Negative Binomial Matrix Factorization

Olivier Gouvert, Thomas Oberlin, Cédric Févotte

▶ **To cite this version:**

Olivier Gouvert, Thomas Oberlin, Cédric Févotte. Negative Binomial Matrix Factorization. IEEE Signal Processing Letters, 2020, 27, ⟨10.1109/LSP.2020.2991613⟩. ⟨hal-02871905⟩

HAL Id: hal-02871905

https://hal.science/hal-02871905v1

Submitted on 17 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Negative Binomial Matrix Factorization

Olivier Gouvert, Thomas Oberlin, *Member, IEEE*, and Cédric Févotte, *Senior Member, IEEE*

***Abstract*—We introduce negative binomial matrix factorization (NBMF), a matrix factorization technique specially designed for analyzing over-dispersed count data. It can be viewed as an extension of Poisson factorization (PF) perturbed by a multiplicative term which models exposure. This term brings a degree of freedom for controlling the dispersion, making NBMF more robust to outliers. We describe a majorization-minimization (MM) algorithm for a maximum likelihood estimation of the parameters. We provide results on a recommendation task and demonstrate the ability of NBMF to efficiently exploit raw data.**

***Index Terms*—Non-negative matrix factorization, Poisson factorization, majorization-minimization, over-dispersion, collaborative filtering**

I. Introduction

Poisson factorization (PF) is a special case of non-negative matrix factorization (NMF) with applications in dictionary learning for signal & image processing [1], [2], [3], text information retrieval [4], [5] or recommender systems [6], [7]. In this setting, the data is assumed to be drawn from the Poisson distribution making it specially well suited for count/integer-valued data. More precisely, each entry y_{ui} of the data matrix $\mathbf{Y}$ is generated from the process:

$$y_{ui} \sim \text{Poisson}([\mathbf{W}\mathbf{H}^T]_{ui}), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}_+^{U\times K}$ is the dictionary matrix and $\mathbf{H} \in \mathbb{R}_+^{I\times K}$ is the activation matrix. Usually, $K \ll \min(U, I)$ which implies a low-rank data approximation. A limitation of using the Poisson distribution is that the variance is fixed and equal to the mean: $\text{var}(y_{ui}) = \mathbb{E}(y_{ui})$, making it poorly adapted for over-dispersed data.

As such, we propose in this letter a new probabilistic matrix factorization (MF) model, coined negative binomial matrix factorization (NBMF), which is especially designed for over-dispersed count data. In particular, NBMF offers an additional degree of freedom for controlling data dispersion, which is beneficial when analyzing raw data.

In particular, we illustrate this ability by applying NBMF to collaborative filtering (CF). CF is a common application of MF methods, which aims at analyzing user preferences in order to make recommendations. Since the Netflix Prize [8], CF has been giving state-of-the-art results for recommender systems by exploiting user historical data. These data can either be explicit (ratings given by users to items) or implicit (count data from users listening to songs, clicking on web pages, watching videos, etc). Count data can be summarized into a matrix $\mathbf{Y} \in \mathbb{N}^{U\times I}$, where y_{ui} corresponds to the number of times a user u interacts with an item i, U and I are the number of users and items respectively, and $\mathbb{N}$ is the set of non-negative integer values. This type of data, easy to collect, is known to be very sparse, noisy and bursty [9], [10]. Therefore, to remain robust to outliers, a pre-processing stage is often carried out before applying PF [7], [11]: all positive values are thresholded to 1, producing binary data, i.e., $\mathbf{Y} \in \{0, 1\}^{U\times I}$. On the contrary, we propose to process the raw data with NBMF, avoiding the loss of information induced by any pre-processing stage.

This work was supported by the European Research Council (ERC FACTORY-CoG-6681839) and the ANR-3IA (ANITI).

O. Gouvert and C. Févotte are with the IRIT, Université de Toulouse, CNRS, France (e-mail: firstname.lastname@irit.fr). T. Oberlin is with ISAE-SUPAERO, Université de Toulouse, France (e-mail: firstname.lastname@isae.fr).

