A majorization-minimization algorithm for nonnegative binary matrix factorization

Paul Magron, Cédric Févotte

A majorization-minimization algorithm for nonnegative binary matrix factorization[*]

Paul Magron[†], Cédric Févotte[‡]

Abstract

This paper tackles the problem of decomposing binary data using matrix factorization. We consider the family of mean-parametrized Bernoulli models, a class of generative models that are well suited for modeling binary data and enables interpretability of the factors. We factorize the Bernoulli parameter and consider an additional Beta prior on one of the factors to further improve the model's expressive power. While similar models have been proposed in the literature, they only exploit the Beta prior as a proxy to ensure a valid Bernoulli parameter in a Bayesian setting; in practice it reduces to a uniform or uninformative prior. Besides, estimation in these models has focused on costly Bayesian inference. In this paper, we propose a simple yet very efficient majorization-minimization algorithm for maximum a posteriori estimation. Our approach leverages the Beta prior whose parameters can be tuned to improve performance in matrix completion tasks. Experiments conducted on three public binary datasets show that our approach offers an excellent trade-off between prediction performance, computational complexity, and interpretability.

Keywords— Binary data, nonnegative matrix factorization, mean-parametrized Bernoulli model, majorization-minimization.

1 Introduction

Binary data are encountered in a variety of research fields such as paleontology [1], electoral data analysis [2], recommender systems [3], or binary image classification [4].

A popular approach for decomposing tabular data is matrix factorization (MF) [5], which consists in expressing a data matrix $\mathbf{Y}$ as the approximate product (of lower rank) of two matrices $\mathbf{W}$ and $\mathbf{H}$, along with constraints such as nonnegativity [6]. MF can be cast in a probabilistic framework [7], where the data is assumed to follow some distribution whose parameter is structured as $\mathbf{WH}$. Over the years, many distributions have been used in order to account for specific properties of the data at hand. For instance, Poisson generative models are appropriate for modeling count data encountered in recommender systems [8, 9]. However, Poisson or Gaussian [10] models are commonly used to analyze binary data for practical reasons, even though they are not tailored for this task.

In order to explicitly account for the binary nature of the data, models based on the Bernoulli distribution have been proposed. These can be divided into two categories. On the one hand, models in the logistic principal component analysis (PCA) [11] family exploit a *link* function in order to map the factorization to the space of Bernoulli parameters $[0, 1]$, i.e., $\mathbb{E}(\mathbf{Y}|\mathbf{WH}) = \sigma(\mathbf{WH})$. On the other hand, *mean-parametrized* models [12] directly factorize the Bernoulli parameter, i.e., $\mathbb{E}(\mathbf{Y}|\mathbf{WH}) = \mathbf{WH}$. Mean-parameterization is a useful property because it readily allows to interpret the factors and the approximation $\mathbf{WH}$. Nonetheless, to ensure a valid Bernoulli parameter, it is required to additionally constrain $\mathbf{W}$ and $\mathbf{H}$, e.g., using Dirichlet and Beta priors in a Bayesian setting [12]. However, in these approaches the Beta prior's sole purpose it to serve as a proxy to ensure a valid parameter: in practice it simplifies to a uniform or uninformative prior [4, 12], thus its full potential remains to be assessed. Besides, variational [4] or sampling [12] estimation schemes are computationally costly, as pointed out in [12]. Finally, the expectation-maximization (EM) algorithm from [1] does not use any prior, and relies on an augmented model with hidden variables, which somehow complicates the derivations.

In this paper, we propose a new approach for estimating a mean-parametrized Bernoulli model, which alleviates the aforementioned issues. We summarize hereafter our contributions and their advantages:

1. We consider a Beta prior on $\mathbf{H}$ with tunable hyperparameters: this allows to optimally exploit this prior, whose impact on performance was until now left to explore.

[*] This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 681839 (project FACTORY).

[†] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France (e-mail: firstname.lastname@inria.fr).

