Non-negative Matrix Factorization for Binary Data

Jacob Søgaard Larsen and Line Katrine Harder Clemmensen
DTU Compute, Technical University of Denmark, Richard Petersens Plads, 2800, Lyngby, Denmark

Keywords: Non-negative Matrix Factorization, Binary Data, Binary Matrix Factorization, Text Modelling.

Abstract: We propose the Logistic Non-negative Matrix Factorization for decomposition of binary data. Binary data are frequently generated in e.g. text analysis, sensory data, market basket data etc. A common method for analysing non-negative data is the Non-negative Matrix Factorization, though this is in theory not appropriate for binary data, and thus we propose a novel Non-negative Matrix Factorization based on the logistic link function. Furthermore we generalize the method to handle missing data. The formulation of the method is compared to a previously proposed logistic matrix factorization without non-negativity constraint on the features. We compare the performance of the Logistic Non-negative Matrix Factorization to Least Squares Non-negative Matrix Factorization and Kullback-Leibler (KL) Non-negative Matrix Factorization on sets of binary data: a synthetic dataset, a set of student comments on their professors collected in a binary term-document matrix and a sensory dataset. We find that choosing the number of components is an essential part in the modelling and interpretation, that is still unresolved.

1 INTRODUCTION

Non-negative matrices are found in many different forms, from a general matrix with non-negative entries to the case with only binary entries. The latter is an interesting case used in many fields e.g. text data, sensory data etc. A common tool for preprocessing data by unsupervised decompsition is the Non-negative Matrix Factorization (NMF) proposed by Lee and Seung (Lee and Seung, 1999; Lee and Seung, 2001). One issue with the general NMF is that the resulting approximation is not bounded above, and hence not suitable for the binary case. Zhang et al. proposed the Binary Matrix Factorization that factorizes the binary data matrix X into two binary matrices W and H (Zhang et al., 2010). The interpretation of such a decomposition may be difficult, since the method does not estimate how important an entry in the components are, and therefore we will not consider this method for our purpose. Gillis proposed that when NMF is used on text data, the components are interpreted as topics (Gillis, 2014). The model also describes how important a topic is for each document and how important a term is for a topic. We adapt this approach and propose a logistic non.negative matrix factorization. Recently, Tomé et al. proposed a logistic but only partially non-negative matrix factorization, where the model allows for negative feature components (Tomé et al., 2015), whereas our method is strictly non-negative and explicit modelling of the threshold in the logistic sigmoidal. Tomé et al. further extended the model with a Lagrangian penalty on the two norm of the columns of W and H. Both our method and the methods by Tomé et al. uses a gradient based update scheme. Tomé et al. uses a constant step length, where we use an adaptive scheme to ensure non-negativity. Tomé et al. ensures non-negativity by projection. In order to evaluate how well the model generalizes the data, Tomé et al. uses a set-up with a test- and training-set, while we have generalized our method to handle missing data, thus enabling the use of cross-validation. The methods proposed by Tomé et al. is tested on synthetic data with binary basis vectors and the USPS digits. The emphasis is put on how well the model reconstructs data, while we focus on estimating the correct number of feature vectors and the interpretation of the model. Furthermore we test our model on sensory data and text data.

The training process and selecting the model complexity is another issue regarding NMF. Nielsen and Mørup proposed to marginalize missing data in order to perform cross-validation (CV) to choose the number of components in the model (Nielsen and Mørup, 2014), and we will use this approach in the paper.

Larsen, J. and Clemmensen, L..
Non-negative Matrix Factorization for Binary Data.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 555-563
ISBN: 978-989-758-158-8

2 METHOD

Non-negative Matrix Factorization (NMF) belongs to the family of Factor Analysis methods and was proposed by Lee and Seung (Lee and Seung, 1999; Lee and Seung, 2001). As the name states, it computes a low rank approximation of a $M \times N$ data matrix, X, consisting of the non-negative matrices $W \in \mathbb{R}^{M \times K}$, with entries $w_{i,d}$, and $H \in \mathbb{R}^{K \times N}$, with entries $h_{d,j}$, such that $x_{i,j} \approx \sum_{d=1}^{K} w_{i,d} h_{d,j}$. Generally the problem is formulated as in (1), where $D(\cdot,\cdot)$ is a distance measure. The cost function of interest obtained by letting $C(W,H) = D(X,WH)$.

$$\min_{W \in \mathbb{R}^{M \times K}, H \in \mathbb{R}^{K \times N}} D(X,WH), \quad st.\ W \geq 0, H \geq 0 \quad (1)$$

In this article, three different measures are used: Least Squares, KL-divergence and cross-entropy. The matrices W and H are computed using Multiplicative Updates (MU) described in Section 2.2, 2.3 and 2.4.1. Other methods for computing W and H are described in (Gillis, 2014, §3.1).

