ISSN 1066-5307, Mathematical Methods of Statistics, 2017, Vol. 26, No. 1, pp. 55–67. © Allerton Press, Inc., 2017.

An Oracle Inequality for Quasi-Bayesian Nonnegative Matrix Factorization

P. Alquier[1*] and B. Guedj[2]**

[1]*CREST, ENSAE, Univ. Paris Saclay, France*

[2]*Modal Project-Team, Inria Lille — Nord Europe Research Center, France*

Received June 7, 2016; in final form, August 24, 2016

Abstract—The aim of this paper is to provide some theoretical understanding of quasi-Bayesian aggregation methods of nonnegative matrix factorization. We derive an oracle inequality for an aggregated estimator. This result holds for a very general class of prior distributions and shows how the prior affects the rate of convergence.

Keywords: nonnegative matrix factorization, oracle inequality, PAC-Bayesian bounds.

2000 Mathematics Subject Classification: primary 62H99; secondary 62F35, 68T05, 65C05.

DOI: 10.3103/S1066530717010045

1. INTRODUCTION

Nonnegative matrix factorization (NMF) is a set of algorithms in high-dimensional data analysis which aims at factorizing a large matrix M with nonnegative entries. If M is an $m_1 \times m_2$ matrix, NMF consists in decomposing it as a product of two matrices of smaller dimensions: $M \simeq UV^T$, where U is $m_1 \times K$, V is $m_2 \times K$, $K \ll m_1 \wedge m_2$ and both U and V have nonnegative entries. Interpreting the columns $M_{\cdot,j}$ of M as (nonnegative) signals, NMF amounts to decompose (exactly, or approximately) each signal as a combination of the "elementary" signals $U_{\cdot,1}, \dots, U_{\cdot,K}$:

$$M_{\cdot,j} \simeq \sum_{\ell=1}^{K} V_{j,\ell} U_{\cdot,\ell}. \qquad (1)$$

Since the seminal paper [28], NMF was successfully applied to various fields such as image processing and face classification [23], separation of sources in audio and video processing [38], collaborative filtering and recommender systems on the Web [26], document clustering [46, 42], medical image processing [1] or topics extraction in texts [39]. In all these applications, it has been pointed out that NMF provides a decomposition which is usually interpretable. A theoretical foundation to this interpretatibility by exhibiting conditions under which the decomposition $M \simeq UV^T$ is unique was given in [16]. However, let us stress that even when this is not the case, the results provided by NMF are still sensibly interpreted by practitioners.

Since a prior knowledge on the shape and/or magnitude of the signal is available in many settings, Bayesian tools have extensively been used for (general) matrix factorization [13, 32, 40, 27, 50] and have been adapted for the Bayesian NMF problem ([37, 12, 17, 41, 45, 49] among others).

The aim of this paper is to provide some theoretical analysis of the performance of an aggregation method for NMF inspired by the aforementioned Bayesian works. We propose a quasi-Bayesian estimator for NMF. By quasi-Bayesian, we mean that the construction of the estimator relies on a prior distribution π, however, it does not rely on any parametric assumptions — that is, the likelihood used to

*E-mail: pierre.alquier@ensae.fr

**E-mail: benjamin.guedj@inria.fr

build the estimator does not have to be well-specified (it is usually referred to as a quasi-likelihood). The use of quasi-likelihoods in Bayesian estimation is advocated by [6] using decision-theoretic arguments. This methodology is also popular in machine learning, and various authors developed a theoretical framework to analyze it ([43, 36, 9, 10, 11], this is known as the PAC-Bayesian theory). It is also related to recent works on exponentially weighted aggregation in statistics [14, 19]. Using these theoretical tools, we derive an oracle inequality for our quasi-Bayesian estimator. The message of this theoretical bound is that our procedure is able to adapt to the unknown rank of M under very general assumptions on the noise.

The paper is organized as follows. Notation for the NMF framework and the definition of our quasi-Bayesian estimator are given in Section 2. The oracle inequality, which is our main contribution, is given in Section 3 and its proof is postponed to Section 5. The computation of our estimator being completely similar to the computation of a (proper) Bayesian estimator, we end the paper with a short discussion and references to state-of-the-art computational methods for Bayesian NMF in Section 4.

2. NOTATION

For any $p \times q$ matrix A we denote by $A_{i,j}$ its (i,j)th entry, by $A_{i,\cdot}$ its ith row and by $A_{\cdot,j}$ its jth column. For any $p \times q$ matrix B we define

$$\langle A, B\rangle_F = \mathrm{Tr}(AB^\top) = \sum_{i=1}^{p}\sum_{j=1}^{q} A_{i,j}B_{i,j}.$$

We define the Frobenius norm $\|A\|_F$ of A by $\|A\|_F^2 = \langle A, A\rangle_F$. Let $A_{-i,\cdot}$ denote the matrix A where the ith column is removed. In the same way, for a vector $v \in \mathbb{R}^p$, $v_{-i} \in \mathbb{R}^{p-1}$ is the vector v with its ith coordinate removed. Finally, let $\mathrm{Diag}(v)$ denote the $p \times p$ diagonal matrix given by $[\mathrm{Diag}(v)]_{i,i} = v_i$.

2.1. Model

The object of interest is an $m_1 \times m_2$ target matrix M possibly polluted with some noise $\mathcal{E}$. So we actually observe

$$Y = M + \mathcal{E}, \tag{2}$$

and we assume that $\mathcal{E}$ is random with $\mathbb{E}(\mathcal{E}) = 0$. The objective is to approximate M by a matrix UV^T where U is $m_1 \times K$, V is $m_2 \times K$ for some $K \ll m_1 \wedge m_2$, and where U, V and M all have nonnegative entries. Note that, under (2), depending on the distribution of $\mathcal{E}$, Y might have some negative entries (the nonnegativity assumption is on M rather than on Y). Our theoretical analysis only requires the following assumption on $\mathcal{E}$.

