (\mathbf{Q}\mathbf{F}_n)^T(\mathbf{Q}\mathbf{F}_n) = \mathbf{F}_n^T\mathbf{Q}^T\mathbf{Q}\mathbf{F}_n = \mathbf{F}_n^T\mathbf{F}_n = \mathbf{I}$, which preserves orthogonality throughout the entire solution path.

The inversion of $(\mathbf{I} + \frac{\tau}{2}\mathbf{S})$ is computationally expensive due to its $n \times n$ dimension. To address this, we apply the Sherman–Morrison–Woodbury (SMW) formula, given as $(\mathbf{B} + \alpha\mathbf{U}\mathbf{V}^T)^{-1} = \mathbf{B}^{-1} - \alpha\mathbf{B}^{-1}\mathbf{U}(\mathbf{I} + \alpha\mathbf{V}^T\mathbf{B}^{-1}\mathbf{U})^{-1}\mathbf{V}^T\mathbf{B}^{-1}$, to reduce this inversion process down to a $2k \times 2k$ matrix by rewriting $\mathbf{S}$ as a product of two low-rank matrices. Let $\mathbf{U} = [\nabla\mathbf{F}|\mathbf{F}]$ and $\mathbf{V} = [\mathbf{F}| - \nabla\mathbf{F}]$ (where $[\mathbf{A}|\mathbf{B}]$ is a block matrix), then we can rewrite $\mathbf{S}$ as $\mathbf{S} = \mathbf{U}\mathbf{V}^T$. Substituting $\mathbf{B}$ with $\mathbf{I}$, α with $\frac{\tau}{2}$, and $\mathbf{S}$ with $\mathbf{U}\mathbf{V}^T$ yields $(\mathbf{I} + \frac{\tau}{2}\mathbf{S})^{-1} = \mathbf{I} - \frac{\tau}{2}\mathbf{U}(\mathbf{I} + \frac{\tau}{2}\mathbf{V}^T\mathbf{U})^{-1}\mathbf{V}^T$. Along with $(\mathbf{I} - \frac{\tau}{2}\mathbf{S}) = (\mathbf{I} - \frac{\tau}{2}\mathbf{U}\mathbf{V}^T)$, the final update rule for $\mathbf{F}$ is

$$\begin{aligned} \mathbf{F}_{n+1} &= (\mathbf{I} + \frac{\tau}{2}\mathbf{S})^{-1}(\mathbf{I} - \frac{\tau}{2}\mathbf{S})\mathbf{F}_n \\ &= (\mathbf{I} - \frac{\tau}{2}\mathbf{U}(\mathbf{I} + \frac{\tau}{2}\mathbf{V}^T\mathbf{U})^{-1}\mathbf{V}^T)(\mathbf{I} - \frac{\tau}{2}\mathbf{U}\mathbf{V}^T)\mathbf{F}_n \\ &= \mathbf{F}_n - \tau\mathbf{U}(\mathbf{I} + \frac{\tau}{2}\mathbf{V}^T\mathbf{U})^{-1}\mathbf{V}^T\mathbf{F}_n. \end{aligned} \tag{3}$$

The enforcement of $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ throughout provides a direct computational benefit in updating $\mathbf{G}$. We use the idea of the hierarchical alternating least squares (HALS) updating scheme (Cichocki, Zdunek, and Amari 2007; Kimura, Tanaka, and Kudo 2015) to update $\mathbf{G}$, since they show that updating each column sequentially is more efficient than a matrix-wise update. By fixing $\mathbf{F}$, the objective function given in Cichocki, Zdunek, and Amari (2007) is: $\text{argmin}_{\mathbf{g}_j} ||\mathbf{X}^{(j)} - \mathbf{f}_j\mathbf{g}_j^T||_F^2$, where $\mathbf{X}^{(j)} = \mathbf{X} - \sum_{k\neq j}\mathbf{f}_k\mathbf{g}_k^T = \mathbf{X} - \mathbf{F}\mathbf{G}^T + \mathbf{f}_j\mathbf{g}_j^T$ is the residual matrix without the jth component. The column-wise update for $\mathbf{G}$ is $\mathbf{g}_j \leftarrow$

$\{(\mathbf{X}^T\mathbf{F})_j - [\mathbf{G}(\mathbf{F}^T\mathbf{F})]_j + \mathbf{g}_j\mathbf{f}_j^T\mathbf{f}_j\}_+$. Since $\mathbf{F}$ is constrained to be strictly orthogonal in our formulation, we have $\mathbf{G}(\mathbf{F}^T\mathbf{F})_j = \mathbf{g}_j\mathbf{f}_j^T\mathbf{f}_j$. Hence, the update rule is simply $\mathbf{g}_j \leftarrow [(\mathbf{X}^T\mathbf{F})_j]_+$, which is essentially a simplified matrix-wise ALS update scheme,

$$\mathbf{G} = [\mathbf{X}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}]_+ = [\mathbf{X}^T\mathbf{F}]_+. \tag{4}$$

The proposed updating method for $\mathbf{G}$ is thus extremely efficient, and it is noteworthy to acknowledge that the matrix-wise and column-wise updating schemes are equivalent under our formulation.

Details of the mathematical derivations of the updates are given in Section A.1 of the Appendix.

4.2. The SONMF for Binary Matrix

In this subsection, we illustrate our proposed method for factorizing a binary matrix, as it requires a different strategy (Schein, Saul, and Ungar 2003; Zhang et al. 2007; Schachtner, Poppel, and Lang 2010; Slawski, Hein, and Lutsik 2013; Tomé et al. 2015). The NMF methods for binary data structures can be formulated using either a logistic regression approach (Schachtner, Poppel, and Lang 2010; Tomé et al. 2015) or a box constrained approach (Slawski, Hein, and Lutsik 2013; Zhang et al. 2014). For our method, we consider the logistic formulation due to its flexibility and ease of optimization. This allows us to have a similar interpretation of the topic vectors as in the continuous case, where each document is represented as a linear combination of orthogonal basis vectors. Furthermore, the objective of enforcing strict orthogonality is much simpler under this formulation since we can follow a similar optimization strategy as in the continuous case.

Analogous to the logistic regression, we utilize the Bernoulli likelihood to capture the underlying probabilistic structure of the binary matrix. In this formulation, we assume that each $\mathbf{X}_{ij}$ follows an independent Bernoulli distribution with parameter p_{ij}, where each $p_{ij} = \sigma([\mathbf{F}\mathbf{G}^T]_{ij}) = \frac{e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}{1+e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}$. The likelihood function is then

$$P(\mathbf{X}_{ij}|\mathbf{F},\mathbf{G}) = \sigma([\mathbf{F}\mathbf{G}^T]_{ij})^{\mathbf{X}_{ij}}(1-\sigma([\mathbf{F}\mathbf{G}^T]_{ij}))^{1-\mathbf{X}_{ij}}. \tag{5}$$

The objective is to find $\mathbf{F}$ and $\mathbf{G}$ such that they maximize the log-likelihood function in Equation (5), or equivalently, minimize the negative log-likelihood,

$$\begin{aligned}\operatorname{argmin}_{\mathbf{F},\mathbf{G}} C(\mathbf{F},\mathbf{G}) &= \operatorname{argmin}_{\mathbf{F},\mathbf{G}} -\mathcal{L}(\mathbf{X}|\mathbf{F},\mathbf{G}) \\ &= -\sum_{i,j}\log\left\{\left(\frac{e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}{1+e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}\right)^{\mathbf{X}_{ij}}\left(\frac{1}{1+e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}\right)^{1-\mathbf{X}_{ij}}\right\} \\ &= \sum_{i,j}\left\{\mathbf{X}_{ij}[\mathbf{F}\mathbf{G}^T]_{ij} - \log(1+e^{[\mathbf{F}\mathbf{G}^T]_{ij}})\right\}. \end{aligned} \tag{6}$$

We update $\mathbf{F}$ in a similar fashion as in the continuous case, but consider a coordinate-wise Newton's method for $\mathbf{G}$. We do not implement the full Newton's method here as the Hessian matrix for $\mathbf{G}$ has a dimension of $nk \times nk$ and is inefficient to compute. Note that the second derivative of the cost function is well-defined, and the first and second derivatives of the cost function with respect to $\mathbf{G}$ are given as

$$\begin{aligned}\frac{\partial C(\mathbf{F},\mathbf{G})}{\partial \mathbf{G}_{jk}} &= \sum_i \frac{e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}{1+e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}\mathbf{F}_{ik} - \mathbf{X}_{ij}\mathbf{F}_{ik} \\ &= \sum_i\left(\frac{1}{1+e^{-[\mathbf{F}\mathbf{G}^T]_{ij}}} - \mathbf{X}_{ij}\right)\mathbf{F}_{ik}\end{aligned}$$

and

$$\frac{\partial^2 C(\mathbf{F},\mathbf{G})}{\partial \mathbf{G}_{jk}^2} = \sum_i\left(\frac{e^{[\mathbf{F}\mathbf{G}^T]_{ij}}}{(1+e^{[\mathbf{F}\mathbf{G}^T]_{ij}})^2}\right)\mathbf{F}_{ik}^2.$$

Following Newton's method, the updating rule for $\mathbf{G}$ in matrix notation is given by

$$\mathbf{G} \leftarrow \mathbf{G} - \eta\frac{\left(\frac{\mathbf{1}}{\mathbf{1}+e^{-(\mathbf{F}\mathbf{G}^T)}} - \mathbf{X}\right)^T\mathbf{F}}{\left(\frac{e^{(\mathbf{F}\mathbf{G}^T)}}{(\mathbf{1}+e^{(\mathbf{F}\mathbf{G}^T)})^2}\right)^T\mathbf{F}^2},$$

where η is a step size and $\mathbf{1}$ is the matrix of all 1's. The quotient and exponential function here are element-wise operations for matrices. In the updating step of $\mathbf{F}$, the only difference from the continuous case is the gradient, whereas the orthogonal-preserving scheme remains the same. Following a similar derivation for $\mathbf{G}$, the gradient of $\mathbf{F}$ is

$$\nabla\mathbf{F} = \frac{\partial C(\mathbf{F},\mathbf{G})}{\partial \mathbf{F}_{ik}} = \left(\frac{\mathbf{1}}{\mathbf{1}+e^{-(\mathbf{F}\mathbf{G}^T)}} - \mathbf{X}\right)\mathbf{G}.$$

However, an over-fitting problem may arise since the algorithm seeks to maximize the probability that $\mathbf{X}_{ij}$ is either 0 or 1 by approximating the corresponding entries of the probability matrix close to 0 or 1. Since $\mathbf{F}$ is constrained to be orthonormal, the approximation scale is solely dependent on $\mathbf{G}$. Thus, larger values in $\mathbf{G}$ increase the risk of over-fitting. To avoid this issue, the step size for updating $\mathbf{G}$ needs to be relatively small. We recommend either 0.05 or 0.01 as a good step size for our algorithm.