The derivative with respect to an arbitrary element $w_{m,n}$ or $h_{m,n}$ of either W or H, which for a general purpose will be termed $\beta_{m,n}$, with respect to a general cost function $C(\cdot)$, can be decomposed into a positive and a negative part (2a). Using a gradient descent method with the step size (2b), this results in the generic update formula (2c). This formula can be shown to converge towards a non-negative solution (Lee and Seung, 2001). The problem of finding the optimal solution is NP-hard (Gillis, 2014), this mean that an update procedure will converge towards a local minimum.

$$\frac{\partial C(\beta)}{\partial \beta_{m,n}} = \frac{\partial C(\beta)^{+}}{\partial \beta_{m,n}} - \frac{\partial C(\beta)^{-}}{\partial \beta_{m,n}} \quad (2a)$$

$$\beta_{m,n} = \beta_{m,n} - \eta_{m,n} \frac{\partial C(\beta)}{\partial \beta_{m,n}}, \quad \eta_{m,n} = \frac{\beta_{m,n}}{\frac{\partial C(\beta)}{\partial \beta_{m,n}}^{+}} \quad (2b)$$

$$\beta_{m,n} = \beta_{m,n} \frac{\frac{\partial C(\beta)}{\partial \beta_{m,n}}^{-}}{\frac{\partial C(\beta)}{\partial \beta_{m,n}}^{+}} \quad (2c)$$

It has been shown that NMF-algorithms improves convergence speed by updating the same factor multiple times (Gillis and Glineur, 2012). In this paper each factor was updated 10 times for each iteration.

2.1 Selecting the Number of Components

In order to determine the optimal number of components K in W and H, Nielsen and Mørup proposed a marginalization approach for handling missing data as an alternative to Expectation Maximization (Nielsen and Mørup, 2014). The method uses an indicator matrix R, with entries $r_{i,j}$, that is 1 if $x_{i,j}$ is present, and 0 otherwise. This enables the use of Cross Validation (CV) to estimate the generalization error of a model, the "one standard-error rule" (Hastie et al., 2009) can then be used to select the optimal number of components.

2.2 Least Squares

The Least Squares NMF uses the Frobenius norm as distance measure, resulting in the objective (3a). Together with the non-negativity constraints of W and H this constitutes the problem formulation. The multiplicative update formulas for the Least Squares formulation of the Non-negative Matrix Factorization (3b) and (3c) are based on taking gradient steps and step sizes of (3a) as defined in (2b).

$$C_{LS} = \sum_{i,j} (x_{i,j} - (WH)_{i,j})^2 \quad (3a)$$

$$w_{i,d} \leftarrow w_{i,d} \frac{(XH^T)_{i,d}}{(WHH^T)_{i,d}} \quad (3b)$$

$$h_{d,j} \leftarrow h_{d,j} \frac{(W^T X)_{d,j}}{(W^T WH)_{d,j}} \quad (3c)$$

The marginalization approach uses the slightly modified objective function (4a). (4b) and (4c) defines the gradients. By using step size 2b, the resulting MU formulas are (4d) and (4e).

$$C_{LS}^{R} = \sum_{i,j} r_{i,j} (x_{i,j} - (WH)_{i,j})^2 \quad (4a)$$

$$\nabla_{w_{i,d}} = \sum_{j} r_{i,j} ((WH)_{i,j} - x_{i,j}) h_{d,j} \quad (4b)$$

$$\nabla_{h_{d,j}} = \sum_{i} r_{i,j} ((WH)_{i,j} - x_{i,j}) w_{i,d} \quad (4c)$$

$$w_{i,d} \leftarrow w_{i,d} \frac{\sum_j r_{i,j} x_{i,j} h_{d,j}}{\sum_j r_{i,j} (WH)_{i,j} h_{d,j}} \quad (4d)$$

$$h_{d,j} \leftarrow h_{d,j} \frac{\sum_i r_{i,j} x_{i,j} w_{i,d}}{\sum_i r_{i,j} (WH)_{i,j} w_{i,d}} \quad (4e)$$

2.3 KL-divergence

Non-negative data are poorly approximated by a normal distribution, (Lee and Seung, 1999) proposes to use the divergence (5e) instead. The term 'divergence' is used instead of distance, since the measure is not symmetric in X and WH. This corresponds to a model in which $x_{i,j}$ has a Poisson distribution with mean $(WH)_{i,j}$ (Hastie et al., 2009, §14.6) and has received

its name since it reduces to the Kullback-Leibler divergence for $\sum_{i,j} x_{i,j} = \sum_{i,j} (WH)_{i,j} = 1$ (Lee and Seung, 2001).