[‡] IRIT, Université de Toulouse, CNRS, Toulouse, France (e-mail: firstname.lastname@irit.fr).

2. We consider maximum a posteriori (MAP) estimation of the model's parameters with majorization-minimization (MM) [13, 14]. This yields easy-to-implement updates with no extra parameters (such as step sizes) to adjust. It is considerably more straightforward to derive than previous approaches. It notably generalizes the prior-free EM method from [1], without using hidden variables and data augmentation.
3. We compare our method with state-of-the-art logistic PCA for matrix completion. We show that our method exhibits an excellent trade-off between prediction performance, computational complexity, and interpretability.

The rest of this paper is structured as follows. Section 2 introduces the generative model and underlines its connection with related works. Section 3 presents the MAP estimation procedure with MM. Experiments are reported in Section 4. Finally, Section 5 draws some concluding remarks.

Mathematical notations:

- a (regular): scalar.
- $\mathbf{a}$ (lower case, bold font): vector.
- $\mathbf{A}$ (capital, bold font): matrix. The (m, n)-th entry of $\mathbf{A}$ is denoted $[\mathbf{A}]_{m,n} = a_{m,n}$.

2 Binary data models

In this section we briefly present Bernoulli-based MF models for binary data. For a more detailed overview, we refer the interested reader to Table 1 from [12].

2.1 Logistic PCA

Let us consider a binary data matrix $\mathbf{Y} \in \{0, 1\}^{M \times N}$. Logistic PCA builds upon the following generative model:

$$y_{m,n} \sim \text{Bernoulli}(\sigma([\mathbf{WH}]_{m,n})), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{M \times K}$, $\mathbf{H} \in \mathbb{R}^{K \times N}$, K is the rank of the factorization, and $\sigma : x \to 1/(1 + e^{-x})$ is the logistic function,[1] which maps the factorization to the range $[0, 1]$. Logistic PCA has been computed using a variety of techniques, including variational approaches [16], gradient descent [17], and alternating least squares [18].

Thanks to the mapping function, no constraint is needed on the factors to ensure a valid Bernoulli parameter. Nonetheless, several approaches have additionally enforced the nonnegativity of one [19] or two [20] factors, or leveraged Gaussian priors [21, 22]. Despite its popularity and performance, a drawback of logistic PCA stems from the fact that the link function hampers interpretability of the decomposition, which is often a desired feature, e.g., for analyzing econometric data [23].

2.2 Mean-parametrized Bernoulli models

To alleviate the aforementioned issue, mean-parametrized Bernoulli models [12] have been proposed, such that:

$$y_{m,n} \sim \text{Bernoulli}([\mathbf{WH}]_{m,n}). \tag{2}$$

In order to guarantee that $[\mathbf{WH}]_{m,n} \in [0, 1]$, it is mandatory to impose additional constraints on the factors, such as:

$$\forall m, \sum_k w_{m,k} = 1 \quad \text{and} \quad \forall (k, n), h_{k,n} \leq 1, \tag{3}$$

along with the nonnegativity of $\mathbf{W}$ and $\mathbf{H}$.[2] As a result, we call this family of models NBMF, which stands for *nonnegative binary matrix factorization.* Maximum likelihood estimation in a such a model was proposed in [1] by introducing hidden variables and deriving an EM algorithm, yielding a variant that we name NBMF-EM. In a Bayesian setting, it is common to consider the following priors instead of (3):

$$\mathbf{w}_m \sim \text{Dirichlet}(\boldsymbol{\gamma}) \quad \text{and} \quad h_{k,n} \sim \text{Beta}(\alpha_k, \beta_k), \tag{4}$$

where $\mathbf{w}_m$ denotes the m-th row of $\mathbf{W}$. Such models have been estimated using variational Bayesian approaches [4] or collapsed Gibbs sampling [12]. In these approaches, the priors only serve as proxy to ensure valid Bernoulli parameters: in practice the Beta prior reduces to a uniform ($\alpha_k = \beta_k = 1$) [12] or uninformative ($\alpha_k = \beta_k = 1/2$) [4] prior. Therefore, these approaches have not actually assessed the potential of carefully tuning the parameters of this prior. We will show that tuning these parameters can lead to significant improvements.