4.3. Implementation

In this section, we discuss the implementation of the proposed model, including the initialization, convergence criterion, and algorithms.

The initialization of NMF methods is crucial and has been extensively studied for better numerical stability and convergence (Xue et al. 2008; Langville et al. 2014). In particular, the SVD-based initialization has been studied to work well in practice (Boutsidis and Gallopoulos 2008; Qiao 2015) for the NMF. This is because the truncated SVD provides the best rank-K approximation of any given matrix (Eckart and Young 1936; Wall, Rechtsteiner, and Rocha 2003; Qiao 2015). Our formulation further benefits from the SVD-based initialization because we can simply use the left singular matrix $\mathbf{U}$ directly. In contrast, a nonnegative projection of the SVD solution is required for the existing NMF methods (Langville et al. 2006). We apply the SVD to decompose $\mathbf{X}$ to its best rank-K factorization, that is,

$$\mathbf{X}_k \approx \mathbf{U}_{p\times k}\mathbf{D}_{k\times k}\mathbf{V}_{k\times n}^T,$$

where k is the rank of the target factorization. Our formulation does not require the initialization of $\mathbf{G}$, since the update rule for $\mathbf{G}$ given in (4) is only dependent on $\mathbf{X}$ and $\mathbf{F}$. We apply the same initialization for both the continuous case and the binary case. For more discussion on the initialization, we have conducted a numerical analysis using three different initialization methods, given in Section A.2 of the Appendix.

The convergence criterion is either a predefined number of iterations that is reached, or the difference of the objective function values between two iterations is less than a certain threshold.

$$f(\mathbf{F}^{((i-1)}, \mathbf{G}^{(i-1)}) - f(\mathbf{F}^{((i)}, \mathbf{G}^{(i)}) \leq \epsilon,$$

where any sufficiently small value ϵ could be a feasible choice, such as 10^{-4}.

In the following, we provide the proposed algorithm for continuous and binary design matrices.

Algorithm 1 The semiorthogonal NMF for continuous X

Input: Arbitrary matrix $\mathbf{X}$, number of basis vectors K
Output: Mixed-sign matrix $\mathbf{F}$ and nonnegative matrix $\mathbf{G}$ such that $\mathbf{X} \approx \mathbf{F}\mathbf{G}^T$ and $\mathbf{F}^T\mathbf{F} = \mathbf{I}$.
Initialization: Initialize $\mathbf{F}$ with orthonormal columns and $\tau = 0.5$.

repeat
  $\mathbf{G} = [\mathbf{X}^T\mathbf{F}]_+$
  $\mathbf{R} = 2\mathbf{F}\mathbf{G}^T\mathbf{G} - 2\mathbf{X}\mathbf{G}$
  $\mathbf{U} = [\mathbf{R}, \mathbf{F}]$
  $\mathbf{V} = [\mathbf{F}, -\mathbf{R}]$
  repeat
    $\mathbf{Y}(\tau) \leftarrow \mathbf{F} - \tau\mathbf{U}(\mathbf{I} + \frac{\tau}{2}\mathbf{V}^T\mathbf{U})^{-1}\mathbf{V}^T\mathbf{F}$
    if $E > 0$ **then**
      $\tau = \tau \times 2$
      $\mathbf{F} = \mathbf{Y}(\tau)$
    else if $E \leq 0$ **then**
      $\tau = \frac{\tau}{2}$
    end if
  until $E > 0$
until Convergence criterion is satisfied.

Algorithm 2 The semiorthogonal NMF for binary X

Input: Arbitrary matrix $\mathbf{X}$ with binary elements, number of basis vectors K
Output: Mixed sign matrix $\mathbf{F}$ and nonnegative matrix $\mathbf{G}$ such that $\mathbf{X}_{ij} \sim \text{Bern}\left(\frac{e^{(\mathbf{F}\mathbf{G}^T)_{ij}}}{1+e^{(\mathbf{F}\mathbf{G}^T)_{ij}}}\right)$
Initialization: Initialize $\mathbf{F}$ with orthonormal columns, $\mathbf{G}$ arbitrary, $\eta = 0.05$, and $\tau = 2$.

repeat
  $\mathbf{D}_1 = (\frac{1}{1+e^{-\mathbf{F}\mathbf{G}^T}} - \mathbf{X})^T\mathbf{F}$
  $\mathbf{D}_2 = (\frac{e^{\mathbf{F}\mathbf{G}^T}}{(1+e^{\mathbf{F}\mathbf{G}^T})^2})^T\mathbf{F}^2$
  $\mathbf{G} \leftarrow [\mathbf{G} - \eta\frac{\mathbf{D}_1}{\mathbf{D}_2}]_+$
  $\mathbf{R} = (\frac{1}{1+e^{-\mathbf{F}\mathbf{G}^T}} - \mathbf{X})\mathbf{G}$
  $\mathbf{U} = [\mathbf{R}, \mathbf{F}]$
  $\mathbf{V} = [\mathbf{F}, -\mathbf{R}]$
  repeat
    $\mathbf{Y}(\tau) \leftarrow \mathbf{F} - \tau\mathbf{U}(\mathbf{I} + \frac{\tau}{2}\mathbf{V}^T\mathbf{U})^{-1}\mathbf{V}^T\mathbf{F}$
    if $E > 0$ **then**
      $\tau = \tau \times 2$
      $\mathbf{F} = \mathbf{Y}(\tau)$
    else if $E \leq 0$ **then**
      $\tau = \frac{\tau}{2}$
    end if
  until $E > 0$
until Convergence criterion is satisfied.

The R package "MatrixFact" (Li, Zhu, and Qu 2019) is available at "*cralo31/MatrixFact*" on Github which implements the proposed method for both continuous and binary cases, along with the original NMF (Lee and Seung 2001), ONMF (Kimura, Tanaka, and Kudo 2015), semi-NMF (Ding, Li, and Jordan 2010), and logNMF (Tomé et al. 2015). The existing R packages only include various algorithms for the NMF (Lee and Seung 2001), but lack access to other methods, while our package bridges this gap.

5. Simulated Data Experiments

In this section, we evaluate the performance of our model through various simulated data experiments. We first compare the performance of our model with several well-established variants of the NMF for the continuous case under different simulation settings. For the binary version, we show that both the cost function and difference between the true and estimated probability matrices monotonically converge under the algorithm, along with a comparison with another state of the art model.

5.1. Simulation for the Continuous Case

For the continuous case, we evaluate the average residual and orthogonal residual, where

$$\text{Average residual} = \frac{||\mathbf{X} - \mathbf{F}\mathbf{G}^T||_F^2}{n \times p} = \frac{1}{n \times p}\sum_{i,j}(\mathbf{X} - [\mathbf{F}\mathbf{G}^T])_{ij}^2, \tag{7}$$

and

$$\text{Orthogonal residual} = ||\mathbf{F}^T\mathbf{F} - \mathbf{I}||_F^2. \tag{8}$$

We simulate the true $\mathbf{F}$ and $\mathbf{G}$ matrices and evaluate how the algorithms perform on recovering them. Thus, we also calculate the difference between the column space of $\mathbf{F}$, $\mathbf{G}$ and $\hat{\mathbf{F}}$, $\hat{\mathbf{G}}$ in which $\mathbf{F}$ and $\mathbf{G}$ are the true underlying matrices, and $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$ are the approximated matrices from the factorization. That is,

$$\epsilon_{\mathbf{F}} = ||\mathbf{H}_{\mathbf{F}} - \mathbf{H}_{\hat{\mathbf{F}}}||_F^2 \quad \text{and} \quad \epsilon_{\mathbf{G}} = ||\mathbf{H}_{\mathbf{G}} - \mathbf{H}_{\hat{\mathbf{G}}}||_F^2, \tag{9}$$

where $\mathbf{H}_{\mathbf{F}}$, $\mathbf{H}_{\mathbf{G}}$, $\mathbf{H}_{\hat{\mathbf{F}}}$, and $\mathbf{H}_{\hat{\mathbf{G}}}$ are the projection matrices of their respective counterparts, that is, $\mathbf{H}_{\mathbf{F}} = \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T$. In addition, we evaluate the sparsity of the solutions by measuring the proportion of 0's within $\mathbf{F}$ and $\mathbf{G}$ after shrinking elements to 0 with 10^{-10} as a cutoff.

We compare our method with three other popular NMF methods, that is, NMF with multiplicative updates (Lee and Seung 2001), semi-NMF (Ding, Li, and Jordan 2010), and ONMF (Kimura, Tanaka, and Kudo 2015). The simulations are conducted under an i7-7700HQ with eight cores at 3.8 GHz. Three different scenarios are considered:

1. $\mathbf{F}_{p\times k}$ where $\mathbf{F}_{ik} \sim \text{Unif}(0, 1)$ and $\mathbf{G}_{n\times k}$ where $\mathbf{G}_{jk} \sim \text{Unif}(0, 2)$.
2. Nonnegative and orthogonal $\mathbf{F}_{p\times k}$ and $\mathbf{G}_{n\times k}$ where $\mathbf{G}_{jk} \sim \text{Unif}(0, 2)$.
3. Orthonormal $\mathbf{F}_{p\times k}$ and $\mathbf{G}_{n\times k}$ where $\mathbf{G}_{jk} \sim \text{Unif}(0, 2)$.