The formulas for the multiplicative update scheme are given in Equation (5d) and (5e), they are derived from (5a) using the derivative and step size given in (5b) and (5c).

$$C_{KL} = -\sum_{i,j} x_{i,j} \log(WH)_{i,j} - (WH)_{i,j} \tag{5a}$$

$$\nabla_{w_{i,d}} = \sum_j h_{d,j} - \sum_j \frac{x_{i,j} h_{d,j}}{(WH)_{i,j}}, \quad \eta_{i,d} = \frac{w_{i,d}}{\sum_j h_{d,j}} \tag{5b}$$

$$\nabla_{h_{d,j}} = \sum_i w_{i,d} - \sum_i \frac{x_{i,j} w_{i,d}}{(WH)_{i,j}}, \quad \eta_{d,j} = \frac{h_{d,j}}{\sum_i w_{i,d}} \tag{5c}$$

$$w_{i,d} \leftarrow w_{i,d} \frac{\sum_j \frac{x_{i,j}}{(WH)_{i,j}} h_{d,j}}{\sum_j h_{d,j}} \tag{5d}$$

$$h_{d,j} \leftarrow h_{d,j} \frac{\sum_i w_{i,d} \frac{x_{i,j}}{(WH)_{i,j}}}{\sum_i w_{i,d}} \tag{5e}$$

Similar to the Least Squares setting, the marginalization approach can also be applied to the KL-divergence. The modified cost function is given in (6a). The multiplicative update formulas (6d) and (6e) are derived using the generic approach (2a)-(2c) with derivatives given in (6b) and (6c).

$$C^R_{KL} = \sum_{i,j} r_{i,j} \left(x_{i,j} \log(WH)_{i,j} - (WH)_{i,j} \right) \tag{6a}$$

$$\nabla_{w_{i,d}} = \sum_j r_{i,j} \left(h_{d,j} - \frac{x_{i,j} h_{d,j}}{(WH)_{i,j}} \right) \tag{6b}$$

$$\nabla_{h_{d,j}} = \sum_i r_{i,j} \left(w_{i,d} - \frac{x_{i,j} w_{i,d}}{(WH)_{i,j}} \right) \tag{6c}$$

$$w_{i,d} \leftarrow w_{i,d} \frac{\sum_j \frac{r_{i,j} x_{i,j}}{(WH)_{i,j}} h_{d,j}}{\sum_j r_{i,j} h_{d,j}} \tag{6d}$$

$$h_{d,j} \leftarrow h_{d,j} \frac{\sum_i w_{i,d} \frac{r_{i,j} x_{i,j}}{(WH)_{i,j}}}{\sum_i r_{i,j} w_{i,d}} \tag{6e}$$

2.4 Logistic NMF

For binary data a general NMF is not optimal in the sense that it maps onto the entire positive real space of numbers. The Logistic Non-negative Matrix Factorization is therefore proposed. The model is a generative model formed by a combination of NMF and logistic regression. The model is given in Equation (7) with $y_{i,j} = p(1|\sum_d w_{i,d} h_{d,j}, c_{i,j})$ being the probability of $y_{i,j}$ being 1 and $\sigma(\cdot)$ being the logistic sigmoid function (8a). The threshold $c_{i,j}$ is estimated in two different ways: a global constant applicable for all $y_{i,j}$ and a rank 1 approximation $u_i v_j$. See the description below.

$$a_{i,j} = \sum_d w_{i,d} h_{d,j} - c_{i,j}, \quad w_{i,d},\ h_{d,j}, c_{i,j} \geq 0 \tag{7a}$$

$$y_{i,j} = \sigma(a_{i,j}) = p(1|\sum_d w_{i,d} h_{d,j}, c_{i,j}) \tag{7b}$$

$$p(0|\sum_d w_{i,d} h_{d,j}, c_{i,j}) = 1 - p(1|\sum_d w_{i,d} h_{d,j}, c_{i,j}) \tag{7c}$$

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{8a}$$

$$\frac{d\sigma}{da} = \sigma(1-\sigma) \tag{8b}$$

General Cost Function. The model (7) applied to a 0-1 coded matrix X, lead to the Likelihood function (9a) and the negative log Likelihood (9b).

$$p(X|W,H,c_{i,j}) = \prod_{i,j} y_{i,j}^{x_{i,j}} (1-y_{i,j})^{1-x_{i,j}} \tag{9a}$$

$$\begin{aligned} C_{LL}(W,H,c_{i,j}) &= -\log\left(p(X|W,H,c_{i,j})\right) \\ &= -\sum_{i,j} x_{i,j} \log(y_{i,j}) + (1-x_{i,j}) \log(1-y_{i,j}) \end{aligned} \tag{9b}$$