C1. *The entries $\mathcal{E}_{i,j}$ of $\mathcal{E}$ are i.i.d. with $\mathbb{E}(\varepsilon_{i,j}) = 0$. With the notation $m(x) = \mathbb{E}[\varepsilon_{i,j}\mathbf{1}_{(\varepsilon_{i,j}\le x)}]$ and $F(x) = \mathbb{P}(\varepsilon_{i,j} \le x)$, assume that there exists a nonnegative and bounded function g with $\|g\|_\infty \le 1$ and*

$$\int_u^v m(x)\,dx = \int_u^v g(x)\,dF(x). \tag{3}$$

First, note that if (3) is satisfied for a function g with $\|g\|_\infty = \sigma^2 > 1$, we can replace (2) by the normalized model $Y/\sigma = M/\sigma + \varepsilon/\sigma$ for which C1 is satisfied. The introduction of this rather involved condition is due to the technical analysis of our estimator which is based on Theorem 2 in Section 5. Theorem 2 has first been proved in [15] using Stein's formula with a Gaussian noise. However, Dalalyan and Tsybakov [14] have shown that C1 is actually sufficient to prove Theorem 2. For the sake of understanding, note that (3) is fulfilled when the noise is Gaussian ($\varepsilon_{i,j} \sim \mathcal{N}(0,\sigma^2)$ with $\|g\|_\infty = \sigma^2$) or uniform ($\varepsilon_{i,j} \sim \mathcal{U}[-b,b]$ with $\|g\|_\infty = b^2/2$).

2.2. Prior

We are going to define a prior $\pi(U, V)$, where U is $m_1 \times K$ and V is $m_2 \times K$, for a fixed K. Regarding the choice of K, we prove in Section 3 that our quasi-Bayesian estimator is adaptive, in the sense that if K is chosen much larger than the actual rank of M, the prior will put very little mass on many columns of U and V, automatically shrinking them to 0. This seems to advocate setting a large K prior to the analysis, say $K = m_1 \wedge m_2$. However, keep in mind that the algorithms discussed below have a computational cost growing with K. Anyhow, the following theoretical analysis only requires $2 \leq K \leq m_1 \wedge m_2$.

With respect to the Lebesgue measure on $\mathbb{R}_+$, let us fix a density f such that

$$S_f := 1 \vee \int_0^{\infty} x^2 f(x)\, dx < +\infty.$$

For any $a, x > 0$, let

$$g_a(x) := \frac{1}{a} f\left(\frac{x}{a}\right).$$

We define the prior on U and V by

$$U_{i,\ell}, V_{i,\ell} \text{ indep. } \sim g_{\gamma_\ell}(\cdot),$$

where

$$\gamma_\ell \text{ indep. } \sim h(\cdot)$$

and h is a density on $\mathbb{R}_+$. With the notation $\gamma = (\gamma_1, \ldots, \gamma_K)$, define π by

$$\pi(U, V, \gamma) = \prod_{\ell=1}^{K} \left(\prod_{i=1}^{m_1} g_{\gamma_\ell}(U_{i,\ell})\right)\left(\prod_{j=1}^{m_2} g_{\gamma_\ell}(V_{j,\ell})\right) h(\gamma_\ell) \tag{4}$$

and

$$\pi(U, V) = \int_{\mathbb{R}_+^K} \pi(U, V, \gamma)\, d\gamma.$$

The idea behind this prior is that under h, many γ_ℓ should be small and lead to insignificant columns $U_{\cdot,\ell}$ and $V_{\cdot,\ell}$. In order to do so, we must assume that a non-negligible proportion of the mass of h is located around 0. On the other hand, a non-negligible probability must be assigned to significant values. This is the meaning of the following assumption.

C2. *There exist constants* $0 < \alpha < 1$, $\beta \geq 0$ *and* $\delta > 0$ *such that for any* $0 < \varepsilon \leq \frac{1}{2\sqrt{2} S_f}$,

$$\int_0^{\varepsilon} h(x)\, dx \geq \alpha \varepsilon^{\beta} \qquad \text{and} \qquad \int_1^2 h(x)\, dx \geq \delta.$$

Finally, the following assumption on f is required to prove our main result.

C3. *There exist a non-increasing density* $\widetilde{f}$ *w.r.t. the Lebesgue measure on* $\mathbb{R}_+$ *and a constant* $\mathcal{C}_f > 0$ *such that for any* $x > 0$,

$$f(x) \geq \mathcal{C}_f \widetilde{f}(x).$$

As shown in Theorem 1, the heavier the tails of $\widetilde{f}(x)$, the better the performance of Bayesian NMF.

Note that the general form of (4) encompasses as special cases almost all priors used in the papers mentioned in the Introduction. We end this subsection with classical examples of functions f and h. Regarding f:

1. Exponential prior $f(x) = \exp(-x)$ with $\widetilde{f} = f$, $\mathcal{C}_f = 1$ and $S_f = 2$. This is the choice made by [41]. A generalization of the exponential prior is the gamma prior used in [12].

2. Truncated Gaussian prior $f(x) \propto \exp(2ax - x^2)$ with $a \in \mathbb{R}$.

3. Heavy-tailed prior $f(x) \propto \frac{1}{(1+x)^\zeta}$ with $\zeta > 1$. This choice is inspired by [14] and leads to better theoretical properties.

Regarding h:

1. The uniform distribution on $[0, 2]$ obviously satisfies C2 with $\alpha = 1/2$, $\beta = 1$ and $\delta = 1/2$.

2. The inverse gamma prior $h(x) = \frac{b^a}{\Gamma(a)} \frac{1}{x^{a+1}} \exp\big(-\frac{b}{x}\big)$ is classical in the literature for computational reasons (see for example, [40, 2]), but note that it does not satisfy C2.

3. Alquier, et al. [3] discuss the $\Gamma(a, b)$ choice for $a, b > 0$: both gamma and inverse gamma lead to explicit conditional posteriors for γ (under a restriction on a in the second case), but the gamma distribution led to better numerical performance. When h is the density of the $\Gamma(a, b)$, assumption C2 is satisfied with $\beta = a$, $\alpha = b^a \exp[-b/(2\sqrt{2}S_f)]/\Gamma(a+1)$ and $\delta = \int_1^2 b^a x^{a-1} \exp(-bx)\, dx/\Gamma(a)$.