[1] Note that alternative link functions have been considered, e.g., in [15].

[2] Other constraints are possible, as will be detailed in Section 3.3.

3 Proposed method

Let us consider the NBMF generative model with the Beta prior in (4) for $\mathbf{H}$ and the sum-to-one constraint in (3) for $\mathbf{W}$. We now derive our MM algorithm for MAP estimation.

3.1 Objective

We seek to minimize $f(\mathbf{W},\mathbf{H})+g(\mathbf{H})$ under the constraints (3), where f is the negative Bernoulli log-likelihood:

$$f(\mathbf{W},\mathbf{H}) = -\log p(\mathbf{Y}|\mathbf{WH}) = -\sum_{m,n} y_{m,n}\log([\mathbf{WH}]_{m,n}) + (1-y_{m,n})\log(1-[\mathbf{WH}]_{m,n}), \tag{5}$$

and g is the negative Beta log-prior:

$$g(\mathbf{H}) = -\log p(\mathbf{H}) = -\sum_{k,n}(\alpha_k-1)\log(h_{k,n}) + (\beta_k-1)\log(1-h_{k,n}). \tag{6}$$

We account for the constraint on $\mathbf{W}$ via the method of Lagrange multipliers (denoted λ_m), thus the problem becomes that of finding a stationary point for:

$$\mathcal{L}(\mathbf{W},\mathbf{H},\boldsymbol{\lambda}) = f(\mathbf{W},\mathbf{H}) + g(\mathbf{H}) + \sum_m \lambda_m\left(\sum_k w_{m,k}-1\right). \tag{7}$$

Note that we do not explicitly consider the constraint $h_{k,n}\leq 1$ nor the nonnegativity of the factors, since we will prove in Section 3.3 that these automatically hold in our algorithm.

3.2 Estimation with majorization-minimization

We consider MM [14], which has shown powerful for estimating MF models in many settings [24, 25, 26]. In a nutshell, if we consider minimization of a function ϕ with parameters θ and current estimate $\tilde{\theta}$, MM consists in constructing and minimizing a tight upper bound ψ such that:

$$\forall\theta,\quad \phi(\theta)\leq\psi(\theta,\tilde{\theta}) \quad\text{and}\quad \phi(\tilde{\theta})=\psi(\tilde{\theta},\tilde{\theta}). \tag{8}$$

Then, it can easily be shown [13] that ϕ is non-increasing under the following update scheme: $\tilde{\theta}\leftarrow\arg\min_\theta\psi(\theta,\tilde{\theta})$. Here we consider a block-descent strategy in which $\mathbf{H}$ and $\mathbf{W}$ are updated in turn, which produces a valid descent algorithm.

3.2.1 Update on H

Let us first present the update for $\mathbf{H}$. To that end, we seek to majorize f defined in (5) with respect to $\mathbf{H}$, with $\mathbf{W}$ fixed. Denoting the current estimate by $\tilde{\mathbf{H}}$, the first term in (5), denoted $f_1(\mathbf{H})$, can be rewritten as:

$$\log\left(\sum_k w_{m,k}h_{k,n}\right) = \log\left(\sum_k \tilde{\rho}_{m,n,k}\frac{w_{m,k}}{h_{k,n}\tilde{\rho}_{m,n,k}}\right), \tag{9}$$

where $\tilde{\rho}_{m,n,k}=w_{m,k}\tilde{h}_{k,n}/\tilde{y}_{m,n}$, and $\tilde{y}_{m,n}=\sum_l w_{m,l}\tilde{h}_{l,n}$. Since the weights $\tilde{\rho}_{m,n,k}$ are nonnegative and $\sum_k\tilde{\rho}_{m,n,k}=1$, and since the function $x\to-\log x$ is convex, we can majorize f_1 using Jensen inequality, such that $f_1(\mathbf{H})\leq\psi_1(\mathbf{H},\tilde{\mathbf{H}})$ with:

$$\psi_1(\mathbf{H},\tilde{\mathbf{H}}) = -\sum_{m,n,k}\frac{y_{m,n}w_{m,k}\tilde{h}_{k,n}}{\tilde{y}_{m,n}}\log\left(\frac{h_{k,n}\tilde{y}_{m,n}}{\tilde{h}_{k,n}}\right). \tag{10}$$

To obtain a majorization for the second term $f_2(\mathbf{H})$ in (5), a first naive approach consists in exploiting the convexity of $x\to-\log(1-x)$ similarly as above. However, this leads to an intractable minimization step. Instead, we exploit the constraint $\sum_k w_{m,k}=1$ to rewrite $f_2(\mathbf{H})$ as follows:

$$\log(1-\sum_k w_{m,k}h_{k,n}) = \log(\sum_k w_{m,k}(1-h_{k,n})) \tag{11}$$

$$= \log\left(\sum_k\tilde{\mu}_{m,n,k}\frac{w_{m,k}(1-h_{k,n})}{\tilde{\mu}_{m,n,k}}\right) \tag{12}$$

where we have introduced the nonnegative weights $\tilde{\mu}_{m,n,k}=w_{m,k}(1-\tilde{h}_{k,n})/(1-\tilde{y}_{m,n})$, which also sum up to 1. Using again Jensen inequality, we obtain an upper bound for this second term, i.e., $f_2(\mathbf{H})\leq\psi_2(\mathbf{H},\tilde{\mathbf{H}})$ with:

$$\psi_2(\mathbf{H},\tilde{\mathbf{H}}) = -\sum_{m,n,k}\frac{(1-y_{m,n})w_{m,k}(1-\tilde{h}_{k,n})}{1-\tilde{y}_{m,n}}\times\log\left(\frac{(1-h_{k,n})(1-\tilde{y}_{m,n})}{1-\tilde{h}_{k,n}}\right). \tag{13}$$

```
Algorithm 1: NBMF-MM
Inputs: Data matrix $\mathbf{Y} \in \{0,1\}^{M \times N}$, prior parameters $\alpha \geq 1$ and $\beta \geq 1$
Initialize $\mathbf{W}$ and $\mathbf{H}$ such that they comply with (3).
while convergence not reached do
    $\mathbf{C} = \mathbf{H} \odot \left( \mathbf{W}^\mathsf{T} \frac{\mathbf{Y}}{\mathbf{WH}} \right) + \alpha - 1$
    $\mathbf{D} = (1 - \mathbf{H}) \odot \left( \mathbf{W}^\mathsf{T} \frac{1 - \mathbf{Y}}{1 - \mathbf{WH}} \right) + \beta - 1$
    $\mathbf{H} = \frac{\mathbf{C}}{\mathbf{C} + \mathbf{D}}$
    $\mathbf{W} = \mathbf{W} \odot \left( \frac{\mathbf{Y}}{\mathbf{WH}} \mathbf{H}^\mathsf{T} + \frac{1 - \mathbf{Y}}{1 - \mathbf{WH}} (1 - \mathbf{H})^\mathsf{T} \right) / N$
end
Outputs: $\mathbf{W}$, $\mathbf{H}$
```

Combining (10) and (13) leads to $f \leq \psi_1 + \psi_2$, which yields an upper bound of the Lagrangian $\mathcal{L}$ given by (7). Note that the upper bound is tight, since it is straightforward to prove that equality holds when $\mathbf{H} = \tilde{\mathbf{H}}$. Minimizing the upper-bound (which is separable, smooth and convex) results in

$$h_{k,n} = \frac{\tilde{c}_{k,n}}{\tilde{c}_{k,n} + \tilde{d}_{k,n}}, \tag{14}$$

where

$$\tilde{c}_{k,n} = \tilde{h}_{k,n} \sum_m \frac{y_{m,n} w_{m,k}}{\tilde{y}_{m,n}} + \alpha_k - 1, \tag{15}$$

$$\tilde{d}_{k,n} = (1 - \tilde{h}_{k,n}) \sum_m \frac{(1 - y_{m,n}) w_{m,k}}{1 - \tilde{y}_{m,n}} + \beta_k - 1. \tag{16}$$