Marginalized Cost Function. In case of missing data, a generalization of (9) is given as the marginalized Likelihood function in (10a) and the corresponding negative log Likelihood (10b).

$$p_r(X|W,H,c_{i,j}) = \prod_{i,j} \left(y_{i,j}^{x_{i,j}} (1-y_{i,j})^{1-x_{i,j}} \right)^{r_{i,j}} \tag{10a}$$

$$\begin{aligned} C^R_{LL}(W,H,c_{i,j}) &= -\log\left(p_r(X|W,H,c_{i,j})\right) \\ &= -\sum_{i,j} r_{i,j} \left(x_{i,j} \log(y_{i,j}) + (1-x_{i,j}) \log(1-y_{i,j}) \right) \end{aligned} \tag{10b}$$

2.4.1 Determining the Parameters

In order to determine the parameters of the model, the optimization problem (11) is formulated for the marginalized negative log Likelihood, as:

$$\min_{W,H,c_{i,j}} C^R_{LL}(W,H,c_{i,j}) \tag{11a}$$

$$\text{subject to. } W \in \mathbb{R}^{M \times K},\ W \geq 0 \tag{11b}$$

$$H \in \mathbb{R}^{K \times N},\ H \geq 0 \tag{11c}$$

$$c_{i,j} \in \mathbb{R},\ c_{i,j} \geq 0 \tag{11d}$$

Global Threshold. Using a global constant $c_{i,j} = c$, the optimization problem is convex (Boyd and Vandenberghe, 2009, §7.1) separately in W, H and c, henceforth this variant is known as Global thresh NMF. Therefore it is solved using MU based on a gradient descend method. The partial derivatives with respect to $w_{i,d}$, $h_{d,j}$ and c are given in (12). They are derived using the Chain Rule, and the derivative of the logistic sigmoid function given in (8b).

$$\frac{\partial C_{LL}R(W,H,c)}{\partial w_{i,d}} = \sum_j r_{i,j}(y_{i,j} - x_{i,j})h_{d,j} \tag{12a}$$

$$\frac{\partial C^R_{LL}(W,H,c)}{\partial h_{d,j}} = \sum_i r_{i,j}(y_{i,j} - x_{i,j})w_{i,d} \tag{12b}$$

$$\frac{\partial C^R_{LL}(W,H,c)}{\partial c} = \sum_{i,j} r_{i,j}(x_{i,j} - y_{i,j}) \tag{12c}$$

Using the step size defined in (2b), this leads to the multiplicative update formulas (13).

$$w_{i,d} \leftarrow w_{i,d}\frac{\sum_j r_{i,j}x_{i,j}h_{d,j}}{\sum_j r_{i,j}y_{i,j}h_{d,j}} \tag{13a}$$

$$h_{d,j} \leftarrow h_{d,j}\frac{\sum_i r_{i,j}x_{i,j}w_{i,d}}{\sum_i r_{i,j}y_{i,j}w_{i,d}} \tag{13b}$$

$$c \leftarrow c\frac{\sum_{i,j} r_{i,j}y_{i,j}}{\sum_{i,j} r_{i,j}x_{i,j}} \tag{13c}$$

Rank 1 Approximation of Threshold. When introducing a rank 1 approximation of the threshold for each (i,j), such that $c_{i,j} = u_i v_j$, the problem is still convex in u_i and v_j separately, henceforth this method is known as Max thresh NMF. Let u_i be the i'th row of U and v_j be the j'th column of V, the partial derivatives are then as shown in (14)

$$\frac{\partial C^R_{LL}(W,H,U,V)}{\partial u_i} = \sum_j r_j(x_{i,j} - y_{i,j})v_j \tag{14a}$$

$$\frac{\partial C^R_{LL}(W,H,U;V)}{\partial v_j} = \sum_i r_j(x_{i,j} - y_{i,j})u_i \tag{14b}$$

Using the step size defined in (2b), the update formulas are then given in (15).

$$u_i \leftarrow u_i\frac{\sum_j r_{i,j}y_{i,j}v_j}{\sum_j r_{i,j}x_{i,j}v_j} \tag{15a}$$

$$v_j \leftarrow v_j\frac{\sum_i r_{i,j}y_{i,j}u_i}{\sum_i r_{i,j}x_{i,j}u_i} \tag{15b}$$

2.4.2 Constraints

The update formulas introduced in (13c), (15a) and (15b) introduce a risk of dividing by zero, or in the case of a sparse matrix, the numerator may be much larger than the denominator. This, together with (7a) introduces the risk of both WH and c_{ij} exploding in size. This is avoided by adding the constraint (16) to the optimization problem (11) with λ being a positive constant.

$$c_{i,j} \leq \lambda \tag{16}$$

3 SIMULATED DATA

3.1 Generating Simulated Data

The four NMF variants are compared on simulated data where we know the true underlying components. The data simulates text data, which is ensured by putting the following conditions on W and H.