Based on the generated true $\mathbf{F}$ and $\mathbf{G}$ matrices, we construct the observed $\mathbf{X} = \mathbf{F}\mathbf{G}^T + \mathbf{E}$, where $\mathbf{E}$ is a matrix of random error such that $\mathbf{E}_{ij} \sim N(0, 0.3)$. We then further truncate all negative values in the observed $\mathbf{X}$ to 0 to enforce nonnegativity to implement NMF and ONMF. In this simulation experiment, we consider $n = p = 500$ and $k = 10, 30, 50$.

We implement the K-means initialization for the semi-NMF (Ding, Li, and Jordan 2010). Lee and Seung (2001) and Kimura, Tanaka, and Kudo (2015) proposed using random initialization for the NMF and ONMF, respectively. For a fair comparison, we initialize $\mathbf{F}$ and $\mathbf{G}$ using a slightly modified SVD approach, where we truncate all the negative values in $\mathbf{U}$ to a small positive constant $\delta = 10^{-10}$ to enforce nonnegativity and avoid the zero-locking problem for the NMF. We then apply our update rule for $\mathbf{G}$ as the initialization for $\mathbf{G}$, that is,

$$\mathbf{F}_0 = [\mathbf{U}]_\delta, \quad \mathbf{G}_0 = [\mathbf{X}^T\mathbf{F}_0]_\delta,$$

where $[x]_\delta = \max(x, \delta)$. The average values of the above four criterion over 100 simulation trials with different underlying true $\mathbf{F}$ and $\mathbf{G}$ are reported under three scenarios in Tables 1–3, respectively, where each trial is set to run 500 iterations. We display the convergence plot of the objective function in Figure 2–4 for all four methods, where the convergence criterion under consideration is

$$0 \le f(\mathbf{F}^{((i-1)}, \mathbf{G}^{(i-1)}) - f(\mathbf{F}^{((i)}, \mathbf{G}^{(i)}) \le 0.0001.$$

Table 1. Comparisons of the proposed method with the other NMF methods on factorization accuracy, sparsity of the solutions, computation time, and convergence speed under simulation scenario (1).

K	Average residual	Orthogonal residual	ϵ_F	ϵ_G	F sparsity	G sparsity	Time (sec)	Iterations until threshold
				SONMF				
10	0.0874	3.47×10^{-29}	0.345	0.409	0	1.64	0.59	12.6
30	0.0809	1.07×10^{-28}	0.634	0.699	0	1.02	0.98	15.5
50	0.0749	2.19×10^{-28}	0.839	0.913	0	0.91	1.21	11.3
				ONMF				
10	0.2912	0.475	3.808	1.040	83.13	0.01	0.52	14.4
30	0.7620	3.367	6.988	3.181	85.90	0.14	1.00	31.4
50	1.1925	9.625	8.972	4.168	83.83	0.35	2.07	50.5
				NMF				
10	0.1742	N/A	1.894	1.893	26.87	27.35	1.20	121.7
30	0.3498	N/A	3.326	3.330	28.37	28.76	7.03	364.3
50	0.5291	N/A	4.337	4.335	29.63	30.15	16.21+	500+
				Semi-NMF				
10	0.1217	N/A	0.880	1.756	0	0	2.89	413.0
30	0.1862	N/A	1.858	3.479	0	0	10.34+	500+
50	0.2751	N/A	2.611	5.049	0	0	17.26+	500+

NOTE: The sparsity measures are given in percentages. The plus sign in columns 8 and 9 indicates that the convergence threshold has not been satisfied after 500 iterations.

Table 2. Comparisons of the proposed method with the other NMF methods on factorization accuracy, sparsity of the solutions, computation time, and convergence speed under simulation scenario (2).

K	Average residual	Orthogonal residual	ϵ_F	ϵ_G	F sparsity	G sparsity	Time (sec)	Iterations until threshold
				SONMF				
10	0.0774	1.95×10^{-24}	0.274	0.486	0	4.80	0.55	11.0
30	0.0728	3.84×10^{-24}	0.840	1.169	0	4.96	0.83	10.8
50	0.0664	2.57×10^{-24}	1.400	1.765	0	4.87	1.16	11.1
				ONMF				
10	0.0827	0.447	0.420	0.535	62.5	1.35	0.59	26.0
30	0.0711	0.669	0.536	1.007	72.8	2.28	1.26	42.6
50	0.0671	1.775	0.806	1.660	75.6	4.71	2.32	58.0
				NMF				
10	0.1026	N/A	0.207	1.162	30.1	21.4	2.44	307.2
30	0.1233	N/A	0.669	2.886	40.3	28.3	7.55	379.4
50	0.1374	N/A	1.412	4.312	44.8	32.2	13.03	399.8
				Semi-NMF				
10	0.0746	N/A	0.274	0.373	0	0	0.27	30.0
30	0.0673	N/A	0.849	0.941	0	0	0.88	40.1
50	0.0628	N/A	1.509	1.747	0	0	1.73	49.7

Table 3. Comparisons of the proposed method with the other NMF methods on factorization accuracy, sparsity of the solutions, computation time, and convergence speed under simulation scenario (3).

K	Average residual	Orthogonal residual	ϵ_F	ϵ_G	F sparsity	G sparsity	Time (sec)	Iterations until threshold
				SONMF				
10	0.0859	3.1×10^{-27}	3.717	3.725	0	19.55	0.50	6.6
30	0.0771	2.4×10^{-25}	6.348	6.376	0	17.26	0.69	8.0
50	0.0693	1.7×10^{-24}	8.039	8.070	0	16.13	0.96	8.7
				Semi-NMF				
10	0.0854	N/A	3.741	3.732	0	0	0.16	8.2
30	0.0759	N/A	6.342	6.340	0	0	0.40	14.6
50	0.0671	N/A	8.007	8.008	0	0	0.76	18.6

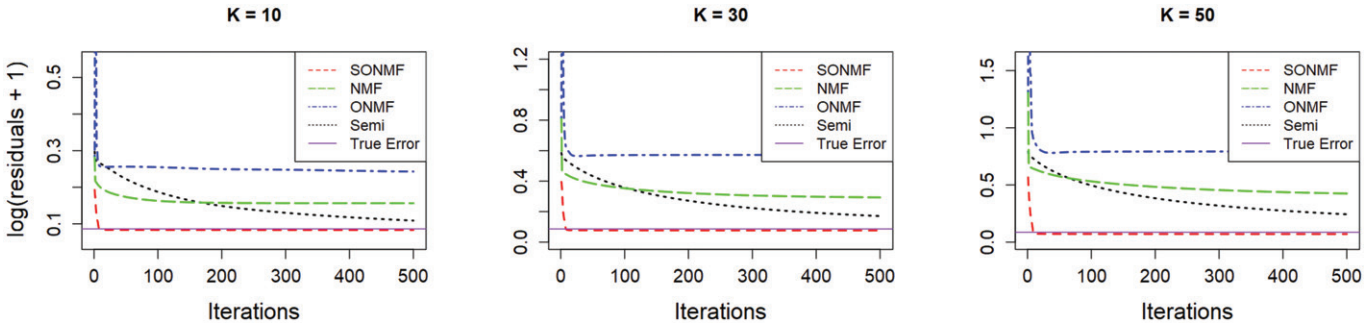

Figure 2. Convergence plots for the average residual in (7) under scenario (1) for 4 the NMF variants. The SONMF is the only method to have converged to the true error.

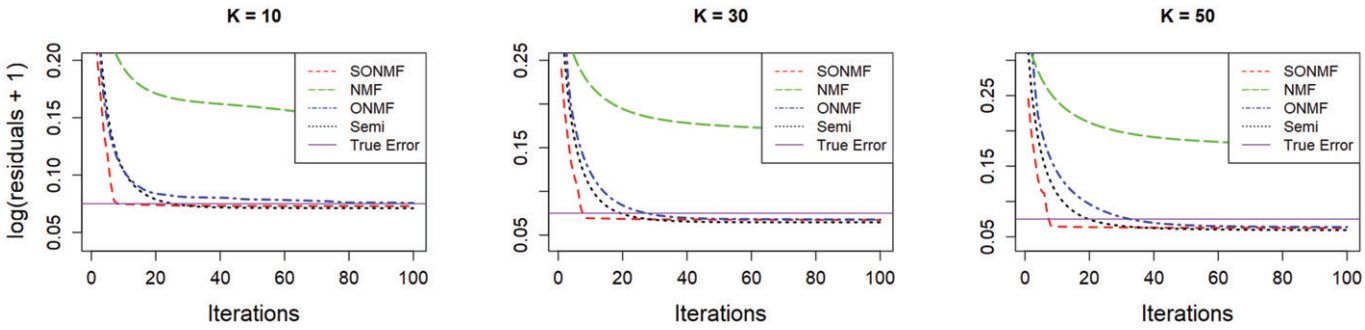

Figure 3. Convergence plots for the average residual in (7) under scenario (2) for all 4 of the NMF variants. Only the first 100 iterations are shown as all methods apart from the NMF have converged to the true error.

For better visibility between the convergence trends, we plot $log(\text{residuals} + 1)$ instead of the original scale. The last two columns of Tables 1–3 indicate the time and the number of iterations for each algorithm to reach this criterion.

5.1.1. Simulation Results for the Continuous Case

Tables 1–3 show that the SONMF has several advantages over the other NMF methods. First, our model converges quickly and consistently regardless of the structure of the true matrices, reaching the convergence criterion and true error in only 10 iterations, greatly surpassing the rate of convergence of the other models (Figures 2–4).