2.3. Quasi-Posterior and Estimator

We define the quasi-likelihood as

$$\widehat{L}(U, V) = \exp\big[-\lambda \|Y - UV^\top\|_F^2\big]$$

for some fixed parameter $\lambda > 0$. Note that under the assumption that $\varepsilon_{i,j} \sim \mathcal{N}(0, 1/(2\lambda))$, this would be the actual likelihood up to a multiplicative constant. As already pointed out, the use of quasi-likelihoods to define quasi-posteriors is becoming rather popular in Bayesian statistics and machine learning literature. Here, the Frobenius norm is to be viewed as a fitting criterion rather than as a ground truth. Note that other criteria were used in the literature: the Poisson likelihood [28], or the Itakura–Saito divergence [17].

Definition 1. We define the quasi-posterior as

$$\widehat{\rho}_\lambda(U, V, \gamma) = \frac{1}{Z}\widehat{L}(U, V)\pi(U, V, \gamma) = \frac{1}{Z}\exp\big[-\lambda\|Y - UV^\top\|_F^2\big]\pi(U, V, \gamma),$$

where

$$Z := \int \exp\big[-\lambda\|Y - UV^\top\|_F^2\big]\pi(U, V, \gamma)\, d(U, V, \gamma)$$

is a normalization constant. The posterior mean will be denoted by

$$\widehat{M_\lambda} = \int UV^T \widehat{\rho}_\lambda(U, V, \gamma)\, d(U, V, \gamma).$$

Section 3 is devoted to the study the theoretical properties of $\widehat{M_\lambda}$. A short discussion on the implementation will be provided in Section 4.

3. AN ORACLE INEQUALITY

Most likely, the rank of M is unknown in practice. So, as recommended above, we usually choose K much larger than the expected order for the rank, with the hope that many columns of U and V will be shrunk to 0. The following set of matrices is introduced to formalize this idea. For any $r \in \{1, \ldots, K\}$, let $\mathcal{M}_r$ be the set of pairs of matrices (U^0, V^0) with nonnegative entries such that

$$U^0 = \begin{pmatrix} U^0_{11} & \ldots & U^0_{1r} & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ U^0_{m_1 1} & \ldots & U^0_{m_1 r} & 0 & \ldots & 0 \end{pmatrix}, \qquad V^0 = \begin{pmatrix} V^0_{11} & \ldots & V^0_{1r} & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ V^0_{m_2 1} & \ldots & V^0_{m_2 r} & 0 & \ldots & 0 \end{pmatrix}.$$

We also define $\mathcal{M}_r(L)$ as the set of matrices $(U^0, V^0) \in \mathcal{M}_r$ such that, for any (i, j, ℓ), $U^0_{i,\ell}, V^0_{j,\ell} \leq L$.

We are now in a position to state our main theorem, in the form of the following oracle inequality.

Theorem 1. *Fix $\lambda = 1/4$. Under Assumptions C1, C2 and C3,*

$$\mathbb{E}\big(\|\widehat{M}_\lambda - M\|_F^2\big) \leq \inf_{1 \leq r \leq K} \inf_{(U^0,V^0) \in \mathcal{M}_r} \big\{\|U^0 V^{0\top} - M\|_F^2 + \mathcal{R}(r, m_1, m_2, U^0, V^0, \beta, \alpha, \delta, K, S_f, \tilde{f})\big\},$$

where

$$\begin{aligned} &\mathcal{R}\big(r, m_1, m_2, U^0, V^0, \beta, \alpha, K, S_f, \tilde{f}\big) \\ &= 8(m_1 \vee m_2) r \log\left(\sqrt{\frac{2(m_1 \vee m_2)}{r}} \frac{\big(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\big)^2}{\mathcal{C}_f}\right) \\ &\quad + 4 \sum_{\substack{1 \leq i \leq m_1 \\ 1 \leq \ell \leq r}} \log\left(\frac{1}{\widetilde{f}(U^0_{i\ell} + 1)}\right) + 4 \sum_{\substack{1 \leq j \leq m_2 \\ 1 \leq \ell \leq r}} \log\left(\frac{1}{\widetilde{f}(V^0_{j\ell} + 1)}\right) \\ &\quad + r\left[8 + \log\left(\frac{1}{\delta}\right)\right] + 4\log(4) + 4K \log\left(\frac{1}{\alpha}\right) \\ &\quad + 4\beta K \log\left(\frac{2 S_f \sqrt{m_1 m_2}\big(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\big)^2}{r}\right). \end{aligned}$$

We remind the reader that the proof is given in Section 5. The main message of the theorem is that $\widehat{M}_\lambda$ is as close to M as would be an estimator designed with the actual knowledge of its rank (i.e., $\widehat{M}_\lambda$ is adaptive to r), up to remainder terms. These terms might be difficult to read. In order to demonstrate the rate of convergence, we now provide a weaker version, where we only compare $\widehat{M}_\lambda$ with the best factorization in $\mathcal{M}_r(L)$.

Corollary 1. *Fix $\lambda = 1/4$. Under Assumptions C1, C2 and C3,*

$$\begin{aligned} \mathbb{E}\big(\|\widehat{M}_\lambda - M\|_F^2\big) \leq \inf_{1 \leq r \leq K} \inf_{(U^0,V^0) \in \mathcal{M}_r(L)} &\left\{\|U^0 V^{0\top} - M\|_F^2 + 8(m_1 \vee m_2) r \log\left(\frac{2(L+1)^2 m_1 m_2}{\mathcal{C}_f \tilde{f}(L+1)}\right)\right. \\ &\left. + r\left[8 + \log\left(\frac{1}{\delta}\right)\right] + 4\log(4) + 4\beta K \log\big(2 S_f (L+1)^2 m_1 m_2\big) + 4K \log\left(\frac{1}{\alpha}\right)\right\}. \end{aligned}$$

First, note that when $L^2 = \mathcal{O}(1)$, the magnitude of the error bound is

$$(m_1 \vee m_2) r \log(m_1 m_2),$$

which is roughly the variance multiplied by the number of parameters to be estimated in any $(U^0, V^0) \in \mathcal{M}_r(L)$. Alternatively, when $M \in \mathcal{M}_r(L)$ only for huge L, the log term in

$$8(m_1 \vee m_2) r \log\left(\frac{(L+1)^2 m_1 m_2}{\tilde{f}(L+1)}\right)$$

becomes significant. Indeed, in the case of the truncated Gaussian prior $f(x) \propto \exp(2ax - x^2)$, the previous quantity is in

$$8(m_1 \vee m_2) r L^2 \log(L m_1 m_2)$$

which is terrible for large L. On the contrary, with the heavy-tailed prior $f(x) \propto (1+x)^{-\zeta}$ (as in [14]), the leading term is

$$8(m_1 \vee m_2) r (\zeta + 2) \log(L m_1 m_2)$$

which is way more satisfactory. Still, this prior has not received much attention from practitioners.