1. Columns of W are sparse
2. $\sum_k W_{ik} = 1$
3. Rows of H are exponentially distributed

1.: Each component represent a topic (Gillis, 2014), but each topic is not necessarily present in all documents. 2.: The pure documents do not contain any noise, and are therefore represented only by the constructed topics. 3.: Terms are exponentially distributed in natural languages (Paukkeri, 2012).

To investigate how the two Logistic NMF methods behave, compared to LS NMF and KL NMF when applied to text data, three problems have been created, with varying degree of sparsity in the score vectors

P. 1 25 % entries in score vectors

P. 2 50 % entries in score vectors

P. 3 75 % entries in score vectors

The data matrix $\tilde{X}$ is then generated as described in (17), with τ being a positive constant.

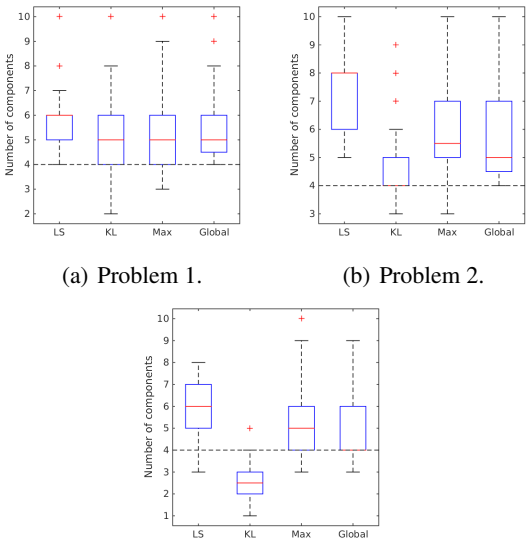
$$a = WH - \tau \tag{17a}$$

$$X = \frac{1}{1 + exp(-a)} \tag{17b}$$

$$\tilde{X} = \text{round}(X) \tag{17c}$$

3.2 Analysing Simulated Data

In all the problems, 4 topics are simulated with columns of W and rows of H having length 50. The resulting data matrix is therefore of size 50×50. The problems are simulated 100 times and modelled by each NMF method. For each simulation, the dataset is



(a) Problem 1. (b) Problem 2.

(c) Problem 3.

Figure 1: Box plots of estimated number of components. Length of whiskers are 1.5 IQR. Dashed line indicate true number of components.

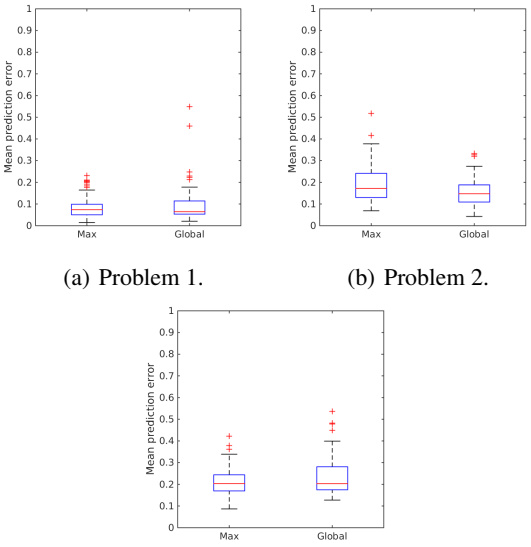
divided into 11 parts, the first 10 parts are used to estimate the optimal number of components by 10 fold CV. The last part is used to report how well the model predicts unknown data.

Analysing the Estimated Number of Components. Figure 1 shows box plots of the estimated number of components for each method. By performing a t-test with significance level $\alpha = 0.05$ it is concluded that none of the methods estimate the correct number of components (4). By using a Welch t-test at significance level $\alpha = 0.05$ it is tested whether the methods estimate different number of components. Instances where the test level is not significant are written with bold face in Table 1.

Table 1: P-values for problems and methods where the methods estimate the same number of components.