For scenario (1), our model is significantly better in terms of the factorization accuracy and recreating the true matrices, as shown by the smallest average residual, $\epsilon_{\mathbf{F}}$ and $\epsilon_{\mathbf{G}}$ values, especially for larger K's. For the semi-NMF and NMF, the mean value over 100 trials fails to converge to the true error. We believe this is due to the large number of saddle points that the true $\mathbf{F}$ possesses, as shown by the large $\epsilon_{\mathbf{F}}$ values. The semi-NMF has the least constraints among these four models, and converges to the true error eventually, but has a much slower rate of convergence.

When the underlying structure of $\mathbf{F}$ is more well-defined as in scenario (2), all four models converge to the true error. For factorization accuracy, our model outperforms the NMF, but performs slightly worse than the ONMF and semi-NMF. This is expected as scenario (2) is tailored toward the ONMF's formulation, evident in the low $\epsilon_{\mathbf{F}}$. The semi-NMF has the lowest error because it has the least amount of constraints. For scenario (3), our model and the semi-NMF have similar performances, but our model has a faster convergence rate. Note that the estimated error is lower than the true error for some models in the above scenarios. This is due to over-fitting in factorizing $\tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{E}$, in the sense that we also factorize the error matrix $\mathbf{E}$ along with the true $\mathbf{X}$. This issue becomes even more evident

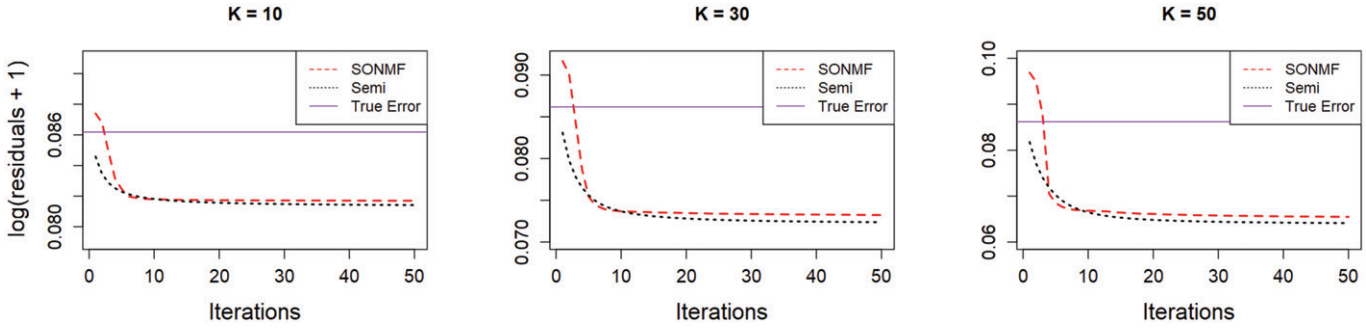

Figure 4. Convergence plots for the average residual in (7) under scenario (3) for all 4 of the NMF variants. Only the first 50 iterations are shown, as both models have already converged to the true error and reached the convergence criteria.

when the true rank of $\mathbf{F}$ and $\mathbf{G}$ is less than the specified target rank, especially for the nonconstrained semi-NMF. We refer to Section A.3 of the Appendix for additional discussion.

Our algorithm successfully preserves strict orthogonality throughout. This is contrasted with the increasing orthogonality residual that the ONMF yields as K increases. The strict nonnegative versions yield sparser solutions compared to the mixed variants, which aligns with previous studies (Lee and Seung 2001; Guan and Dy 2009; Ding, Li, and Jordan 2010). The semi-NMF returns little to no sparse elements at all in its solution for all scenarios, which is consistent with Ding, Li, and Jordan's (2010) finding. Our model has a slight advantage in this criteria over the semi-NMF, with a moderate degree of sparsity in the third scenario. However, our analyses of the triage datasets provided in Section 6.1 show that our algorithm yields a moderate degree of sparsity in $\mathbf{G}$ (document-topic matrix), which is beneficial for interpretation.

5.2. Simulation for the Binary Case

For the binary response, we use the mean value of the cost function $C(\mathbf{F}, \mathbf{G})$ in Equation (6) as our evaluation criterion instead of the normalized residual. That is,

$$C(\mathbf{F}, \mathbf{G}) = \frac{1}{N} \sum_{i,j} \mathbf{X}_{ij}(\mathbf{F}\mathbf{G}^T)_{ij} - \log(1 + e^{[\mathbf{F}\mathbf{G}^T]_{ij}}), \qquad (10)$$

where N is the total number of elements in $\mathbf{X}$. We also consider the orthogonal residual, $\epsilon_{\mathbf{F}}$ and $\epsilon_{\mathbf{G}}$ given in Equations (8) and (9), respectively. Additionally, we evaluate the difference between the true and estimated probability matrices, $\epsilon_{\mathbf{P}} = ||\mathbf{P} - \hat{\mathbf{P}}||_F^2$.

For the binary simulation setting, we generate mixed-sign $\mathbf{F}$ and nonnegative $\mathbf{G}$ such that $\mathbf{F}_{ij} \sim N(0, 1)$ and $\mathbf{G}_{ij} \sim \text{Unif}(0, 1)$. We then construct the true probability matrix $\mathbf{P}$ using the logistic sigmoid function, $\mathbf{P} = \frac{e^{[\mathbf{F}\mathbf{G}^T]}}{\mathbf{1}+e^{[\mathbf{F}\mathbf{G}^T]}}$. We then add a matrix of random error $\mathbf{E}$ to $\mathbf{P}$ where $\mathbf{E}_{ij} \sim N(0, 0.1)$. Finally, we generate the true $\mathbf{X}$ where each $\mathbf{X}_{ij} \sim \text{Bernoulli}([\mathbf{P} + \mathbf{E}]_{ij})$ and has dimension 500-by-500.

Similar to the continuous case, we consider $K = 10, 30, 50$.The average values of the above five criterion over 100 simulation trials are reported. For our method, we use a step size of 0.01 for Newton's update of $\mathbf{G}$. We compare the performance of our method with logNMF (Tomé et al. 2015), where they set their step size for the gradient ascent to be 0.001. For our model, the initialization for $\mathbf{F}$ and $\mathbf{G}$ are the same as in the continuous case. However, this initialization resulted in severe numerical issues for logNMF in our experiments. Therefore, we initialize $\mathbf{F}$ and $\mathbf{G}$ with each $\mathbf{F}_{ij} \sim \text{Unif}(0, 1)$ and $\mathbf{G}_{ij} \sim N(0, 1)$. We compare the results of both models after running 500 iterations.

5.2.1. Simulation Results for the Binary Case

The result above shows that our model has a faster convergence rate toward the true cost and, ultimately, a lower mean cost than the logNMF (Figure 5). Additionally, our model has a lower error for $\epsilon_{\mathbf{P}}$, $\epsilon_{\mathbf{F}}$, and $\epsilon_{\mathbf{G}}$, respectively, when both algorithms reach the convergence criteria (Table 4). Due to the implementation of a line search and Newton's method in our update scheme, the computation cost is higher, as reflected by the time required to run 500 iterations. However, our model reaches the true error in about half the iterations compared to the logNMF, which compensates for the longer computation time. Furthermore, our model yields a sparser solution, which is beneficial for interpretation.

Unlike the continuous case, the SVD-based initialization does not provide a rapid convergence to the true error for our model. The reason here is because the SVD is applied on $\mathbf{X}$, but $\mathbf{F}$ and $\mathbf{G}$ are estimating the underlying probability matrix of $\mathbf{X}$ and not $\mathbf{X}$ itself. For $\epsilon_{\mathbf{P}}$, the difference between $\mathbf{P}$ and $\hat{\mathbf{P}}$ converges once the average cost for the factorization reaches the true cost. An important caveat to note here is that the rate of convergence of our model is very sensitive to the step size of $\mathbf{G}$. In our numerical experiment, we discovered that the degree of over-fitting increases as the number of basis vectors increases, and thus the step size should be adjusted accordingly. For larger K's, it is recommended to use a smaller step size. We refer the readers to Section A.4 in the Appendix for additional discussion on the step size of $\mathbf{G}$.

6. Triage Notes Case Study

In this section, we focus on the analyses of the triage notes dataset. The vocabulary used in the notes is relatively different across different medical complaints. Thus, it is necessary to consider each complaint separately. For this study, we consider the analyses of the triage notes under seven different medical complaints. Our main objective is a binary classification problem where we aim to classify whether a patient is either (a)

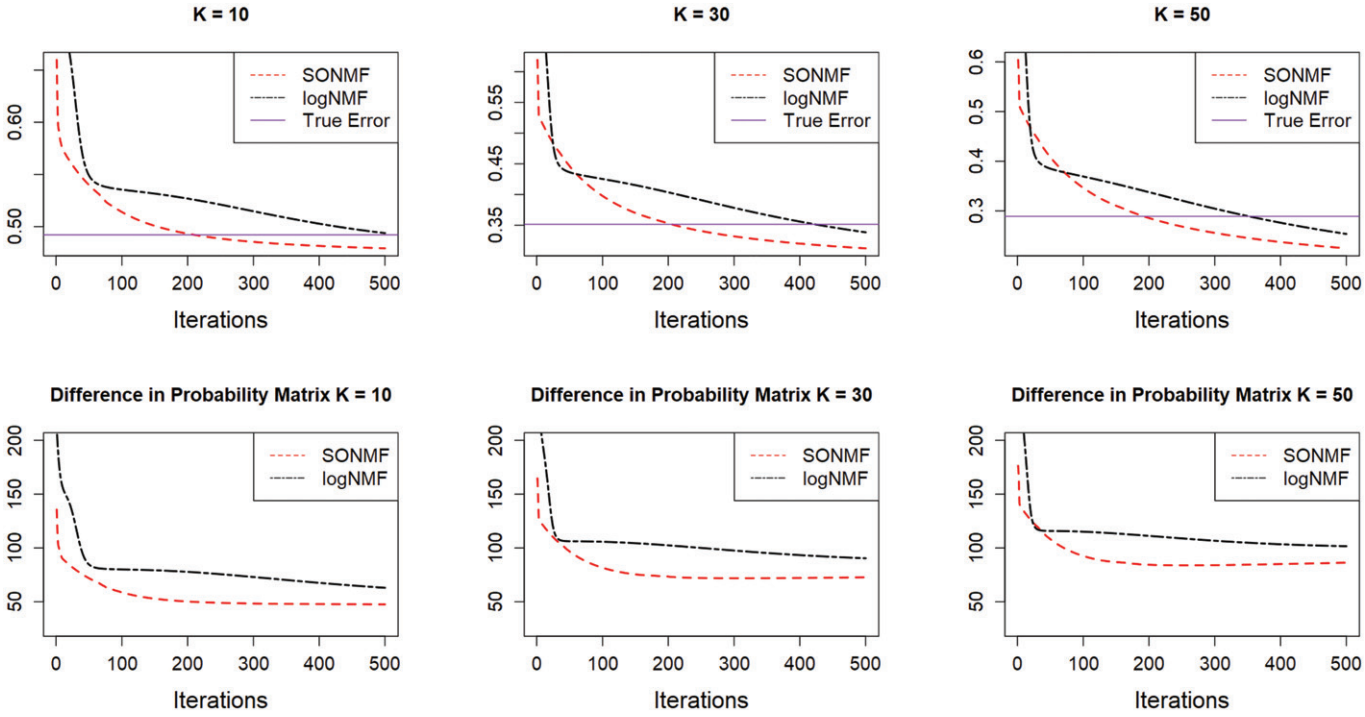

Figure 5. Comparison of the mean cost (top) and recovery of the probability matrix (bot) between our method and the logNMF.