Remark 1. When (3) in C1 is satisfied with $\|g\|_\infty = \sigma^2 > 1$ we already remarked that it is necessary to use the normalized model $Y/\sigma = M/\sigma + \mathcal{E}/\sigma$ in order to apply Theorem 1. Going back to the original model, we get that, for $\lambda = 1/(4\sigma^2)$,

$$\mathbb{E}\big(\|\widehat{M}_\lambda - M\|_F^2\big) \leq \inf_{1 \leq r \leq K} \inf_{(U^0, V^0) \in \mathcal{M}_r} \big\{ \|U^0 V^{0\top} - M\|_F^2 + \sigma^2 \mathcal{R}(r, m_1, m_2, U^0, V^0, \beta, \alpha, \delta, K, S_f, \tilde{f}) \big\}.$$

4. ALGORITHMS FOR BAYESIAN NMF

As the quasi-Bayesian estimator takes the form of a Bayesian estimator in a special model, we can obviously use tools from computational Bayesian statistics to compute it. The method of choice for computing Bayesian estimators for complex models is Monte-Carlo Markov Chain (MCMC). In the case of Bayesian matrix factorization, the Gibbs sampler was considered in the literature: see for example [40, 3] for the general case and [37, 41, 49] for NMF. The Gibbs sampler (described in its general form in [5], for example), is given by Algorithm 1.

Algorithm 1 Gibbs sampler.

Input Y, λ.

Initialization $U^{(0)}, V^{(0)}, \gamma^{(0)}$.

For $k = 1, \ldots, N$:

- **For** $i = 1, \ldots, m_1$: draw $U_{i,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(U_{i,\cdot} | V^{(k-1)}, \gamma^{(k-1)}, Y)$.
- **For** $j = 1, \ldots, m_2$: draw $V_{j,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(V_{j,\cdot} | U^{(k)}, \gamma^{(k-1)}, Y)$.
- **For** $\ell = 1, \ldots, K$: draw $\gamma_\ell^{(k)} \sim \widehat{\rho}_\lambda(\gamma_\ell | U^{(k)}, V^{(k)}, Y)$.

In the aforementioned papers, there are discussions on the choice of f and h that lead to explicit forms for the conditional posteriors of $U_{i,\cdot}$, $V_{j,\cdot}$ and γ_ℓ, leading to fast algorithms. We refer the reader to these papers for detailed descriptions of the algorithm in this case, and for exhaustive simulation studies.

Optimization methods used for (non-Bayesian) NMF are much faster than the MCMC methods used for Bayesian NMF though: the original multiplicative algorithm [28, 29], projected gradient descent [33, 20], second order schemes [25], linear progamming [7], ADMM (alternative direction method of multipliers [8, 48]), block coordinate descent [47] among others.

We believe that an efficient implementation of Bayesian and quasi-Bayesian methods will be based on fast optimisation methods, like Variational Bayes (VB) or Expectation–Progapation (EP) methods [24, 34, 5]. VB was used for Bayesian matrix factorization [32, 3] and more recently in Bayesian NMF [39] with promising results. Still, there is no proof that these algorithms provide valid results. To the best of our knowledge, the first attempt to study the convergence of the VB to the target distribution is made in [4] for a family of problems, that do not include NMF. We believe that further investigation in this direction is necessary.

5. PROOFS

This section contains the proof to the main theoretical claim of the paper (Theorem 1).

5.1. A PAC-Bayesian Bound from [14]

The analysis of quasi-Bayesian estimators with PAC bounds started with [43]. McAllester improved on the initial method and introduced the name "PAC-Bayesian bounds" [36]. Catoni also improved these results to derive sharp oracle inequalities [9, 10, 11]. This methods were used in various complex models of statistical learning [21, 2, 44, 35, 22, 18, 31]. Dalalyan and Tsybakov [14] proved a different PAC-Bayesian bound based on the idea of unbiased risk estimation (see [30]). We first recall its form in the context of matrix factorization.

Theorem 2. *Under* C1*, as soon as* $\lambda \leq 1/4$*,*

$$\mathbb{E}\|\widehat{M}_\lambda - M\|_F^2 \leq \inf_\rho \left\{ \int \|UV^\top - M\|_F^2 \rho(U,V,\gamma)\, d(U,V,\gamma) + \frac{\mathcal{K}(\rho,\pi)}{\lambda} \right\},$$

where the infimum is taken over all probability measures ρ *absolutely continuous with respect to* π*, and* $\mathcal{K}(\mu,\nu)$ *denotes the Kullback–Leibler divergence between two measures* μ *and* ν*.*

We let the reader check that the proof in [14], stated for vectors, is still valid for matrices (also, the result [14] is actually stated for any σ^2, we only use the case $\sigma^2 = 1$).

The end of the proof of Theorem 1 is organized as follows. First, we define in Section 5.2 a parametric family of probability distributions ρ:

$$\left\{\rho_{r,U^0,V^0,c} \colon c > 0, 1 \leq r \leq K, (U^0,V^0) \in \mathcal{M}_r\right\}.$$

We then upper bound the infimum over all ρ by the infimum over this parametric family. So, we have to calculate, or upper bound

$$\int \|UV^\top - M\|_F^2 \rho_{r,U^0,V^0,c}(U,V,\gamma)\, d(U,V,\gamma)$$

and

$$\mathcal{K}(\rho_{r,U^0,V^0,c},\pi).$$

This is done in two lemmas in Sections 5.2 and 5.4, respectively. We finally gather all the pieces together in Section 5.5, and optimize with respect to c.