	P. 1	P. 2	P. 3
LS vs. Glob. thresh	**0.19**	$1.2 \cdot 10^{-6}$	$< 10^{-10}$
KL vs. Max thresh	**0.087**	$2.7 \cdot 10^{-8}$	$< 10^{-10}$
Glob. vs. Max	**0.22**	**0.23**	**0.62**

Analysing the Error Level. Neither Least Squares error (3a), KL-divergence (5a) or Cross-entropy (9b) are directly comparable. Hence only the two Logistic NMF methods are comparable in terms of prediction error. Figure 2 shows box plot of the mean cross-entropy of prediction when using the optimal number of components. Using a Welch t-test, a significance level $\alpha = 0.05$ reveal that for all of the problems, the

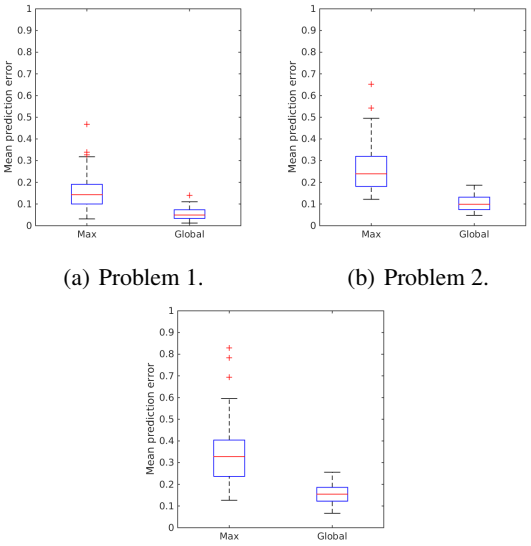


(a) Problem 1. (b) Problem 2.

(c) Problem 3.

Figure 2: Comparison of the mean cross-entropy of prediction for Max NMF and Const NMF with optimal number of components. Length of whiskers are 1.5 IQR.

prediction error is considered the same for the two methods. The p-values are 0.19, 0.12 and 0.08 for Problem 1, 2 and 3 respectively.

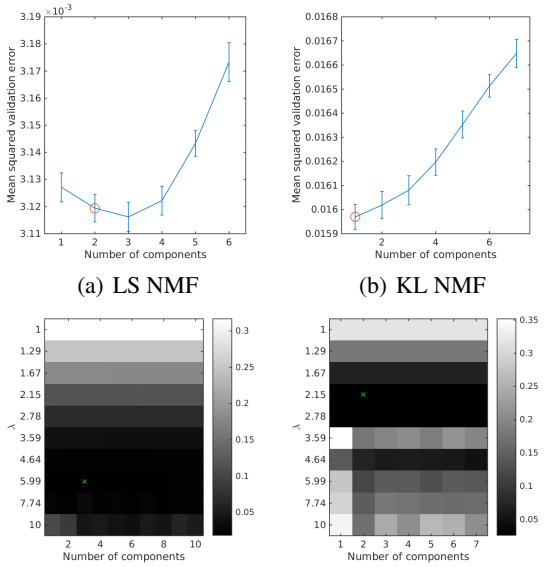


(a) Problem 1. (b) Problem 2.

(c) Problem 3.

Figure 3: Comparison of the mean cross-entropy of prediction for Max NMF and Const NMF with 4 component models. Length of whiskers are 1.5 IQR.

Figure 3 shows box plot of the mean cross-entropy of prediction when 4 components are used for both Global thresh NMF and Max thresh NMF. A Welch t-test with significance level $\alpha = 0.05$ is performed to



(a) LS NMF (b) KL NMF

(c) Global thresh NMF (d) Max thresh NMF

Figure 4: CV error for text data. (a) and (b): Mean CV error with standard error bars. (c) and (d): Mean CV error subtracted one standard error. Circles and crosses mark the optimal model complexity with regard to "one standard-error rule" (Hastie et al., 2009).

test whether the error levels are significantly different. The test value for all three problems are $p < 10^{-10}$, revealing that for all three problems Global thresh NMF have lower prediction error than Max thresh NMF.

4 REAL DATA

Two different real datasets are analysed using the four NMF variants: A dataset consisting of student comments and a sensory dataset (Randall, 1989).

4.1 Text Data

A collection of 10579 student comments on professors at various American universities collected at www.ourumd.com is being analysed. After preprocessing (including correcting misspellings, stemming and replacing numbers and names with relevant tags) the collection holds 9400 different words. The resulting Term Document Matrix (TDM) X is therefore of size 10579×9400. The estimated models are analysed using the procedure described in (Gillis, 2014). The interpretation of W and H are swapped, since the i'th document is collected in the i'th row of X.

The ability of each method to generalize the data, is estimated using 5-fold CV. In order to avoid local minima, each fold is re-estimated 10 times. Figure 4 shows the estimated generalization error and the chosen model complexity for each method. The dominant words of each topic are shown in Tables 2-5.