Table 4. Comparisons between the proposed method and the logNMF on factorization accuracy, sparsity of solutions, and computation time.

K	Average cost	Orthogonal residual	ϵ_P	ϵ_F	ϵ_G	F sparsity	G sparsity	Time (sec)
			SONMF (binary)					
10	0.4792	2.177×10^{-25}	47.61	1.876	2.030	0	18.06	67.11
30	0.3124	1.577×10^{-23}	72.74	4.394	4.249	0	21.22	81.06
50	0.2246	5.284×10^{-23}	86.32	6.462	6.110	0	23.19	99.15
			logNMF					
10	0.4938	N/A	63.03	2.789	2.938	8.14	0	39.75
30	0.3385	N/A	90.41	5.649	5.639	7.56	0	46.21
50	0.2536	N/A	101.65	7.606	7.499	8.67	0	53.38

admitted or (b) discharged. To do so, we first fit a supervised learning model on the bag-of-words after preprocessing the text information. Then, we show that the classification error can be improved by performing a linear transformation of basis with our model and using the topics as features instead. In addition, we also present the interpretation for both the word-topic matrix $\mathbf{F}$ and document-topic matrix $\mathbf{G}$ identified by our model for selected datasets.

We perform standard text preprocessing methods to clean the triage notes. Specifically, we first convert the data into a vector-space model (Salton, Wong, and Yang 1975), after removing numbers, punctuation, stop words, and stemming words to their root form. Given n documents, we construct a word-document matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$, where $\mathbf{X}_{ij}$ corresponds to the occurrence or significance of word w_i in document d_j, depending on the weighting scheme. We then consider the TF-IDF and binary weighing (Gupta and Lehal 2009) for the continuous and binary cases, respectively.

For classification, we denote the training and testing bag-of-words as $\mathbf{X}_{\text{train}}$ and $\mathbf{X}_{\text{test}}$, respectively. Applying the matrix factorization method yields a word-topic matrix $\mathbf{F}_{\text{train}}$ and document-topic matrix $\mathbf{G}_{\text{train}}$, such that the mixed and orthogonal constraint is imposed on the word-topic vectors in $\mathbf{F}_{\text{train}}$. After obtaining the factorized solution, we project both $\mathbf{X}_{\text{train}}$ and $\mathbf{X}_{\text{test}}$ onto the column space of $\mathbf{F}_{\text{train}}$. Let $\mathbf{G}_{\text{proj}} = \mathbf{X}_{\text{train}}^T \mathbf{F}_{\text{train}}$ and $\mathbf{G}_{\text{test}} = \mathbf{X}_{\text{test}}^T \mathbf{F}_{\text{train}}$, then $\mathbf{G}_{\text{proj}}$ and $\mathbf{G}_{\text{test}}$ are reduced dimension representations of $\mathbf{X}_{\text{train}}$ and $\mathbf{X}_{\text{test}}$, respectively, and replace $\mathbf{X}_{\text{train}}$ and $\mathbf{X}_{\text{test}}$ as the new features. Intuitively, we can regard $\mathbf{F}_{\text{train}}$ as a summary device, where each topic consists of a linear combination of different words. After applying the projection, $\mathbf{G}_{\text{proj}}$ can be viewed as a summary of the original bag-of-word matrix, where each document is now a linear combination of the topics from $\mathbf{F}_{\text{train}}$.

We apply a 5-fold cross-validation (Hastie, Tibshirani, and Friedman 2009) for classification, and the results are averaged over 20 different runs, where the observations in each run are randomly assigned to different folds with stratified sampling using the *caret* package (Kuhn 2008). On a side note, we also consider both the 2-fold and 10-fold cross-validation. The former led to worse performance due to the smaller sample size

Table 5. Datasets considered in this study.

Datasets	Dimension (Docs × Words)	Baseline	Linear	KNN	RF	SVM
Altered Level of Consciousness (ALC)	5220 × 5128	51.15	75.01 (75.02)	53.26 (54.79)	75.06 (75.06)	75.32 (75.40)
Cough	13084 × 5876	87.21	87.65 (87.21)	84.55 (85.61)	87.51 (87.57)	87.63 (87.54)
Fever	7302 × 4770	77.20	81.52 (81.69)	77.31 (77.43)	82.07 (82.20)	82.61 (82.79)
General Weakness	7442 × 5455	52.21	69.31 (69.52)	53.97 (62.02)	69.40 (69.48)	69.47 (69.51)
Lower Extremity Injury	12377 × 5180	82.50	88.60 (88.67)	83.11 (84.26)	88.87 (88.75)	88.96 (89.01)
Shortness of Breath	9322 × 4659	55.04	74.29 (74.17)	55.20 (57.22)	74.31 (74.32)	74.40 (74.43)
Symptoms of Stroke	5036 × 3869	54.83	74.21 (74.20)	55.54 (56.58)	74.27 (74.21)	74.33 (74.38)

NOTE: The dimension of the document-word matrix, the proportion of the majority class (baseline for classification accuracy) and classification accuracy on the bag-of-words using the considered models are shown. For each dataset, the classification accuracy for both the continuous and binary case (in parenthesis) are provided.

of the training set in each fold, while the latter yielded similar results as the 5-fold. Thus, we only report the results for the 5-fold cross-validation.

We compare our model with four other NMF methods. For TF-IDF weighting, we apply our continuous model and compare it with the NMF (Lee and Seung 2001), ONMF (Kimura, Tanaka, and Kudo 2015), and semi-NMF (Ding, Li, and Jordan 2010). For binary weighting, we consider the comparison with the logNMF (Tomé et al. 2015). The stopping criteria of the NMF algorithms are set to 100 iterations. During our experiments, we noticed that the default step size $\eta = 0.001$ used in Tomé et al. (2015) was too large for the logNMF on these datasets and caused unstable performances. Thus, we use $\eta = 0.0005$ for the logNMF in the following experiments.

We implement 4 different supervised learning models, the penalized linear regression (Tibshirani 1996), random forest (Breiman 2001), K-nearest neighbor (Dudani 1976), and support vector machine (Suykens and Vandewalle 1999). We consider different tuning parameters for each model (alpha for linear models, nodesize for random forests, number of nearest neighbors for KNN, and kernel/gamma/cost for SVM), and report the best result for each model. These models are implemented using the R packages *glmnet* (Friedman, Hastie, and Tibshirani 2010), *randomforest* (RColorBrewer and Liaw 2018), *class* (Ripley, Venables, and Ripley 2015), and *e1071* (Meyer et al. 2014), respectively. We show that our method of factorization is able to improve the classification performance over the naive bag-of-words, while also outperforming other matrix factorization methods. In addition, we present the average residual of the factorization and the sparsity of the solutions. Note that these two measurements are computed from the full bag-of-words, and not the training sets from cross-validation.

6.1. Results for the Classification of Triage Notes

The seven triage datasets we consider in this study are given in the table below. The dimensions of each dataset, the baseline accuracy of classifying all observations as the majority class, and the classification accuracy using different supervised learning models on the document-word matrix are also presented.

Table 5 shows that the triage notes contain predictive signals toward the disposition of patients to varying degrees. The penalized linear regression, random forest, and SVM share similar performances. The KNN performs considerably worse than the others, likely due to the ineffectiveness of a distance-based classification approach on such a large feature space. On the other hand, there is no clear advantage between either the continuous or the binary formulations. From our observation, the medical complaints that consist of more severe symptoms (i.e., ALC, stroke) have a much more significant improvement from baseline compared to those that are relatively common (cough, fever). This may be because the underlying causes for these common symptoms are much more diverse, and may require further examinations (i.e., x-ray) by the doctors to assess the severity of these patients. Meanwhile, the causes for the severe symptoms may be easier to identify and evaluate (e.g., ALC due to alcohol consumption vs. ALC due to head trauma). Thus, the verbal description from patients themselves may be sufficient for a doctor to make a diagnosis.

Next, we present the classification performance after transforming the basis with the various NMF methods. We present the results for the ALC dataset in Table 6, while the results for the remaining datasets are given in Section A.5 of the Appendix.

We observe that applying factorization and basis projection improves the classification performances by up to 1% compared to the bag-of-words model across different datasets. Our proposed model achieves the best performance under the radial kernel SVM, which outperforms the best result of any other NMF models. The penalized linear regression also yields comparable performances. Additionally, our method has a slight advantage in terms of precision, recall, and F1 compared to other models (Table 7). This shows that a strict orthogonal basis is beneficial for improving prediction by reducing the potential multicollinearity among the features. The classification performance for the continuous case is better than the binary case, with the logNMF having a significantly worse performance than the other methods. Thus, we recommend using the continuous formulation in practice. Our model yields an increasingly sparse solution in the $\mathbf{G}$ matrix as the number of topics increases. This is advantageous for interpretation compared to the dense solutions obtained by the semi-NMF, despite both methods

Table 6. Classification results for the "Altered Level of Consciousness" dataset.