5.2. A Parametric Family of Factorizations

We define, for any $r \in \{1,\ldots,K\}$ and any pair of matrices $(U^0,V^0) \in \mathcal{M}_r$, for any $0 < c \leq \sqrt{Kr}$, the density

$$\rho_{r,U^0,V^0,c}(U,V,\gamma) = \frac{\mathbf{1}_{\{\|U-U^0\|_F \leq c, \|V-V^0\|_F \leq c\}} \pi(U,V,\gamma)}{\pi\big(\{\|U-U^0\|_F \leq c, \|V-V^0\|_F \leq c\}\big)}.$$

5.3. Upper Bound for the Integral Part

Lemma 5.1. *We have*

$$\int \|UV^\top - M\|_F^2 \rho_{r,U^0,V^0,c}(U,V,\gamma)\, d(U,V,\gamma) \leq \|U^0V^{0\top} - M\|_F^2 + 4c^2\big(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\big)^2.$$

Proof. Note that (U,V) belonging to the support of $\rho_{r,U^0,V^0,c}$ implies that

$$\begin{aligned}\|UV^\top - U^0V^{0\top}\|_F &= \|U(V^\top - V^{0\top}) + (U-U^0)V^{0\top}\|_F \\ &\leq \|U(V^\top - V^{0\top})\|_F + \|(U-U^0)V^{0\top}\|_F \\ &\leq \|U\|_F\|V-V^0\|_F + \|U-U^0\|_F\|V^0\|_F \\ &\leq (\|U^0\|_F + c)c + c\|V^0\|_F = c\big(\|U^0\|_F + \|V^0\|_F + c\big).\end{aligned}$$

Now, let Π be the orthogonal projection on the set

$$\Big\{M^0\colon \|M^0 - U^0V^{0\top}\|_F \leq c\big(\|U^0\|_F + \|V^0\|_F + c\big)\Big\}$$

with respect to the Frobenius norm. Note that

$$\begin{aligned}\|UV^\top - M\|_F^2 &\leq \|UV^\top - \Pi(M)\|_F^2 + \|\Pi(M) - M\|_F^2 \\ &\leq \big[2c\big(\|U^0\|_F + \|V^0\|_F + c\big)\big]^2 + \|U^0V^{0\top} - M\|_F^2.\end{aligned}$$

Integrate with respect to $\rho_{r,U^0,V^0,c}$ and use $c \leq \sqrt{Kr}$ to get the result. □

5.4. Upper Bound for the Kullback–Leibler Divergence

Lemma 5.2. *Under C2 and C3,*

$$\begin{aligned}\mathcal{K}(\rho_{r,U^0,V^0,c},\pi) \leq{}& 2(m_1 \vee m_2) r \log\left(\frac{\sqrt{2(m_1\vee m_2)r}}{c\mathcal{C}_f}\right) \\ &+ \sum_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(U^0_{i\ell}+1)}\right) + \sum_{\substack{1\leq j\leq m_2\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(V^0_{j\ell}+1)}\right) \\ &+ \beta K \log\left(\frac{2S_f\sqrt{2Km_1m_2}}{c}\right) + K\log\left(\frac{1}{\alpha}\right) + r\log\left(\frac{1}{\delta}\right) + \log(4).\end{aligned}$$

Proof. By definition

$$\begin{aligned}\mathcal{K}(\rho_{r,U^0,V^0,c},\pi) &= \int \rho_{r,U^0,V^0,c}(U,V,\gamma)\log\left(\frac{\rho_{r,U^0,V^0,c}(U,V,\gamma)}{\pi(U,V,\gamma)}\right) d(U,V,\gamma) \\ &= \log\left(\frac{1}{\int \mathbf{1}_{\{\|U-U^0\|_F\leq c, \|V-V^0\|_F\leq c\}}\pi(U,V,\gamma)\,d(U,V,\gamma)}\right).\end{aligned}$$

Then, note that

$$\begin{aligned}&\int \mathbf{1}_{\{\|U-U^0\|_F\leq c,\|V-V^0\|_F\leq c\}}\pi(U,V,\gamma)\,d(U,V,\gamma) \\ &= \int\left(\int \mathbf{1}_{\{\|U-U^0\|_F\leq c,\|V-V^0\|_F\leq c\}}\pi(U,V|\gamma)\,d(U,V)\right)\pi(\gamma)\,d\gamma \\ &= \underbrace{\int\left(\int \mathbf{1}_{\{\|U-U^0\|_F\leq c\}}\pi(U|\gamma)\,dU\right)\pi(\gamma)\,d\gamma}_{=:I_1}\underbrace{\int\left(\int \mathbf{1}_{\{\|V-V^0\|_F\leq c\}}\pi(V|\gamma)\,dV\right)\pi(\gamma)\,d\gamma}_{=:I_2}.\end{aligned}$$

So we have to lower bound I_1 and I_2. We deal only with I_1, as the method to lower bound I_2 is exactly the same. We define the set $E \subset \mathbb{R}^K$ as

$$E = \left\{\gamma \in \mathbb{R}^K\colon \gamma_1,\ldots,\gamma_r \in (1,2] \text{ and } \gamma_{r+1},\ldots,\gamma_K \in \left(0, \frac{c}{2S_f\sqrt{2Km_1}}\right]\right\}.$$