Table 2: Most dominant words for LS NMF.

Component 1			
exam	lectur	num	grade
question	final	test	class
help	name	book	studi
good	materi	point	homework
hard	problem	note	read

Component 2			
num	class	name	professor
cours	interest	great	!
student	teacher	easi	lot
recommend	teach	learn	good
paper	materi	read	make

Table 3: Most dominant words for KL NMF.

Component 1			
num	class	name	professor
cours	lectur	exam	grade
good	easi	lot	student
help	materi	test	read
teacher	teach	!	question

Table 4: Most dominant words for Global thresh NMF

Component 1			
num	lectur	exam	class
question	studi	easi	onlin
slide	answer	post	attent
test	note	choic	pai
multipl	hard	name	attend

Component 2			
num	class	name	cours
professor	exam	good	materi
help	lectur	cover	start
student	prepar	level	come
found	teach	will	school

Component 3			
num	class	name	particip
essai	group	present	grade
easi	fun	paper	interest
core	read	project	!
short	page	midterm	requir

4.2 Wine Data

The four variants of NMF are applied to the 'Bitterness of Wine' dataset (Randall, 1989). The data is

Table 5: Most dominant words for Max thresh NMF.

Component 1			
offic	ve	look	complet
tell	averag	feel	major
hour	practic	high	sai
ask	school	minut	num
show	mean	gener	give

Component 2			
multipl	post	answer	question
choic	final	slide	onlin
concept	attend	class	num
midterm	grade	point	requir
exam	semest	studi	consist

represented by introducing the 7 binary variables

- Contact (1 if ”yes”, 0 if ”no”)
- Temperature (1 if ”warm”, 0 if ”cold”)
- Rating 1
- Rating 2
- Rating 3
- Rating 4
- Rating 5

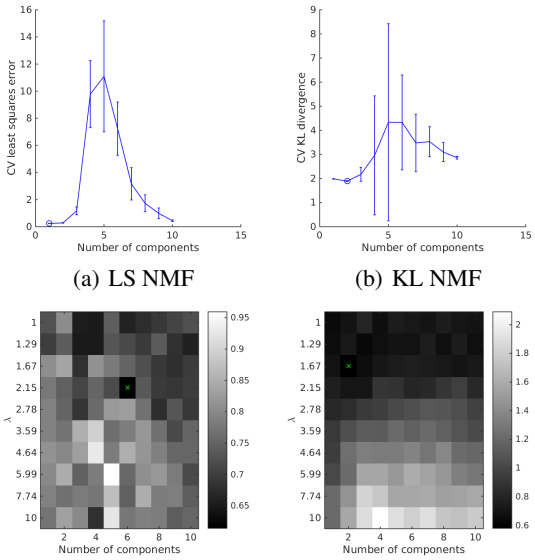
The generalization error of each of the models is estimated using 10-fold Cross validation. In order to avoid local minima in the training process, each fold is trained 30 times and the model with lowest validation error is reported. The model complexity is chosen using ”one standard-error rule” (Hastie et al., 2009). The estimated generalization error and chosen model complexities are shown in Figure 5.

Figure 6 shows the components of the estimated 6 component model using Global thresh NMF. From the rows of H it is seen the components basically describe variable each - except for Contact which is described by both component 5 and 6, further component 4 describe both Temperature and Rating 5.

5 DISCUSSION

5.1 Simulated Data

On the simulated problems we saw that the assumptions regarding the distribution of the noise, and thereby the choice of method, influences the how many components is estimated as being optimal. We observed that all four methods have a tendency to estimate too many components as being optimal even though the ”one standard-error” rule was used to choose the number of components in the models.



(a) LS NMF (b) KL NMF

(c) Global thresh NMF (d) Max thresh NMF

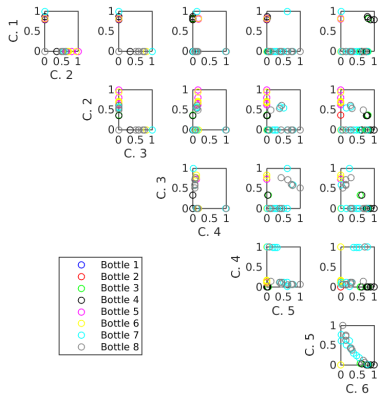
Figure 5: CV error for wine data. (a) and (b): Mean CV error with standard error bars. (c) and (d): Mean CV error subtracted one standard error. Circles and crosses mark the optimal model complexity with regard to ”one standard-error rule” (Hastie et al., 2009).