Altered Level of Consciousness						
K	Cost	Sparsity (**F,G**)	Linear	KNN	RF	SVM
SONMF						
10	0.1370	(0, 24.08)	74.03	**73.61**	74.08	74.29
25	0.1315	(0, 34.16)	74.40	72.23	**74.29**	74.47
50	0.1247	(0, 42.06)	74.96	71.40	74.23	75.10
100	0.1145	(0, 47.70)	75.39	67.56	74.22	75.73
150	0.1071	(0, 49.27)	**75.72**	63.71	74.10	***75.89***
NMF						
10	0.1375	(61.89, 42.31)	72.54	72.00	72.91	73.06
25	0.1324	(75.14, 53.99)	73.21	71.29	73.81	73.29
50	0.1258	(81.82, 62.67)	74.12	71.39	74.25	73.96
100	0.1152	(86.65, 72.35)	74.76	70.24	74.24	74.70
150	0.1069	(88.87, 77.66)	75.13	68.01	74.02	74.91
ONMF						
10	0.1374	(66.12, 41.03)	72.92	72.19	73.15	73.23
25	0.1322	(75.95, 57.85)	73.49	71.56	73.83	73.33
50	0.1256	(80.99, 70.31)	74.23	71.65	74.23	74.19
100	0.1153	(83.70, 83.61)	74.89	71.30	74.28	74.93
150	0.1073	(84.64, 89.38)	75.32	70.45	74.22	75.11
Semi						
10	0.1368	(0, 0)	73.68	73.07	74.06	74.35
25	0.1311	(0, 0)	74.37	71.56	74.15	74.59
50	0.1235	(0, 0)	74.69	69.55	73.86	74.93
100	0.1115	(0, 0)	75.08	65.16	73.91	75.47
150	0.1018	(0, 0)	75.30	62.24	73.55	75.62
SONMF (bin)						
10	0.0273	(0, 0)	69.10	68.72	69.78	70.18
25	0.0259	(0, 0)	70.18	67.95	70.39	71.01
50	0.0207	(0, 0.10)	72.19	67.40	71.54	72.70
100	0.0137	(0, 1.40)	73.65	65.56	72.34	74.32
150	0.0110	(0, 2.26)	74.29	63.91	72.56	74.68
logNMF (bin)						
10	0.0179	(0.01, 0)	56.75	54.75	55.74	57.16
25	0.0182	(0.04, 0)	59.41	57.12	58.65	59.92
50	0.0229	(0.97, 0)	60.45	58.01	59.92	61.10
100	0.0347	(8.85, 0)	63.28	60.98	62.26	64.17
150	0.049	(16.70, 0)	65.31	62.37	63.76	66.18

NOTE: The highest classification accuracy for each supervised learning method is marked in bold, and the highest accuracy overall is marked in bold and italics. The first four methods are applied using the continuous bag-of-words, while the last two methods are applied using the binary bag-of-words. The final cost and sparsity of the factorized solutions are also provided.

Table 7. Accuracy, Precision, Recall, and F1 for the best supervised model after clustering using different NMF approaches.

Altered Level of Consciousness				
Method	Accuracy	Precision	Recall	F1
SONMF	**75.89**	**76.53**	**73.10**	**74.75**
NMF	75.13	76.11	71.60	73.76
ONMF	75.32	76.40	71.66	73.93
Semi	75.62	76.49	72.35	74.33
SONMF (bin)	74.68	75.16	71.98	73.51
logNMF (bin)	66.18	66.07	63.37	64.66

allowing mixed values in **F**. On the other hand, neither binary methods yield sparse solutions (Table 6).

Another direct benefit of removing the correlations between features with the SONMF is the topic features in our model is much more efficient at representing the main information of the data. Our model effectively reaches the classification accuracy of the bag-of-words model with only 25–50 topics, whereas other methods require somewhere between 50–100 topics or even more. The classification accuracy also increases with more topics at a diminishing rate, but has a larger increase in other methods, especially under the nonnegative approaches. From our experiment, having more than 150 basis vectors does not provide a noticeable improvement in performance. Aside from over-fitting, the computation cost for factorizing a large dimension bag-of-words matrix increases sharply as K increases, and thus the trade-off is not warranted.

The above results suggest that choosing the appropriate number of topics is a challenging problem. To select the appropriate rank k from a data driven perspective, we have considered the elbow plot (Abdi and Williams 2010) as a metric. However, as shown in Appendix A.5, there were no evident cutoff point, and thus it seems that the elbow plot is ineffective for our application. Regardless, many supervised learning methods in the next stage of the modeling may effectively reduce the impact of over-selecting the rank. Therefore, the selection of the appropriate rank should be considered jointly from both the matrix factorization and supervised learning model perspectives. Nevertheless, a consensus from our numerical studies, including the other 6 datasets, shows that 100 topics are sufficient for classification. For additional discussion on this problem, please refer to Section A.5 of the Appendix.

Although the improvement using topic modeling on top of bag-of-words may not be significant, it is nevertheless numerically consistent and clinically important. Since this study is directly related to people's health, even a small increase in classification accuracy may be clinically significant and critical to some patients. From the medical perspective, improved predictions of disposition at the early phase of triage can lead to better patient streaming strategies in ED. Specifically, more accurate predictions means that more patients are sent to the "right" queue, which can improve operational efficiency and reduce the overall waiting and length of stay (LoS) in ED. In addition, more accurate predictions of admissions can help Intensive Care Units (ICU) coordinate beds in advance to meet the true demands better. This can significantly reduce the boarding time and save crucial ED resources to serve more patients and, consequently, improve clinical outcomes, healthcare efficiency, and revenue. Specifically, less ED crowding has been shown to be associated with lower mortality (Sprivulis et al. 2006; Sun et al. 2013). In addition, each hour of reduction in boarding time would increase daily revenue by $9.693 to $13,298 from patients who left hospitals or are diverted due to overcrowding (Pines et al. 2011).

6.2. Interpretation of the Word-Topic Matrix

In this subsection, we present examples of the word-topic vectors (**F** matrix) illustrated by our method from the "Lower Extremity Injury" and "Symptoms of Stroke" datasets. The uncorrelated word-topic vectors provide us with an immediate interpretation of the main reasons and causes for the hospital visits. The meaning of each topic can be interpreted by

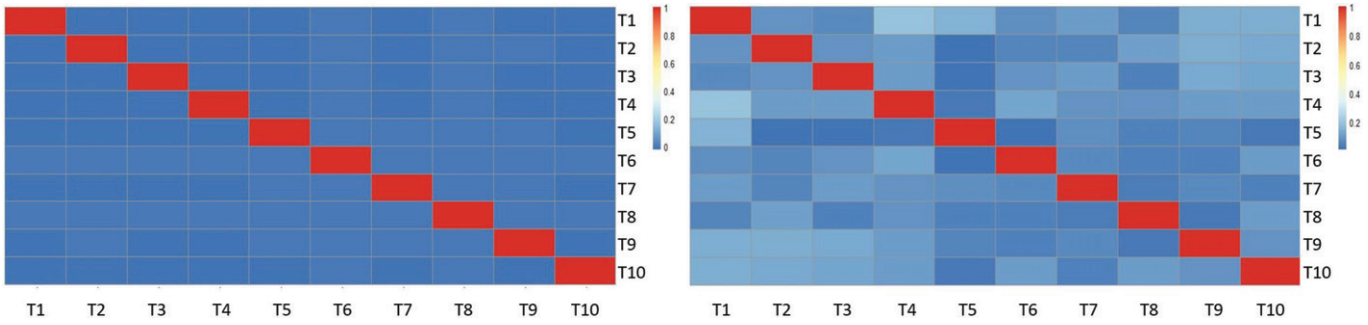

Figure 6. Correlation heatmap of 10 generated word-topic vectors from the SONMF (left) and NMF (right) for Lower Extremity Injury dataset. The "T" stands for "Topic."

Table 8. Words with the largest magnitude under the first five topics generated by the SONMF for the Lower Extremity Injury and Stroke datasets.

(SONMF) Lower Extremity Injury				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Positive				
hip	foot	knee	xray	play
rotate	ankle	fall	fracture	roll
glf	bruise	twist	sent	soccer
break	swell	bend	told	pop
morphin	ago	left	done	heard
Negative				
ankle	toe	play	weight	ago
knee	cap	pop	able	knee
foot	alright	soccer	bear	week
swell	drop	day	note	fall
calf	big	ago	aspect	increase

(SONMF) Symptoms of Stroke				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Positive				
left	team	right	equal	episode
side	stroke	side	grip	min
leg	see	eye	strong	last
arm	aware	facial	steady	resolve
weak	place	face	strength	approx
Negative				
deny	deny	left	gait	day
note	resolve	leg	left	note
right	symptom	weak	unsteady	onset
episode	home	state	arm	side
state	week	confuse	weak	place

Abbreviations: glf (ground-level fall), ubl (Ubiquitin-like protein), gcs (glasgow coma scale).

examining the words with the largest positive weight calculated by the proposed model. On the other hand, the words with the largest negative weight under the same topic indicate that they are negatively correlated with the topic. This implies that words with negative weights tend not to appear together with words with positive weights. Consequently, this provides an insight into identifying and isolating the main causes of admission or discharge for hospital management. In addition, the generated topics also inform us on what symptoms or complaints tend to happen simultaneously, and what complaints tend not to co-exist. To illustrate the above points, we present the heatmap of the correlations among the word-topics in **F**, along with the top 5 words with the largest weights (positive or negative) under each topic vector. For reference, we compare the heatmap and topic vectors generated by our model with the NMF (Figure 6).