3.2.2 Update on W

An upper bound of $\mathcal{L}$ w.r.t to $\mathbf{W}$ for fixed $\mathbf{H}$ and $\boldsymbol{\lambda}$ can be obtained using the same tricks as above. Canceling the gradient of the upper bound now leads to:

$$-\frac{\tilde{w}_{m,k}}{w_{m,k}} \left(\sum_n \frac{y_{m,n} h_{k,n}}{\tilde{y}_{m,n}} + \frac{(1 - y_{m,n})(1 - h_{k,n})}{1 - \tilde{y}_{m,n}} \right) + \lambda_m = 0. \tag{17}$$

To determine λ_m, we multiply (17) by $w_{m,k}$ and sum over k to exploit the constraint (3). This yields:

$$\lambda_m = \sum_n \frac{y_{m,n}}{\tilde{y}_{m,n}} \sum_k \tilde{w}_{m,k} h_{k,n} + \frac{1 - y_{m,n}}{1 - \tilde{y}_{m,n}} \sum_k \tilde{w}_{m,k} (1 - h_{k,n}). \tag{18}$$

Since $\sum_k \tilde{w}_{m,k}(1 - h_{k,n}) = 1 - \tilde{y}_{m,n}$ and $\sum_k \tilde{w}_{k,m} h_{k,n} = \tilde{y}_{m,n}$, the expression of λ_m simplifies to:

$$\lambda_m = \sum_n y_{m,n} + (1 - y_{m,n}) = N. \tag{19}$$

Finally, combining (17) and (19) yields the following update:

$$w_{m,k} = \tilde{w}_{m,k} \left(\sum_n \frac{y_{m,n} h_{k,n}}{\tilde{y}_{m,n}} + \frac{(1 - y_{m,n})(1 - h_{k,n})}{1 - \tilde{y}_{m,n}} \right) / N. \tag{20}$$

3.3 Algorithm

Alternating (14) and (20) leads to the iterative procedure that we name NBMF-MM. It is summarized in Algorithm 1, where the updates are written into matrix form. The operations $.^\mathsf{T}$, $\odot$, and $\frac{\cdot}{\cdot}$ denote matrix transpose, element-wise multiplication, and division, respectively. Note that Algorithm 1 uses constant hyperparameter values $\alpha_k = \alpha$ and $\beta_k = \beta$, which led to satisfactory performance in preliminary experiments.

We remark that if $\mathbf{W}$ and $\mathbf{H}$ are initialized with nonnegative entries that respect the constraints (3), then the proposed updates guarantee that these constraints hold through iterations. Indeed, since initially $0 \leq \mathbf{H} \leq 1$ and $0 \leq \mathbf{WH} \leq 1$, then $1 - \mathbf{H} \geq 0$ and $1 - \mathbf{WH} \geq 0$. Besides, since $\mathbf{Y}$ is binary, then $1 - \mathbf{Y} \geq 0$. Therefore, all the terms involved in the updates are nonnegative, and consequently $\mathbf{W}$ and $\mathbf{H}$ remain nonnegative. Moreover, since the update on $\mathbf{H}$ is of the form $\mathbf{C}/(\mathbf{C}+\mathbf{D})$ with $\mathbf{C}$ and $\mathbf{D}$ nonnegative (as long as $\alpha \geq 1$ and $\beta \geq 1$), then $\mathbf{H} \leq 1$. Thus, the constraints are preserved.