When changing the number of components, the interpretation of a model may change, as the distribution of signal among the components is changed. Thus, the use of extra components compared to the true number of components (4) may hinder the interpretation of the estimated model. A low generalization error is as important as estimating the true number of component when building a model. Furthermore, it is observed that the size of the error in estimating the number of components, is problem dependent.

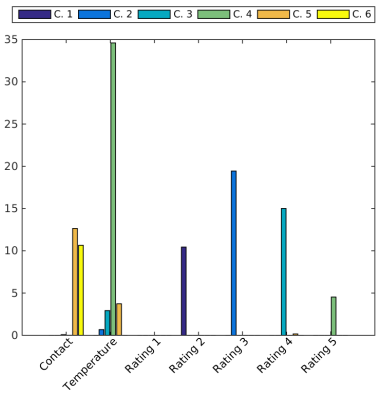
The error levels were compared for the Max thresh NMF and Global thresh NMF. It was seen that when the optimal number of components were used, the two methods performed with the same error level. When the true number of components were used, the Global thresh NMF performed significantly better than Max thresh NMF.

5.2 Text Data

The components or topics extracted from the collection of student comments are clearly related to teaching and exams. The logistic NMF’s are able to extract at least as many components from the data as LS NMF and more components than KL NMF. This indicate that KL NMF are not as suited for binary data as the other methods.



(a) Columns of W plotted against each other.



(b) Bar plot of rows of H.

Figure 6: Visualization of the components of estimated 6 component model using Global thresh NMF.

5.3 Wine Data

When estimating the optimal number of components, it was observed that this issue was dependent on the method. LS NMF, KL NMF and Max thresh NMF estimated that 1-2 components should be used while Global thresh NMF estimated that 6 components were optimal. Further, LS NMF and KL NMF was seen to have large standard errors.

The 6 component model determined with Global thresh NMF was presented. It was seen that each variable to some degree was described by a component. Furthermore it was observed that the variable Rating 1 was not described by any of the components. The variable Rating 1 is 1 in 5 samples while being 0 in 67 samples, i.e. approx. 7% of the samples, which may have influenced the model.

6 CONCLUSION

We presented the two well known variants of Non-negative Matrix Factorization and proposed the logistic Non-negative Matrix Factorization for binary data. The proposed method was presented in two variants; one with a global threshold for the logistic sigmoid function and one with a rank one approximation (max threshold).

The four NMF methods were applied to a collection of student comments regarding professors at various universities. It was seen that all methods were able to extract components that were describing teaching and exam. Global thresh NMF and Max thresh NMF were able to validate more components than the two usual methods: LS NMF and KL NMF.

The "Bitterness of Wine" dataset (Randall, 1989) was analysed using all four NMF methods. It was seen that methods which had a large standard error when doing Cross Validation, either estimated few components (LS, KL and Max) or many components (Global).

The four NMF methods were also compared on simulated data with four underlying components. The methods had good error convergence. However, the methods had a tendency to choose too many components. Furthermore, it was seen that the bias in estimating the number of components was problem dependent.

The interpretation may change with a varying number of components and as this is where we observed the most issues for all four studies, we recommend that future work should investigate methods to estimate the number of components.

ACKNOWLEDGEMENTS

The authors would like to thank Jacob Kogan from Department of Mathematics and Statistics at Univerty of Maryland for collecting the student comments from www.ourumd.com.

REFERENCES

Boyd, S. and Vandenberghe, L. (2009). *Convex Optimization*. Cambridge University Press.

Gillis, N. (2014). The why and how of nonnegative matrix factorization. *ArXiv e-prints*.

Gillis, N. and Glineur, F. (2012). Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization. *NEURAL COMPUTATION*, 24(4):1085–1105.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, 2nd edition.

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

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13(13):556–562.

Nielsen, S. F. V. and Mørup, M. (2014). Non-negative tensor factorization with missing data for the modeling of gene expressions in the human brain. In *2014 IEEE International workshop on Machine Learning for Signal Processing*.

Paukkeri, M.-S. (2012). *Language- and domain- independent text mining*. Doctorial Dissertations. Aalto University.

Randall, J. (1989). The analysis of sensory data by generalised linear model. *Biometrical journal*, 7:pp. 781–793.

Tomé, A. M., Schachtner, R., Vigneron, V., Puntonet, C. G., and Lang, E. W. (2015). A logistic non-negative matrix factorisation approach to binary data sets. *Multidim Syst Sign Process*, 26:125–143.

Zhang, Z., Li, T., Ding, C., and Zhang, X. (2010). Binary matrix factorization with applications. *Data Mining and Knowledge Discovery*, 20(1):28–52.