Figure 6 shows that the correlation between each word-topic generated by the SONMF is 0, as opposed to the correlated features generated by the NMF. Based on Table 8, each topic specifically points out the location and cause of the injury. We can interpret Topic 1 as on injury from falling (falling at ground-level leads to breaking/over-rotation of the hips), Topic 2 on ankle injury, Topic 3 on knee injury, Topic 4 on x-rays, and Topic 5 on soccer by looking at the words with positive weights. Each topic has its distinct interpretation, and the words with negative weights under the same topic refer to a completely different location and cause. For instance, Topic 3 and Topic 5 are almost mirrors to each other. The interpretations of these topics are also sensible, as it is unlikely that patients who injured their knees would also twist their ankles during soccer since people are likely to restrain themselves from further physical activities if any of these conditions happen. The contrast in meaning is more evident in the Symptoms of Stroke dataset. For Topics 1 and 3, we see that our model correctly identifies "left" from "right." For Topic 4, we also observe that "steady" and "unsteady" have been placed in the opposite signs under the same topic. This further exhibits our model's ability to cluster correlated terms while also differentiate between word clusters.

The SONMF and NMF both identified and captured the main topics of the Lower Extremity Injury dataset, and agrees with each other to a certain extent. Tables 8 and 9 show that three pairs of topics generated by the SONMF and NMF have a high degree of overlap. However, we observe that the topics generated by the SONMF are more distinct than the NMF. For instance, Topics 2, 4, and 5 generated by the NMF are related to fall-induced injuries, while Topics 1, 4, and 5 are related to leg-injury. In addition, the SONMF provides more information, both semantically and numerically compared to the NMF due to the additional sub-clustering property that the negative weights provide.

6.3. Interpretation of the Document-Topic Matrix

In this subsection, we present the interpretation of the generated document-topic vectors in **G**. The interpretation of the **G** matrix is the same for all the NMF methods, where each column (patient/document) in **X** is represented as a purely additive linear combination of the columns (topic vectors) generated in **F**. We can thus interpret a patient's record by examining the weights of each topic. Here we first present the violin plot of

Table 9. Words with the largest magnitude under the first five topics generated by the NMF for the Lower Extremity Injury and Stroke datasets.

(NMF) Lower Extremity Injury					(NMF) Symptoms of Stroke				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
leg	hip	xray	fall	day	right	stroke	episode	found	left
state	rotate	fracture	knee	leg	numb	team	resolve	staff	gait
right	glf	sent	twist	fall	arm	see	min	family	unsteady
numb	morphin	told	bend	drop	leg	page	symptom	command	arm
tingle	leg	done	bruise	twist	side	stat	last	lsn	grip

Abbreviations: glf (ground-level fall), ubl (Ubiquitin-like protein), gcs (glasgow coma scale).

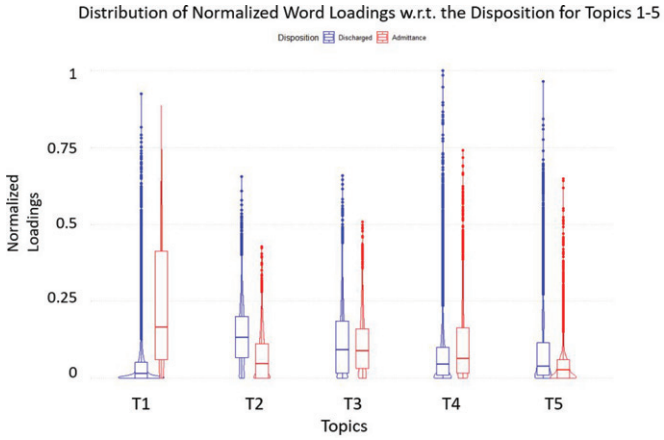

Figure 7. Violin plot of the normalized word loadings generated by the SONMF grouped by dispositions for the Lower Extremity Injury dataset. Blue represents the discharged patients, while red represents the admitted patients.

the loadings on the first five topics in **G** for patients who were in the Lower Extremity Injury dataset in Figure 7. Figure 7 indicates that Topics 1 and 2 are significantly different in their distributions of loadings. This suggests that patients weighted toward Topic 1 (injury due to ground-level falls) are more likely to be admitted, while patients weighted toward Topic 2 (ankle bruising/swelling) are more likely to be discharged. The remaining three topics are less clear to distinguish between these two dispositions, but Topic 4 (x-ray) has a higher rate of admittance, while Topic 5 (soccer) has a higher rate of discharge, which are intuitively sensible.

Lastly, we present the generated document-topic vectors by the SONMF for two representative notes from the Lower Extremity Injury dataset in Table 10, one for "admitted" and one for "discharged."

Table 10 shows that our model has captured the main sentiment of these notes according to the weights of these 5 topics. The above findings indicate our model is able to identify the main reasons behind patients' visits and their disposition status.

7. Conclusion

In this case study, we aim to build a classifier to predict patients' disposition from a triage notes dataset provided by the Alberta Medical Center, and show the potential advantages of using machine learning approaches on triage notes in addressing ED crowding. Additionally, we also intend to understand the main causes of the patients' visits and dispositions. The triage text data is challenging to model and interpret due to its high-dimensional and noisy structures. To address these data challenges, we proposed a SONMF as a topic model to bi-cluster the patients and words jointly into a lower dimension of topics. Our proposed method produces an orthogonal word-topic basis matrix, where each patient can be rerepresented as a strictly additive linear combination of these topics. The benefits of our method over the existing NMF methods are 2-fold. First, our method generates uncorrelated projection bases, which alleviate multicollinearity and over-fitting problems. This provides numerical stability and enhances classification performances using reduced latent features. Second, the generated topics themselves provide a meaningful interpretation, which helps the hospital understand patients' needs for each medical complaint.

We show that the text information contains significant predictive signals toward the final disposition of each patient by performing topic modeling and classification. These predictions can be directly implemented to a streamlined queue process to improve operational efficiency and reduce the overall waiting time and length of stay in ED, which consequently improves clinical outcomes, healthcare efficiency, and revenues. However, extra caution needs to take when implementing a machine learning model. Since most of the patients in an ED are in relatively vulnerable conditions, thus a poor assignment can be extremely dangerous and costly. Therefore, it is recommended that the implementation of these models should be considered cautiously and separately for each medical complaint, and should only be used when there is a high degree of confidence. Nevertheless, regardless of the classification performance of these models, the generated topic vectors are still beneficial in most situations, which can guide hospital administrators, doctors, and nurses in making better decisions for patients from a data-driven perspective.

We believe that this article potentially simulates new interests for further investigation to better aid the quality of emergency health care for hospital patients.

Table 10. Topic representation of selected triage notes (top: discharged, bottom: admitted) from the Lower Extremity Injury dataset.

Notes	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Tried to jump over fence and hit left knee, heard a pop, today pain increasing behind knee. Swollen since yesterday.	0	3.29	1.04	0.51	0.68
Glf on hardwood floor, landing on left hip, now c/o pain to left groin, non radiating, sharp, worse on movement. Given morphine by EMS.	9.14	1.13	1.87	0	0.75

Supplementary Materials

The supplementary material contains detailed derivation of the orthogonal preserving algorithm and additional numerical results. In addition, the details of how to reproduce the results in the manuscript is also provided.

Acknowledgments

The authors would like to acknowledge the editor, associate editor, and four anonymous referees for their critical and insightful comments in improving this article.

Funding

The authors also acknowledge the support for this project from the National Science Foundation grants DMS-1613190 and DMS-1821198. Additionally, Yutong Li and Ruoqing Zhu are supported by the University of Illinois at Urbana-Champaign through the NCSA Faculty Fellows program. Zhankun Sun is supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 21500517 and TRS T32-102/14N).

References

Abdi, H., and Williams, L. J. (2010), "Principal Component Analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 433–459. [1610,1620]

Aggarwal, C. C., and Reddy, C. K. (2013), *Data Clustering: Algorithms and Applications*, Boca Raton, FL: CRC Press. [1609,1610]

Anantharaman, V. (2008), "Impact of Health Care System Interventions on Emergency Department Utilization and Overcrowding in Singapore," *International Journal of Emergency Medicine*, 1, 11–20. [1609]

Barrett, L., Ford, S., and Ward-Smith, P. (2012), "A Bed Management Strategy for Overcrowding in the Emergency Department," *Nursing Economic*, 30, 82. [1609]

Boutsidis, C., and Gallopoulos, E. (2008), "SVD Based Initialization: A Head Start for Nonnegative Matrix Factorization," *Pattern Recognition*, 41, 1350–1362. [1613]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1619]

Chalfin, D. B., Trzeciak, S., Likourezos, A., Baumann, B. M., and Dellinger, R. P. (2007), "Impact of Delayed Transfer of Critically Ill Patients From the Emergency Department to the Intensive Care Unit," *Critical Care Medicine*, 35, 1477–1483. [1609]

Cichocki, A., and Phan, A.-H. (2009), "Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 92, 708–721. [1611,1612]

Cichocki, A., Zdunek, R., and Amari, S.-I. (2007), "Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization," in *International Conference on Independent Component Analysis and Signal Separation*, Springer, pp. 169–176. [1611,1612]

Ding, C. H., Li, T., and Jordan, M. I. (2010), "Convex and Semi-Nonnegative Matrix Factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 45–55. [1611,1612,1614,1615,1617,1619]

Ding, C., Li, T., Peng, W., and 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*, ACM, pp. 126–135. [1610,1611,1612]

Douma, M. J., Drake, C. A., O'Dochartaigh, D., and Smith, K. E. (2016), "A Pragmatic Randomized Evaluation of a Nurse-Initiated Protocol to Improve Timeliness of Care in an Urban Emergency Department," *Annals of Emergency Medicine*, 68, 546–552. [1609]

Dudani, S. A. (1976), "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE Transactions on Systems, Man, and Cybernetics*, 4, 325–327. [1619]

Eckart, C., and Young, G. (1936), "The Approximation of One Matrix by Another of Lower Rank," *Psychometrika*, 1, 211–218. [1613]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [1619]