Let us point out that other sets of constraints can ensure a valid Bernoulli parameter [12]. Indeed, one can switch the role of $\mathbf{W}$ and $\mathbf{H}$ (and the constraints/priors accordingly), which results in switching the corresponding updates in Algorithm 1. Alternatively, it is possible to set $\sum w_{m,k} = 1$ and $\sum_k h_{k,n} = 1$, in which case the update on $\mathbf{H}$ becomes similar to that of $\mathbf{W}$. In this work we only consider (3) for brevity.

Finally, let us outline that if the Beta prior reduces to a uniform prior (i.e., setting $\alpha = \beta = 1$), then the procedure is equivalent to the NBMF-EM algorithm [1]. However, NBMF-MM is obtained in a more straightforward fashion thanks to the MM strategy, does not require to introduce latent variables in an augmented model, and allows to tune the prior parameters.

4 Experiments

In this section, we assess the potential of NBMF-MM for decomposing and predicting binary data in a matrix completion task. Our code is available online for reproducibility.[3]

4.1 Protocol

4.1.1 Datasets

We consider three public binary datasets:

- `animals` [27]: An entry $y_{m,n} = 1$ indicates that the animal m has the attribute n ($M = 50$, $N = 85$).
- `paleo` [1]: An entry $y_{m,n} = 1$ indicates that the gene m has been found at location n ($M = 253$, $N = 902$).
- `lastfm` [28]: An entry $y_{m,n} = 1$ indicates that the user m has listened to the artist n ($M = 1{,}226$, $N = 285$).

Each dataset is split into a training, a validation, and a test subset, containing 70%, 15% and 15% of the data, respectively. The factors are learned on the training subset, and the hyperparameters (rank of the factorization K and prior parameters α and β) are tuned to minimize perplexity (see below) on the validation subset. Finally, the trained model is used to predict the test data in a binary matrix completion task.

4.1.2 Methods

The proposed NBMF-MM is compared against two baselines: NBMF-EM [1], which is equivalent to Algorithm 1 with $\alpha = \beta = 1$; and logistic PCA (logPCA) [11], which is a state-of-the-art method for predicting binary data.[4] The Bayesian methods from [4] and [12] are also relevant baselines, but they have been shown to perform similarly to NBMF-EM on these datasets [12]. Thus, for brevity we do not report these. All methods use the same convergence criterion: the algorithm is stopped when the relative variation of the objective function is lower than 10^{-5} or when a maximum number of 2000 iterations is reached.

4.1.3 Evaluation

Predictions are computed through $\hat{\mathbf{Y}} = \mathbf{W}\mathbf{H}$ for NBMF models, and $\hat{\mathbf{Y}} = \sigma(\mathbf{W}\mathbf{H})$ for logPCA. Prediction performance is then measured using the *perplexity* [29] (lower is better), which is defined as:

$$\text{perplexity} = -\frac{1}{|\vartheta|} \sum_{(m,n)\in\vartheta} \log p(y_{m,n}|\hat{y}_{m,n}) \tag{21}$$

where ϑ denotes the evaluation (validation or test) set, and $|\vartheta|$ denotes the number of elements in ϑ.

4.2 Results

First, we investigate the impact of the Beta prior on the performance on NBMF-MM. We display the perplexity on the validation set in Fig. 1. Overall, while increasing α improves performance (up to a point which depends on the dataset), a different trend is observed regarding β. For instance, performance becomes worse than with prior-free NBMF for large values of β and $\alpha = 1$ on the `animals` dataset. On the other hand, performance on the `lastfm` dataset is less sensitive to these variations, provided that $\alpha > 1$. We select the optimal hyperparameters and display the results on the test set in Fig. 2. NBMF-MM outperforms NBMF-EM by a large magin, which demonstrates the effectiveness of adjusting the Beta prior. This also shows that similar methods, e.g., [12], which have considered such a prior in the model formulation but did not test it experimentally, could actually benefit from this finding in order to fully reveal their potential. NBMF-MM outperforms logPCA on the `animals` dataset, but logPCA yields the best performance on `paleo` and `lastfm`. However, this result can be tempered

[3] The code will be made available when the paper is published.