Then

$$\int\left(\int \mathbf{1}_{\{\|U-U^0\|_F\leq c\}}\pi(U\mid\gamma)\,dU\right)\pi(\gamma)\,d\gamma \geq \int_E\left(\underbrace{\int \mathbf{1}_{\{\|U-U^0\|_F\leq c\}}\pi(U\mid\gamma)\,dU}_{=:I_3}\right)\pi(\gamma)\,d\gamma$$

and we first focus on a lower-bound for I_3 when $\gamma\in E$:

$$\begin{aligned}
I_3 &= \pi\Bigg(\sum_{\substack{1\leq i\leq m_1\\1\leq\ell\leq K}}(U_{i,\ell}-U^0_{i,\ell})^2\leq c^2\mid\gamma\Bigg)\\
&= \pi\Bigg(\sum_{\substack{1\leq i\leq m_1\\1\leq\ell\leq r}}(U_{i,\ell}-U^0_{i,\ell})^2+\sum_{\substack{1\leq i\leq m_1\\r+1\leq\ell\leq K}}U^2_{i,\ell}\leq c^2\mid\gamma\Bigg)\\
&\geq \pi\Bigg(\sum_{\substack{1\leq i\leq m_1\\r+1\leq\ell\leq K}}U^2_{i,\ell}\leq\frac{c^2}{2}\mid\gamma\Bigg)\pi\Bigg(\sum_{\substack{1\leq i\leq m_1\\1\leq\ell\leq r}}(U_{i,\ell}-U^0_{i,\ell})^2\leq\frac{c^2}{2}\mid\gamma\Bigg)\\
&\geq \underbrace{\pi\Bigg(\sum_{\substack{1\leq i\leq m_1\\r+1\leq\ell\leq K}}U^2_{i,\ell}\leq\frac{c^2}{2}\mid\gamma\Bigg)}_{=:I_4}\prod_{\substack{1\leq i\leq m_1\\1\leq\ell\leq r}}\pi\Bigg((U_{i,\ell}-U^0_{i,\ell})^2\leq\frac{c^2}{2m_1r}\mid\gamma\Bigg).
\end{aligned}$$

Now, using Markov's inequality,

$$\begin{aligned}
1-I_4 &= \pi\Bigg(\sum_{\substack{1\leq i\leq m_1\\r+1\leq\ell\leq K}}U^2_{i,\ell}\geq\frac{c^2}{2}\mid\gamma\Bigg)\\
&\leq 2\frac{\mathbb{E}_\pi\big(\sum_{\substack{1\leq i\leq m_1\\r+1\leq\ell\leq K}}U^2_{i,\ell}\mid\gamma\big)}{c^2}=2\frac{\sum_{\substack{1\leq i\leq m_1\\r+1\leq\ell\leq K}}\gamma_j^2S_f^2}{c^2}\leq\frac{1}{2},
\end{aligned}$$

and as on E, for $\ell\geq r+1$, $\gamma_j\leq c/(2S_f\sqrt{2Km_1})$. So

$$I_4\geq\frac{1}{2}.$$

Next, we remark that

$$\begin{aligned}
\pi\Bigg(\big(U_{i,\ell}-U^0_{i,\ell}\big)^2\leq\frac{c^2}{2m_1r}\mid\gamma\Bigg) &\geq \int_{U^0_{i,\ell}}^{U^0_{i,\ell}+\frac{c}{\sqrt{2m_1r}}}\frac{1}{\gamma_j}f\left(\frac{u}{\gamma_j}\right)du\\
&\geq \int_{U^0_{i,\ell}}^{U^0_{i,\ell}+\frac{c}{\sqrt{2m_1r}}}\frac{\mathcal{C}_f}{\gamma_j}\widetilde{f}\left(\frac{u}{\gamma_j}\right)du.
\end{aligned}$$

Recall that $1\leq\gamma_j\leq2$ and $\tilde{f}$ is non-increasing, so

$$\begin{aligned}
\pi\Bigg(\big(U_{i,\ell}-U^0_{i,\ell}\big)^2\leq\frac{c^2}{2m_1r}\mid\gamma\Bigg) &\geq \frac{2c\mathcal{C}_f}{\sqrt{2m_1r}}\widetilde{f}\left(U^0_{i,\ell}+\frac{c}{\sqrt{2m_1r}}\right)\\
&\geq \frac{2c\mathcal{C}_f}{\sqrt{2m_1r}}\widetilde{f}\big(U^0_{i,\ell}+1\big)
\end{aligned}$$

as $c\leq\sqrt{Kr}\leq\sqrt{m_1r}$. We plug this result and the lower-bound $I_4\geq1/2$ into the expression of I_3 to get

$$I_3\geq\frac{1}{2}\left(\frac{2c\mathcal{C}_f}{\sqrt{2m_1r}}\right)^{m_1r}\Bigg[\prod_{\substack{1\leq i\leq m_1\\1\leq\ell\leq r}}\widetilde{f}(U^0_{i,\ell}+1)\Bigg].$$

So

$$\begin{aligned} I_1 &\geq \int_E I_3 \pi(\gamma)\, d\gamma \\ &= \frac{1}{2}\left(\frac{2c\mathcal{C}_f}{\sqrt{2m_1 r}}\right)^{m_1 r} \Bigg[\prod_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \widetilde{f}(U^0_{i,\ell}+1)\Bigg] \int_E \pi(\gamma)\, d\gamma \\ &= \frac{1}{2}\left(\frac{2c\mathcal{C}_f}{\sqrt{2m_1 r}}\right)^{m_1 r} \Bigg[\prod_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \widetilde{f}(U^0_{i,\ell}+1)\Bigg] \left(\int_1^2 h(x)\, dx\right)^r \left(\int_0^{\frac{c}{2S_f\sqrt{2Km_1}}} h(x)\, dx\right)^{K-r} \\ &\geq \frac{1}{2}\left(\frac{2c\mathcal{C}_f}{\sqrt{2m_1 r}}\right)^{m_1 r} \Bigg[\prod_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \widetilde{f}(U^0_{i,\ell}+1)\Bigg] \delta^r \alpha^{K-r} \left(\frac{c}{2S_f\sqrt{2Km_1}}\right)^{\beta(K-r)} \\ &\geq \frac{1}{2}\left(\frac{2c\mathcal{C}_f}{\sqrt{2m_1 r}}\right)^{m_1 r} \Bigg[\prod_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \widetilde{f}(U^0_{i,\ell}+1)\Bigg] \delta^r \alpha^{K} \left(\frac{c}{2S_f\sqrt{2Km_1}}\right)^{\beta K}, \end{aligned}$$

using C2. Proceeding exactly in the same way,

$$I_2 \geq \frac{1}{2}\left(\frac{2c\mathcal{C}_f}{\sqrt{2m_2 r}}\right)^{m_2 r} \Bigg[\prod_{\substack{1\leq j\leq m_2\\ 1\leq \ell\leq r}} \widetilde{f}(V^0_{j,\ell}+1)\Bigg] \delta^r \alpha^{K} \left(\frac{c}{2S_f\sqrt{2Km_2}}\right)^{\beta K}.$$