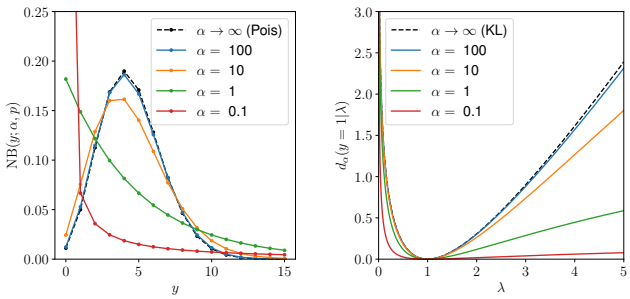

Fig. 1. Left: Probability mass function of the NB distribution: $y \sim \text{NB}(\alpha, p)$, such that $\mathbb{E}(y) = 4.5$. Right: The associated divergence for $y = 1$.

This paper is organized as follows. In Section II, we introduce NBMF. In Section III, we discuss its connections with the state of the art. In Section IV, we study the maximum likelihood estimator of NBMF and discuss the fit function/divergence it implies. Finally, in Section V, we illustrate the benefits of NBMF with experiments on a real dataset.

II. Negative Binomial Matrix Factorization

In this section, we introduce NBMF, which is a NMF method especially designed for over-dispersed count data. We chose to illustrate the concepts and properties of NBMF for data generated by users interacting with items. However, NBMF can be applied to a wider range of data.

A. Model

We assume that, for each user $u \in \{1, \dots, U\}$ and item $i \in \{1, \dots, I\}$, our observations $y_{ui} = [\mathbf{Y}]_{ui}$ are sampled from the generative process

$$y_{ui} \sim \text{NB}\left(\alpha, \frac{[\mathbf{W}\mathbf{H}^T]_{ui}}{[\mathbf{W}\mathbf{H}^T]_{ui} + \alpha}\right), \tag{2}$$

where $\mathbf{W}$ represents the preferences of users and $\mathbf{H}$ represents the attributes of items [12]. $\text{NB}(\alpha, p)$ is the negative binomial (NB) distribution parametrized by a dispersion coefficient $\alpha \in$

$\mathbb{R}_+$ and a probability parameter $p \in [0, 1]$. Its probability mass function, displayed in Figure 1, is given by:

$$\mathbb{P}(Y = y) = \frac{\Gamma(y + \alpha)}{y!\,\Gamma(\alpha)} p^y (1 - p)^\alpha, \tag{3}$$

where $\Gamma(.)$ is the gamma function. When $\alpha \leq 0$, the mode of the NB distribution is located in 0. When α tends to infinity and the average is fixed, we recover the Poisson distribution.

Like in PF and many mean-parametrized matrix factorization models [13], the expected value of the observations is given by: $\mathbb{E}(y_{ui}) = [\mathbf{W}\mathbf{H}^T]_{ui}$, which gives an intuitive understanding of the model. Contrary to the Poisson distribution, the NB distribution has a second parameter α which enables to add variance to the model:

$$\mathrm{var}(y_{ui}) = [\mathbf{W}\mathbf{H}^T]_{ui} \left(1 + \frac{[\mathbf{W}\mathbf{H}^T]_{ui}}{\alpha}\right) > \mathbb{E}(y_{ui}). \tag{4}$$

It is important to note that, contrary to the method of the same name introduced in [14], we place the factorization $[\mathbf{W}\mathbf{H}^T]_{ui}$ on the probability parameter of the NB distribution and not on the shape parameter.

The NB distribution can also be viewed as a Poisson-gamma mixture. Note that this formulation is different from [3] where gamma priors are put on $\mathbf{W}$ and/or $\mathbf{H}$. Using this property, we can write the following equivalent hierarchical model:

$$a_{ui} \sim \mathrm{Gamma}(\alpha, \alpha), \tag{5}$$

$$y_{ui} | a_{ui} \sim \mathrm{Poisson}(a_{ui}[\mathbf{W}\mathbf{H}^T]_{ui}), \tag{6}$$

where the latent variables a_{ui} control local variabilities. We denote by $\mathbf{A}$ the $U \times I$ matrix with coefficients $[\mathbf{A}]_{ui} = a_{ui}$. By construction, we have $\mathbb{E}(a_{ui}) = 1$ and $\mathrm{var}(a_{ui}) = \alpha^{-1}$.