[4] We use the package available at `https://github.com/andland/logisticPCA`

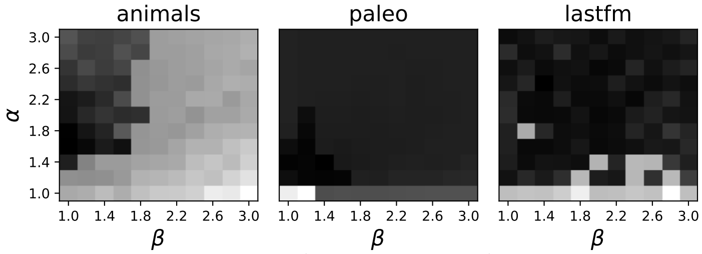

Figure 1: Validation perplexity (darker is better) for the optimal rank K.

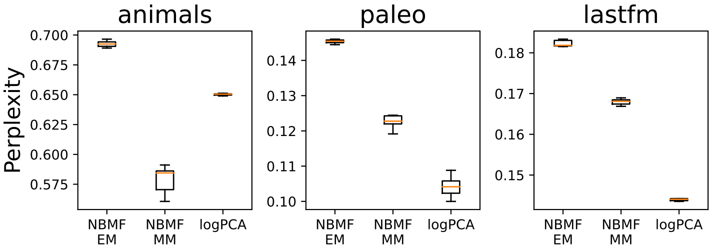

Figure 2: Perplexity of the test set for 10 random initializations. Each box-plot is made up of a central line indicating the median, box edges indicating the 1st and 3rd quartiles, and whiskers indicating the extremal values.

by the following considerations. Firstly, NBMF-MM is roughly 10 times faster than logPCA. Secondly, logPCA relies on using a link function, which hampers its interpretability.

To illustrate this last point, we plot in Fig. 3 the matrices $\mathbf{H}$ obtained on `lastfm`. We observe that much more distinct clusters can be extracted from the NBMF-based factor, which allows to grasp a high-level meaning of the K components. On this example, we can indeed interpret these components as related to the musical genre, e.g., “pop”, “electronic”, “rock” and “punk/metal”. This property is an asset in scenarios where interpretability of the factors is required.

5 Conclusion

We have proposed a novel MF algorithm for decomposing binary data. This method builds upon a mean-parametrized Bernoulli generative model along with a Beta prior. The parameters are estimated with MM, which yields simple and efficient updates. Our method offers an excellent trade-off between prediction performance, computational complexity, and interpretability, compared to state-of-the-art logistic PCA.

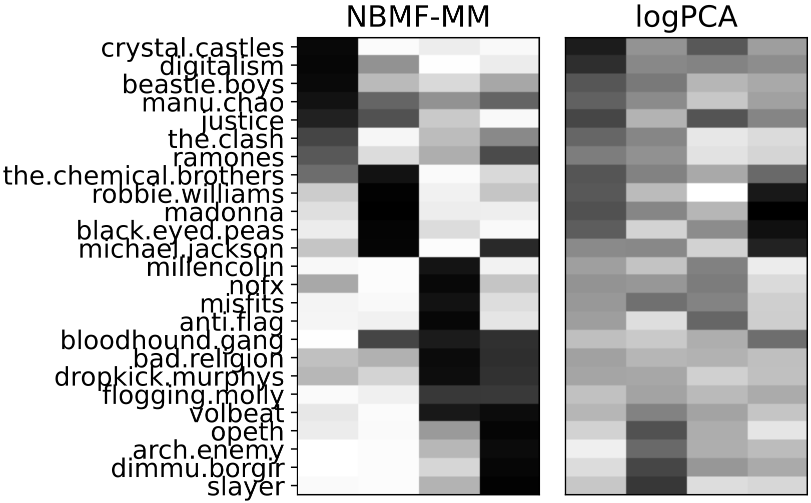

Figure 3: Estimated **H** matrix (transposed) from the `lastfm` dataset.