We combine these inequalities, and we use trivia between m_1, m_2, $m_1 \vee m_2$ and $m_1 + m_2$ to obtain

$$\begin{aligned} \mathcal{K}(\rho_{r,U^0,V^0,c}, \pi) &\leq 2(m_1\vee m_2) r \log\left(\frac{\sqrt{2(m_1\vee m_2)r}}{c\mathcal{C}_f}\right) \\ &\quad + \sum_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(U^0_{i\ell}+1)}\right) + \sum_{\substack{1\leq j\leq m_2\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(V^0_{j\ell}+1)}\right) \\ &\quad + \beta K \log\left(\frac{2S_f\sqrt{2Km_1m_2}}{c}\right) + K\log\left(\frac{1}{\alpha}\right) + r\log\left(\frac{1}{\delta}\right) + \log(4). \end{aligned}$$

This ends the proof of the lemma. □

5.5. Conclusion

We now plug Lemmas 5.1 and 5.2 into Theorem 2. We obtain, under C1, C2 and C3,

$$\begin{aligned} \mathbb{E}\big(\|\widehat{M}_\lambda - M\|_F^2\big) \leq \inf_{1\leq r\leq K}\ \inf_{(U^0,V^0)\in\mathcal{M}_r}\ \inf_{0<c\leq\sqrt{Kr}} \Bigg\{ &\|U^0V^{0\top} - M\|_F^2 \\ &+ \frac{2(m_1\vee m_2)r}{\lambda}\log\left(\frac{\sqrt{2(m_1\vee m_2)r}}{c\mathcal{C}_f}\right) \\ &+ \frac{1}{\lambda}\sum_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(U^0_{i\ell}+1)}\right) + \frac{1}{\lambda}\sum_{\substack{1\leq j\leq m_2\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(V^0_{j\ell}+1)}\right) \\ &+ \frac{\beta K}{\lambda}\log\left(\frac{2S_f\sqrt{2Km_1m_2}}{c}\right) + \frac{K}{\lambda}\log\left(\frac{1}{\alpha}\right) + \frac{r}{\lambda}\log\left(\frac{1}{\delta}\right) + \frac{1}{\lambda}\log(4) \\ &+ 4c\big(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\big)^2 \Bigg\}. \end{aligned}$$

Recall that we fixed $\lambda = \frac{1}{4}$. We finally choose

$$c = \frac{2r}{(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr})^2} \leq \frac{2r}{Kr} = \frac{2}{K}$$

and so the condition $c \leq \sqrt{Kr}$ is always satisfied as we imposed $K \geq 2$. The inequality becomes

$$\begin{aligned}\mathbb{E}\big(\|\widehat{M}_\lambda - M\|_F^2\big) \leq \inf_{1\leq r\leq K} \inf_{(U^0,V^0)\in\mathcal{M}_r} \Bigg\{ &\|U^0 V^{0\top} - M\|_F^2 \\ &+ 8(m_1 \vee m_2) r \log\left(\sqrt{\frac{2(m_1\vee m_2)}{r}} \frac{\big(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\big)^2}{\mathcal{C}_f}\right) \\ &+ 4 \sum_{\substack{1\leq i\leq m_1 \\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(U^0_{i\ell}+1)}\right) + 4 \sum_{\substack{1\leq j\leq m_2 \\ 1\leq \ell\leq r}} \log\left(\frac{1}{\widetilde{f}(V^0_{j\ell}+1)}\right) \\ &+ 4\beta K \log\left(\frac{2S_f\sqrt{m_1 m_2}\big(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\big)^2}{r}\right) \\ &+ r\left[8 + 4\log\left(\frac{1}{\delta}\right)\right] + 4K\log\left(\frac{1}{\alpha}\right) + 4\log(4)\Bigg\},\end{aligned}$$

which ends the proof. □

ACKNOWLEDGMENTS

The authors are grateful to Jialia Mei and Yohann de Castro (Université Paris-Sud) for insightful discussions and for providing many references on NMF, and to the anonymous Referee for helpful comments.

The first author gratefully acknowledges financial support from the research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*, from Labex ECODEC (ANR-11-LABEX-0047) and from Labex CEMPI (ANR-11-LABX-0007-01).

REFERENCES

1. G. I. Allen, L. Grosenick, and J. Taylor, "A Generalized Least-Square Matrix Decomposition", J. Amer. Statist. Assoc. **109** (505), 145–159 (2014).
2. P. Alquier, "Bayesian Methods for Low-Rank Matrix Estimation: Short Survey and Theoretical Study", in *Algorithmic Learning Theory 2013* (Springer, 2013), pp. 309–323.
3. P. Alquier, V. Cottet, N. Chopin, and J. Rousseau, *Bayesian Matrix Completion: Prior Specification*, Preprint arXiv:1406.1440 (2014).
4. P. Alquier, J. Ridgway, and N. Chopin, *On the Properties of Variational Approximations of Gibbs Posteriors*, J. Machine Learning Res., **17** (239), 1-41 (2016).
5. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006), Chapter 10.
6. P. G. Bissiri, C. C. Holmes, and S. G. Walker, *A General Framework for Updating Belief Distributions*, J. Roy. Statist. Soc. Ser. B, **78** (5) (2016).
7. V. Bittorf, B. Recht, C. Re, and J. Tropp, "Factoring Nonnegative Matrices with Linear Programs", in *Advances in Neural Information Processing Systems* (2012), pp. 1214–1222.
8. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", Foundations and Trends in Machine Learning **3**(1), 1–122 (2011).
9. O. Catoni, *A PAC-Bayesian Approach to Adaptive Classification*, Preprint Laboratoire de Probabilités et Modèles Aléatoires, PMA-840 (2003).
10. O. Catoni, *Statistical Learning Theory and Stochastic Optimization*, in *Saint-Flour Summer School on Probability Theory 2001*, Ed. by Jean Picard (Springer, 2004).
11. O. Catoni, *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, in *Inst. of Math. Statist. Lecture Notes—Monograph Series* (IMS, Beachwood, OH, 2007), Vol. 56.
12. A. T. Cemgil, "Bayesian Inference for Nonnegative Matrix Factorization Models", Computational Intelligence and Neuroscience (2009).