B. Interpretation of the Latent Variable $\mathbf{A}$

The matrix $\mathbf{A}$ captures local variations that cannot be explained by the product $\mathbf{W}\mathbf{H}^T$, by attenuating or accentuating them. In the field of recommender systems, $\mathbf{A}$ can be viewed as an exposure variable [11] which models how much a user is exposed to an item. For example, in the context of song recommendation, we have the following interpretations:

• If $a_{ui} \ll 1$, the user is under-exposed to the item. It may be explained by several reasons: the user does not frequent the places/communities where the song is played, he is not aware of the release of a new song, etc.

• If $a_{ui} \gg 1$, the user is over-exposed to the item. This over-exposure can be "active", e.g., the user listens to the song on repeat, or "passive", e.g., the item is heavily broadcasted on the radio, is highlighted on a website, etc.

• If $a_{ui} \approx 1$, the exposure does not affect the listening pattern of the user which is fully described by $\mathbf{W}\mathbf{H}^T$.

III. Related Works

A. Negative Binomial Regression

Regression for count data based on the Poisson distribution has been considered by [15]. It has been augmented by a latent variable a to model over-dispersion in [16], [17], [18]:

$$y_i \sim \mathrm{Poisson}(a_i \exp(\mathbf{x}_i^T \mathbf{b})), \tag{7}$$

where $y_i \in \mathbb{N}$ is the response variable, $\mathbf{x}_i$ is the covariate vector for sample i and $\mathbf{b}$ is the vector of regression coefficients. When a_i is given a gamma prior and marginalized, we get NB regression:

$$y_i \sim \mathrm{NB}\left(\alpha, (1 + \alpha \exp(-\mathbf{x}_i^T \mathbf{b}))^{-1}\right). \tag{8}$$

Equation (8) defines a generalized linear model [19] in which the data expectation is not linear in the parameters. We work instead with the mean-parametrized form of Equation (2), which is more natural to the MF and dictionary learning settings. Furthermore, we also learn the "covariates" (similar to $\mathbf{W}$ in our case) and assume all variables to be non-negative.

B. Robust NMF

The latent variable $\mathbf{A}$ can be interpreted as a variable that accounts for outliers. Indeed, $\mathbf{A}$ is a multiplicative perturbation which can explain unexpectedly high or low values (see Section II-B). In [20], the authors proposed a different way for handling outliers in NMF models and in particular in Poisson factorization (in the context of hyperspectral image unmixing). The outliers are modeled with an additive latent variable. The data is assumed Poisson-distributed with expectation $[\mathbf{W}\mathbf{H}^T]_{ui} + s_{ui}$ where s_{ui} is imposed to be sparse and non-negative. The non-negativity implies that only unexpectedly large data values can be captured with such a model.

C. Exposure and Poisson Modeling

NBMF can be cast as a particular instance of the following general model:

$$\mathbf{A} \sim p(\mathbf{A}; \Theta), \tag{9}$$

$$y_{ui} | a_{ui} \sim \mathrm{Poisson}(a_{ui}[\mathbf{W}\mathbf{H}^T]_{ui}), \tag{10}$$

where $p(\mathbf{A}; \Theta)$ is a distribution governed by its own parameters Θ.

There are a few examples of such models in the literature, as described next.

• When $\mathbf{A}$ is deterministic with $\forall (u, i), a_{ui} = 1$, we recover the well-known PF model [3], [4], [5], [6], [7].

• Zero-inflated models. In [21], a_{ui} is drawn from a Bernoulli distribution: $a_{ui} \sim \mathrm{Bern}(\mu)$. Marginalizing out this latent variable leads to the zero-inflated Poisson distribution [22]:

$$y_{ui} \sim (1 - \mu)\delta_0 + \mu\, \mathrm{Poisson}([\mathbf{W}\mathbf{H}^T]_{ui}). \tag{11}$$

In practice, it appears that the Bernoulli distribution puts too much weight on 0. The gamma distribution offers a softer alternative. [21] also proposes more sophisticated hierarchical models for μ (which becomes μ_{ui}) to include external sources of knowledge (social network or geographical informations). Such ideas could also be incorporated in our setting. Note that a zero-inflated NB model has also been introduced in [23], but the parametrization of this model differs from Eq. (11).

• Coupled compound PF. In [24] the authors consider matrix completion with PF and missing-not-at-random phenomenas. Their approach relies on the following assumption

$$a_{ui} \sim \mathrm{Poisson}([\mathbf{U}\mathbf{V}^T]_{ui}), \tag{12}$$

which is more restrictive in terms of support and structure than our proposal. The general purpose is also different.