References

[1] E. Bingham, A. Kabán, and M. Fortelius, "The aspect Bernoulli model: Multiple causes of presences and absences," *Pattern Analysis & Applications*, vol. 12, no. 1, p. 55–78, January 2009.

[2] E. Voeten, "Data and analyses of voting in the UN general assembly," *Reinal B (ed) Routledge Handbook of International Organization*, p. 55–78, July 2013.

[3] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, "Music recommender systems," in *Recommender Systems Handbook.* Springer, 2015, pp. 453–492.

[4] A. Kabán and E. Bingham, "Factorisation and denoising of 0-1 data: A variational approach," *Neurocomputing*, vol. 71, no. 10–12, p. 2291–2308, June 2008.

[5] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, p. 30–37, August 2009.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*, December 2007, p. 1257–1264.

[8] P. K. Gopalan, L. Charlin, and D. Blei, "Content-based recommendations with Poisson factorization," in *Proc. of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*, December 2014, p. 3176–3184.

[9] P. K. Gopalan, J. M. Hofman, and D. Blei, "Scalable recommendation with hierarchical Poisson factorization," in *Proc. of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI'15)*, July 2014, p. 326–335.

[10] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. of the IEEE International Conference on Data Mining (ICDM '08)*, December 2008, pp. 263–272.

[11] A. J. Landgraf and Y. Lee, "Dimensionality reduction for binary data through the projection of natural parameters," *Journal of Multivariate Analysis*, vol. 180, p. 18, November 2020.

[12] A. Lumbreras, L. Filstroff, and C. Févotte, "Bayesian mean-parameterized nonnegative binary matrix factorization," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, p. 1898–1935, November 2020.

[13] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, February 2004.

[14] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, February 2017.

[15] M. Zhou, "Infinite edge partition models for overlapping community detection and link prediction," in *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, May 2015, pp. 1135–1143.

[16] M. E. Tipping, "Probabilistic visualisation of high-dimensional binary data," in *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*, December 1998, p. 592–598.

[17] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal component analysis to the exponential family," in *Proc. of the International Conference on Neural Information Processing Systems: Natural and Synthetic*, January 2001, p. 617–624.

[18] A. I. Schein, L. K. Saul, and L. H. Ungar, "A generalized linear model for principal component analysis of binary data," in *Proc. of the Ninth International Workshop on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, January 2003, pp. 240–247.

[19] A. M. Tomé, R. Schachtner, V. Vigneron, C. G. Puntonet, and E. W. Lang, "A logistic non-negative matrix factorization approach to binary data sets," *Multidimensional Systems and Signal Processing*, vol. 26, no. 1, p. 125–143, January 2015.

[20] J. S. Larsen and L. K. H. Clemmensen, "Non-negative matrix factorization for binary data," in *Proc. of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, vol. 01, November 2015, pp. 555–563.

[21] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[22] F. Ortega, R. Lara-Cabrera, A. González-Prieto, and J. Bobadilla, "Providing reliability in recommender systems through bernoulli matrix factorization," *Information Sciences*, vol. 553, pp. 110–128, 2021.

[23] C. A. Hidalgo, "Economic complexity theory and applications," *Nature Reviews Physics*, vol. 3, no. 2, p. 92–113, January 2021.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, September 2011.

[25] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2017, pp. 259–263.

[26] Z. Lin, C. Xu, and H. Zha, "Robust matrix factorization by majorization minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 208–220, January 2018.

[27] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proc. of the 21st National Conference on Artificial Intelligence - Volume 1*, July 2006, p. 381–388.

[28] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. of the 12th International Conference on Music Information Retrieval (ISMIR)*, October 2011.

[29] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of the 15th Conference on Uncertainty in Artificial Intelligence (UAI '99)*, July 1999, p. 50–57.