Gillis, N., and Vavasis, S. A. (2013), "Fast and Robust Recursive Algorithms for Separable Nonnegative Matrix Factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, 698–714. [1611,1612]

Guan, Y., and Dy, J. (2009), "Sparse Probabilistic Principal Component Analysis," in *Artificial Intelligence and Statistics*, pp. 185–192. [1617]

Gupta, V., and Lehal, G. S. (2009), "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence*, 1, 60–76. [1618]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer. [1611,1618]

Hong, W. S., Haimovich, A. D., and Taylor, R. A. (2018), "Predicting Hospital Admission at Emergency Department Triage Using Machine Learning," *PLOS ONE*, 13, e0201016. [1609]

Hoyer, P. O. (2004), "Non-Negative Matrix Factorization With Sparseness Constraints," *Journal of Machine Learning Research*, 5, 1457–1469. [1612]

Hsieh, C.-J., and Dhillon, I. S. (2011), "Fast Coordinate Descent Methods With Variable Selection for Non-Negative Matrix Factorization," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1064–1072. [1612]

Ieraci, S., Digiusto, E., Sonntag, P., Dann, L., and Fox, D. (2008), "Streaming by Case Complexity: Evaluation of a Model for Emergency Department Fast Track," *Emergency Medicine Australasia*, 20, 241–249. [1610]

Kim, J., and Park, H. (2008), "Sparse Nonnegative Matrix Factorization for Clustering," Technical Report, Georgia Institute of Technology. [1610]

Kimura, K., Tanaka, Y., and Kudo, M. (2015), "A Fast Hierarchical Alternating Least Squares Algorithm for Orthogonal Nonnegative Matrix Factorization," in *Proceedings of the Sixth Asian Conference on Machine Learning*, Proceedings of Machine Learning Research (PMLR) (Vol. 39), D. Phung and H. Li, Nha Trang City, Vietnam, pp. 129–141. [1610,1611,1612,1614,1615,1619]

Kuhn, M. (2008), "Building Predictive Models in R Using the caret Package," *Journal of Statistical Software*, 28, 1–26. [1618]

Langville, A. N., Meyer, C. D., Albright, R., Cox, J., and Duling, D. (2006), "Initializations for the Nonnegative Matrix Factorization," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Citeseer, pp. 23–26. [1613]

——— (2014), "Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization," arXiv no. 1407.7299. [1613]

Lee, D. D., and Seung, H. S. (1999), "Learning the Parts of Objects by Non-Negative Matrix Factorization," *Nature*, 401, 788–791. [1610,1611,1612]

——— (2001), "Algorithms for Non-Negative Matrix Factorization," in *Advances in Neural Information Processing Systems*, pp. 556–562. [1610,1612,1614,1615,1617,1619]

Li, Y., Zhu, R., and Qu, A. (2019), "Package 'matrixfact'," available at *https://github.com/cralo31/MatrixFact*. [1614]

Lin, C.-J. (2007), "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation*, 19, 2756–2779. [1611]

Mareiniss, D. P. (2020), "The Impending Storm: Covid-19, Pandemics and Our Overwhelmed Emergency Departments," *The American Journal of Emergency Medicine*, 38, 1293–1294. [1610]

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., and Lin, C.-C. (2014), "e1071: Misc functions of the Department of Statistics (e1071), TU Wien," *R package version*, 1. [1619]

Morley, C., Unwin, M., Peterson, G. M., Stankovich, J., and Kinsman, L. (2018), "Emergency Department Crowding: A Systematic Review of Causes, Consequences and Solutions," *PLOS ONE*, 13, e0203316. [1609]

Paatero, P., and Tapper, U. (1994), "Positive Matrix Factorization: A Non-Negative Factor Model With Optimal Utilization of Error Estimates of Data Values," *Environmetrics*, 5, 111–126. [1610]

Peck, J. S., Benneyan, J. C., Nightingale, D. J., and Gaehde, S. A. (2012), "Predicting Emergency Department Inpatient Admissions to Improve Same-Day Patient Flow," *Academic Emergency Medicine*, 19, E1045–E1054. [1609]

Pines, J. M., Hilton, J. A., Weber, E. J., Alkemade, A. J., Al Shabanah, H., Anderson, P. D., Bernhard, M., Bertini, A., Gries, A., Ferrandiz, S., and Kumar, V. A. (2011), "International Perspectives on Emergency Department Crowding," *Academic Emergency Medicine*, 18, 1358–1370. [1620]

Powell, E. S., Khare, R. K., Venkatesh, A. K., Van Roo, B. D., Adams, J. G., and Reinhardt, G. (2012), "The Relationship Between Inpatient Discharge Timing and Emergency Department Boarding," *The Journal of Emergency Medicine*, 42, 186–196. [1610]

Qiao, H. (2015), "New SVD Based Initialization Strategy for Non-Negative Matrix Factorization," *Pattern Recognition Letters*, 63, 71–77. [1609,1613]

Qiu, S., Chinnam, R. B., Murat, A., Batarse, B., Neemuchwala, H., and Jordan, W. (2015), "A Cost Sensitive Inpatient Bed Reservation Approach to Reduce Emergency Department Boarding Times," *Health Care Management Science*, 18, 67–85. [1610]

RColorBrewer, S., and Liaw, M. A. (2018), "Package 'randomforest'," University of California, Berkeley, Berkeley, CA, USA. [1619]

Recht, B., Re, C., Tropp, J., and Bittorf, V. (2012), "Factoring Nonnegative Matrices With Linear Programs," in *Advances in Neural Information Processing Systems*, pp. 1214–1222. [1611]

Ripley, B., Venables, W., and Ripley, M. B. (2015), "Package 'class'," *The Comprehensive R Archive Network*. [1619]

Saghafian, S., Hopp, W. J., Van Oyen, M. P., Desmond, J. S., and Kronick, S. L. (2012), "Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments," *Operations Research*, 60, 1080–1097. [1610]

Salton, G., Wong, A., and Yang, C.-S. (1975), "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, 18, 613–620. [1609,1610,1618]

Schachtner, R., Poppel, G., and Lang, E. W. (2010), "A Nonnegative Blind Source Separation Model for Binary Test Data," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 57, 1439–1448. [1613]

Schein, A. I., Saul, L. K., and Ungar, L. H. (2003), "A Generalized Linear Model for Principal Component Analysis of Binary Data," in *AISTATS*, (Vol. 2), p. 10. [1613]

Shahnaz, F., Berry, M. W., Pauca, V. P., and Plemmons, R. J. (2006), "Document Clustering Using Nonnegative Matrix Factorization," *Information Processing & Management*, 42, 373–386. [1610]

Slawski, M., Hein, M., and Lutsik, P. (2013), "Matrix Factorization With Binary Components," in *Advances in Neural Information Processing Systems*, pp. 3210–3218. [1613]

Sprivulis, P. C., Da Silva, J.-A., Jacobs, I. G., Jelinek, G. A., and Frazer, A. R. (2006), "The Association Between Hospital Overcrowding and Mortality Among Patients Admitted via Western Australian Emergency Departments," *Medical Journal of Australia*, 184, 208–212. [1620]

Sun, B. C., Hsia, R. Y., Weiss, R. E., Zingmond, D., Liang, L.-J., Han, W., McCreath, H., and Asch, S. M. (2013), "Effect of Emergency Department Crowding on Outcomes of Admitted Patients," *Annals of Emergency Medicine*, 61, 605–611. [1609,1620]

Sun, Y., Heng, B. H., Tay, S. Y., and Seow, E. (2011), "Predicting Hospital Admissions at Emergency Department Triage Using Routine Administrative Data," *Academic Emergency Medicine*, 18, 844–850. [1609]

Suykens, J. A., and Vandewalle, J. (1999), "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, 9, 293–300. [1619]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1619]

Tomé, A., Schachtner, R., Vigneron, V., Puntonet, C., and Lang, E. (2015), "A Logistic Non-Negative Matrix Factorization Approach to Binary Data Sets," *Multidimensional Systems and Signal Processing*, 26, 125–143. [1613,1614,1617,1619]

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003), "Singular Value Decomposition and Principal Component Analysis," in *A Practical Approach to Microarray Data Analysis*, eds. D. P. Berrar, W. Dubitzky, and M. Granzow, Boston, MA: Springer, pp. 91–109. [1613]

Wang, W., Yang, D., Chen, F., Pang, Y., Huang, S., and Ge, Y. (2019), "Clustering With Orthogonal Autoencoder," *IEEE Access*, 7, 62421–62432. [1610]

Wen, Z., and Yin, W. (2013), "A Feasible Method for Optimization With Orthogonality Constraints," *Mathematical Programming*, 142, 397–434. [1611,1612]

Xue, Y., Tong, C. S., Chen, Y., and Chen, W.-S. (2008), "Clustering-Based Initialization for Non-Negative Matrix Factorization," *Applied Mathematics and Computation*, 205, 525–536. [1613]

Yaram, S. (2016), "Machine Learning Algorithms for Document Clustering and Fraud Detection," in *2016 International Conference on Data Science and Engineering (ICDSE)*, IEEE, pp. 1–6. [1609]

Yoo, J., and Choi, S. (2008), "Orthogonal Nonnegative Matrix Factorization: Multiplicative Updates on Stiefel Manifolds," in *IDEAL*, Springer, pp. 140–147. [1610,1611,1612]

Zhang, X., Guan, N., Lan, L., Tao, D., and Luo, Z. (2014), "Box-Constrained Projective Nonnegative Matrix Factorization via Augmented Lagrangian Method," in *2014 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1900–1906. [1613]

Zhang, X., Kim, J., Patzer, R. E., Pitts, S. R., Patzer, A., and Schrager, J. D. (2017), "Prediction of Emergency Department Hospital Admission Based on Natural Language Processing and Neural Networks," *Methods of Information in Medicine*, 56, 377–389. [1609]

Zhang, Z., Li, T., Ding, C., and Zhang, X. (2007), "Binary Matrix Factorization With Applications," in *Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007*, IEEE, pp. 391–400. [1613]