• Random graphs. In reference [25], the exposure is modeled with bipartite random graphs and it is arbitrarily assumed that half of the unconsumed items are missing feedbacks.

D. Exposure and Gaussian modeling

Besides the models with a Poisson likelihood, the notion of exposure was also introduced in the context of Gaussian modeling.

• Exposure matrix factorization. In [11], the authors develop the so-called exposure matrix factorization (ExpoMF). This model posits a Gaussian distribution for the binary observations with factorized matrix expectation. A binary variable modeling exposure is introduced. It models whether a user knows an item or not:

$$a_{ui} \sim \mathrm{Bern}(\mu), \tag{13}$$

$$y_{ui}|a_{ui} \sim \mathcal{N}(0, a_{ui}[\mathbf{W}\mathbf{H}^T]_{ui}). \tag{14}$$

The authors of this paper emphasize the fact that weighted MF (WMF) [9] applied to binary data is a special case of ExpoMF. They developed an EM algorithm to infer the parameters of the model. A similar model has been introduced in [26], in the context of gene expression analysis. Contrary to those works, we choose to work with the Poisson distribution which is better adapted to count data [4]. In addition, we do not apply preprocessing to the data.

• Semi-blind source separation. In semi-blind source separation, a similar model has been developed to allow for more flexibility in the modeling [27], [28]. It assumes that the time-frequency coefficients y_{ui} of each source follow a Student's t distribution. The parameters of the distribution are structured by an NMF model. This model is equivalent to the hierarchical model:

$$a_{ui} \sim \mathrm{IG}(\alpha/2, \alpha/2) \tag{15}$$

$$y_{ui}|a_{ui} \sim \mathcal{N}(0, a_{ui}[\mathbf{W}\mathbf{H}^T]_{ui}). \tag{16}$$

The latent variable $\mathbf{A}$ has a role similar to our exposure variable, allowing to obtain a marginal distribution with a heavier tail.

IV. Maximum Likelihood Estimation

In this section, we study maximum likelihood (ML) estimation in the proposed model (2) and discuss the data fitting term that arises from our model.

A. A New Divergence

The ML estimator of $\mathbf{W}$ and $\mathbf{H}$ is obtained by minimizing the objective function defined by:

$$C_{\mathrm{ML}}(\mathbf{W}, \mathbf{H}) = -\log p(\mathbf{Y}; \mathbf{W}, \mathbf{H}) \tag{17}$$

$$= \sum_{ui} d_\alpha(y_{ui}|[\mathbf{W}\mathbf{H}^T]_{ui}) + cst, \tag{18}$$

where cst is a constant with respect to (w.r.t.) $\mathbf{W}$ and $\mathbf{H}$ and

$$d_\alpha(a|b) = a\log\left(\frac{a}{b}\right) - (\alpha + a)\log\left(\frac{\alpha + a}{\alpha + b}\right). \tag{19}$$

We exhibit a new divergence, denoted by d_α, which is associated to the mean-parametrized NB distribution with fixed dispersion coefficient α. It is displayed in Figure 1 for various values of α. To the best of our knowledge, this divergence does not have a name nor corresponds to a well-known case from the literature. As expected, we recover in the limit case the generalized Kullback-Leibler divergence associated with the Poisson distribution:

$$\lim_{\alpha\to\infty} d_\alpha(a|b) = a\log\left(\frac{a}{b}\right) - a + b = \mathrm{KL}(a|b). \tag{20}$$

B. Block-Descent Majorization-Minimization

As it turns out, maximum likelihood reduces to minimization of

$$C(\mathbf{W}, \mathbf{H}) = D_\alpha(\mathbf{Y}|\mathbf{W}\mathbf{H}^T) \tag{21}$$

where $D_\alpha(\cdot|\cdot)$ is the entry-wise matrix divergence induced by $d_\alpha(\cdot|\cdot)$. Equation (21) defines a new NMF problem. A standard approach to minimize $C(\mathbf{W}, \mathbf{H})$ is alternate block-descent optimization in which $\mathbf{W}$ and $\mathbf{H}$ are updated in turn until convergence of the objective function. The returned solution may only be a local one owing to the non-convexity of $C(\mathbf{W}, \mathbf{H})$. The individual updates for $\mathbf{W}$ and $\mathbf{H}$ can be obtained using majorization-minimization (MM) like in many NMF cases, and such as NMF with the β-divergence [29]. The roles of $\mathbf{W}$ and $\mathbf{H}$ can be exchanged by transposition ($\mathbf{Y} \approx \mathbf{W}\mathbf{H}^T$ is equivalent to $\mathbf{Y}^T \approx \mathbf{H}\mathbf{W}^T$) and we may for example address the update of $\mathbf{H}$ given $\mathbf{W}$. MM amounts to optimizing an upper bound $G(\mathbf{H}|\bar{\mathbf{H}})$ of $C(\mathbf{W}, \mathbf{H})$, constructed so as to be tight at the current iterate $\bar{\mathbf{H}}$ ($G(\bar{\mathbf{H}}|\bar{\mathbf{H}}) = C(\mathbf{W}, \bar{\mathbf{H}})$). This produces a descent algorithm where the objective function is decreased at every iteration [30].

In our setting, a tight upper bound can be constructed by majorizing the convex and concave parts of $C(\mathbf{W}, \mathbf{H})$, following the approach proposed in [29] for NMF with the β-divergence. The convex part (terms in $-\log(x)$) may be majorized using Jensen's inequality. The concave part (terms in $\log(x + c)$) can be majorized using the tangent inequality. This procedure leads to the following multiplicative update that preserves non-negativity given positive initializations:

$$h_{ik} = \bar{h}_{ik}\, \frac{\sum_u \frac{y_{ui}}{[\mathbf{W}\bar{\mathbf{H}}^T]_{ui}} w_{uk}}{\sum_u \frac{y_{ui} + \alpha}{[\mathbf{W}\bar{\mathbf{H}}^T]_{ui} + \alpha} w_{uk}}. \tag{22}$$

Similarly, the update for $\mathbf{W}$ is given by:

$$w_{uk} = \bar{w}_{uk}\, \frac{\sum_i \frac{y_{ui}}{[\bar{\mathbf{W}}\mathbf{H}^T]_{ui}} h_{ik}}{\sum_i \frac{y_{ui} + \alpha}{[\bar{\mathbf{W}}\mathbf{H}^T]_{ui} + \alpha} h_{ik}}. \tag{23}$$

As expected, the multiplicative updates of KL-NMF [31] are obtained in the limit $\alpha \to \infty$. This algorithm scales with the number of entries UI in the matrix $\mathbf{Y}$, and not with the number of non-zero entries like for KL-NMF algorithm. This increase of computational complexity can be a drawback in applications where the matrix $\mathbf{Y}$ is large but sparse.

TABLE I
RECOMMENDATION PERFORMANCE OF THE THREE COMPARED ALGORITHMS ON THE TASTE PROFILE DATASET. IN BOLD, THE BEST NDCG SCORES.

	NDCG@20					
Model	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 6
NBMF ($\alpha = 1$)	0.223 ($\pm$ 0.005)	0.234 ($\pm$ 0.006)	0.242 ($\pm$ 0.006)	0.243 ($\pm$ 0.009)	0.239 ($\pm$ 0.008)	0.246 ($\pm$ 0.009)
NBMF ($\alpha = 10$)	0.224 ($\pm$ 0.005)	0.237 ($\pm$ 0.005)	**0.247 ($\pm$ 0.006)**	**0.250 ($\pm$ 0.008)**	**0.247 ($\pm$ 0.008)**	**0.254 ($\pm$ 0.009)**
NBMF ($\alpha = 100$)	0.222 ($\pm$ 0.005)	0.235 ($\pm$ 0.005)	0.245 ($\pm$ 0.006)	0.248 ($\pm$ 0.008)	0.245 ($\pm$ 0.008)	0.252 ($\pm$ 0.009)
KL-NMF-raw	0.221 ($\pm$ 0.005)	0.235 ($\pm$ 0.005)	0.244 ($\pm$ 0.006)	0.247 ($\pm$ 0.008)	0.244 ($\pm$ 0.009)	0.250 ($\pm$ 0.009)
KL-NMF-bin	**0.265 ($\pm$ 0.006)**	**0.251 ($\pm$ 0.006)**	0.239 ($\pm$ 0.007)	0.226 ($\pm$ 0.009)	0.212 ($\pm$ 0.010)	0.213 ($\pm$ 0.009)

Bayesian inference of NBMF can also be considered, by introducing gamma priors on the latent variables $\mathbf{W}$ and $\mathbf{H}$ [3], [7]. In particular, coordinate ascent variational inference (CAVI) algorithm is detailed in the technical report [32].

V. EXPERIMENTAL RESULTS

A. Experimental Setup

a) Dataset: We apply our algorithm to the Taste Profile dataset provided by The Echo Nest [33]. This dataset contains the listening history of users in the form of song play counts. As in [11], we select a subset of the original data by only keeping users who listened to at least 20 different songs, and songs which have been listened to at least by 50 different users. This pre-processing ensures enough information for each user and item, while it avoids the cold-start problem inherent in CF. This leads to a dataset with a number of users $U = 1509$ and a number of items $I = 805$.

b) Recommendation Task: The goal of recommender system is to propose to each user a personalized list of new items (items he has not consumed yet) that he may like. To evaluate our algorithm, we randomly divide the observed matrix $\mathbf{Y}$ into two matrices $\mathbf{Y}^{\text{train}}$ and $\mathbf{Y}^{\text{test}}$. $\mathbf{Y}^{\text{train}}$ is composed of 80% of the non-zero values of $\mathbf{Y}$ (the other values are set to zero, i.e., are assumed not to have been listened to in the train set), while $\mathbf{Y}^{\text{test}}$ is composed of the remaining 20%. For each user, we propose a list of recommendations composed of m items never consumed in $\mathbf{Y}^{\text{train}}$, i.e., such that $y_{ui}^{\text{train}} = 0$. The list is constructed by decreasing order of the score: $s_{ui} = [\hat{\mathbf{W}}\hat{\mathbf{H}}^T]_{ui}$, where $\hat{\mathbf{W}}$ are the estimated user preferences and $\hat{\mathbf{H}}$ are the estimated item attributes.

c) Normalized discounted cumulative gain: We use the normalized discounted cumulative gain (NDCG) to measure the quality of these lists of recommendations [34]. For each user, we calculate the discounted cumulative gain (DCG), defined by:

$$\text{DCG}_u = \sum_{l=1}^{m} \frac{\text{rel}(u,l)}{\log_2(l+1)}, \tag{24}$$

where $\text{rel}(u,l)$ is the ground-truth relevance of the l-th item of the list of the user u. The denominator penalizes relevant items which are at the end of the proposed list. It accounts for the fact that a user will only browse the beginning of the list, and will not pay attention to items which are ranked at the end. The NDCG is a normalized version of the DCG:

$$\text{NDCG}_u = \frac{\text{DCG}_u}{\text{IDCG}_u} \in [0,1], \tag{25}$$

where IDCG_u is the DCG score obtained by an ideal list of recommendations.

In our experiments, we chose the following definition of the ground-truth relevance:

$$\text{rel}(u,l) = \mathbb{1}[y_{ui(l)} \geq s], \tag{26}$$

where $i(l)$ is the index of the l-th item of the list, and s is a fixed threshold. When $s = 1$, we recover the classic NDCG metric for binary data. When $s > 1$, we focus only on items which have been listened to at least s times. It totally ignores listening counts lower than s for which the confidence may not be high enough. Note that the introduction of the threshold s only affects evaluation and not training.

d) Compared methods: We compare our algorithm (called NBMF), described by the update rules in Eq. (22)-(23), with two versions of KL-NMF [31]. One with pre-processing stage where we binarize $\mathbf{Y}^{\text{train}}$, denoted by KL-NMF-bin; and one without where we work directly on the raw data (as for NBMF), denoted by KL-NMF-raw.

All the three algorithms are run on 10 random splittings of the dataset, with 10 random initializations. For NBMF and KL-NMF-raw, we initialize the algorithms with KL-NMF-bin. We found out that this was the best initialization strategy as compared to using random initialization or the result of KL-NMF-raw. We fix the number of latent factors to $K = 50$. The algorithms are stopped when the relative decrement of the cost function is less than 10^{-5}.

B. Recommendation Results

Table I displays the NDCG score obtained for each algorithm on the Taste Profile dataset. The size of the lists of recommendations is fixed to $m = 20$, and 5 different values of threshold are compared. First, we discuss the difference between KL-NMF algorithm performed on raw data or on binarized data (two last lines of the table). We can see that KL-NMF-bin achieves the best performances for small threshold ($s = 0$ or 1). This confirms the usefulness of the pre-processing stage up to a certain threshold. Then, we display on Table I, the results of NBMF algorithm for three different values of $\alpha \in \{1, 10, 100\}$. We remind that, when α goes to infinity, the NBMF algorithm is equivalent to KL-NMF-raw. The NBMF algorithm seems to reach peak performance for $\alpha = 10$. For this value, NBMF returns better results than KL-NMF-raw for all the values of threshold s. As for KL-NMF-raw, NBMF returns the best NDCG scores for $s \geq 2$. This confirms the loss of information caused by the pre-processing stage.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, 1999.

[2] P. Smaragdis, C. Févotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, 2014.

[3] A. T. Cemgil, "Bayesian Inference for Nonnegative Matrix Factorisation Models," *Computational Intelligence and Neuroscience*, 2009.

[4] J. Canny, "GaP: A Factor Model for Discrete Data," in *Proc. ACM International on Research and Development in Information Retrieval (SIGIR)*, 2004.

[5] W. Buntine and A. Jakulin, "Discrete component analysis," in *Subspace, Latent Structure and Feature Selection*. Springer, 2006.

[6] H. Ma, C. Liu, I. King, and M. R. Lyu, "Probabilistic Factor Models for Web Site Recommendation," in *Proc. ACM International on Research and Development in Information Retrieval (SIGIR)*, 2011.

[7] P. Gopalan, J. M. Hofman, and D. M. Blei, "Scalable Recommendation with Hierarchical Poisson Factorization." in *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

[8] J. Bennett, S. Lanning *et al.*, "The Netflix Prize," in *Proc. KDD Cup and Workshop*, 2007.

[9] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. IEEE International Conference on Data Mining (ICDM)*, 2008.

[10] A. Schein, H. Wallach, and M. Zhou, "Poisson-Gamma dynamical systems," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[11] D. Liang, L. Charlin, J. McInerney, and D. M. Blei, "Modeling User Exposure in Recommendation," in *Proc. International Conference on World Wide Web (WWW)*, 2016.

[12] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer*, 2009.

[13] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the beta-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[14] M. Zhou, "Nonparametric Bayesian negative binomial factor analysis," *Bayesian Analysis*, 2017.

[15] W. Gardner, E. P. Mulvey, and E. C. Shaw, "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." *Psychological bulletin*, 1995.

[16] J. F. Lawless, "Negative binomial and mixed Poisson regression," *Canadian Journal of Statistics*, 1987.

[17] J. M. Hilbe, *Negative Binomial Regression*. Cambridge University Press, 2011.

[18] M. Zhou, L. Li, D. Dunson, and L. Carin, "Lognormal and gamma mixed negative binomial regression," in *Proc. International Conference on Machine Learning (ICML)*, 2012.

[19] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1980.

[20] C. Févotte and N. Dobigeon, "Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization," *IEEE Transactions on Image Processing*, 2015.

[21] M. Simchowitz, "Zero-Inflated Poisson Factorization for Recommendation Systems," 2013.

[22] D. Lambert, "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 1992.

[23] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, "A general and flexible method for signal extraction from single-cell rna-seq data," *Nature communications*, 2018.

[24] M. E. Basbug and B. E. Engelhardt, "Coupled compound poisson factorization," *arXiv:1701.02058*, 2017.

[25] U. Paquet and N. Koenigstein, "One-class Collaborative Filtering with Random Graphs," in *Proc. International Conference on World Wide Web (WWW)*, 2013.

[26] E. Pierson and C. Yau, "Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome biology*, 2015.

[27] S. Leglaive, R. Badeau, and G. Richard, "Semi-blind student's t source separation for multichannel audio convolutive mixtures," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2017.

[28] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[29] C. Févotte and J. Idier, "Algorithms for Nonnegative Matrix Factorization with the beta-divergence," *Neural computation*, 2011.

[30] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, 2004.

[31] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," in *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[32] O. Gouvert, T. Oberlin, and C. Févotte, "Negative Binomial Matrix Factorization for Recommender Systems," *arXiv:1801.01708*, 2018.

[33] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proc. International Society for Music Information Retrieval (ISMIR)*, 2011.

[34] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, 2002.