13. J. Corander and M. Villani, “Bayesian Assessment of Dimensionality in Reduced Rank Regression”, Statistica Neerlandica **58**, 255–270 (2004).
14. A. Dalalyan and A. B. Tsybakov, “Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity”, Machine Learning **72** (1–2), 39–61 (2008).
15. A. S. Dalalyan and A. B. Tsybakov, “Aggregation by Exponential Weighting and Sharp Oracle Inequalities”, in *Lecture Notes in Computer Science*, Vol. 4539: *Learning Theory*, Ed. by N. Bshouty and C. Gentile (Springer, Berlin–Heidelberg, 2007), pp. 97–111.
16. D. Donoho and V. Stodden, “When Does Nonnegative Matrix Factorization Give a Correct Decomposition into Parts?”, in *Advances in Neural Information Processing Systems* (2003).
17. C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative Matrix Factorization with the Itakura–Saito Divergence: With Application to Music Analysis”, Neural Computation **21** (3), 793–830 (2009).
18. I. Giulini, *PAC-Bayesian Bounds for Principal Component Analysis in Hilbert Spaces*, Preprint arXiv:1511.06263 (2015).
19. Y. Golubev and D. Ostrovski, “Concentration Inequalities for the Exponential Weighting Method”, Math. Methods Statist. **23** (1), 20–37 (2014).
20. N. Guan, D. Tao, Z. Luo, and B. Yuan, “NeNMF: An Optimal Gradient Method for Nonnegative Matrix Factorization”, IEEE Trans. on Signal Processing **60** (6), 2882–2898 (2012).
21. B. Guedj and P. Alquier, “PAC-Bayesian Estimation and Prevision in Sparse Additive Models”, Electronic J. Statist, **7**, 264–291 (2013).
22. B. Guedj and S. Robbiano, *PAC-Bayesian High Dimensional Bipartite Ranking*, Preprint arXiv: 1511.02729 (2015).
23. D. Guillamet and J. Vitria, “Classifying Faces with Nonnegative Matrix Factorization”, in *Proc. 5th Catalan Conference for Artificial Intelligence* (2002), pp. 24–31.
24. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An Introduction to Variational Methods for Graphical Models”, Machine Learning **37**, 183–233 (1999).
25. D. Kim, S. Sra, and I. S. Dhillon, “Fast Projection-Based Methods for the Least Squares Nonnegative Matrix Approximation Problem”, Statist. Analysis and Data Mining **1** (1), 38–51 (2008).
26. Y. Koren, R. Bell, and C. Volinsky, “Matrix Factorization Techniques for Recommender Systems”, Computer **42** (8), 30–37 (2009).
27. N. D. Lawrence and R. Urtasun, “Nonlinear Matrix Factorization with Gaussian Processes”, in *Proc. 26th Annual Internat. Conf. on Machine Learning* (ACM, 2009), pp. 601–608.
28. D. D. Lee and H. S. Seung, “Learning the Parts of Objects by Nonnegative Matrix Factorization”, Nature **401** (6755), 788–791 (1999).
29. D. D. Lee and H. S. Seung, “Algorithms for Nonnegative Matrix Factorization”, in *Adv. in Neural Inform. Processing Systems* (2001), pp. 556–562.
30. G. Leung and A. R. Barron, “Information Theory and Mixing Least-Squares Regressions”, IEEE Trans. Inform. Theory **52**(8), 3396–3410 (2006).
31. L. Li, B. Guedj, and S. Loustau, *PAC-Bayesian Online Clustering*, Preprint arXiv:1602.00522 (2016).
32. Y. J. Lim and Y. W. Teh, “Variational Bayesian Approach to Movie Rating Prediction”, in *Proc. KDD Cup and Workshop* (2007), Vol. 7, pp. 5–21.
33. C.-J. Lin, “Projected Gradient Methods for Nonnegative Matrix Factorization”, Neural Computation **19** (10), 2756–2779 (2007).
34. D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge Univ. Press, 2002).
35. T. T. Mai and P. Alquier, “A Bayesian Approach for Matrix Completion: Optimal Rates under General Sampling Distributions”, Electronic J. Statist. **9**, 823–841 (2015).
36. D. McAllester, “Some PAC-Bayesian Theorems”, in *Proc. 11th Annual Conf. on Comput. Learning Theory* (ACM, New York, 1998), pp. 230–234.
37. S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, “Separation of Nonnegative Mixture of Nonnegative Sources Using a Bayesian Approach and MCMC Sampling”, IEEE Trans. on Signal Processing **54** (11), 4133–4145 (2006).
38. A. Ozerov and C. Févotte, “Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation”, IEEE Trans. on Audio, Speech, and Language Processing **18** (3), 550–563 (2010).
39. J. Paisley, D. Blei, and M. I. Jordan, *Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference*, in *Handbook of Mixed Membership Models and Their Applications* (Chapman and Hall/CRC, 2015), Chapter 11.
40. R. Salakhutdinov and A. Mnih, “Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo”, in *Proc. 25th Internat. Conf. on Machine Learning* (ACM, 2008), pp. 880–887.

41. M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian Nonnegative Matrix Factorization", in *Independent Component Analysis and Signal Separation* (Springer, 2009), pp. 540–547.
42. F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document Clustering Using Nonnegative Matrix Factorization", Inform. Processing & Management **42** (2), 373–386 (2006).
43. J. Shawe-Taylor and R. Williamson, "A PAC Analysis of a Bayes Estimator", in *Proc. 10th Annual Conf. on Comput. Learning Theory* (ACM, New York, 1997), pp. 2–9.
44. T. Suzuki, "Convergence Rate of Bayesian Tensor Estimator and Its Minimax Optimality", in *Proc. 32nd Internat. Conf. on Machine Learning (Lille, 2015)* (2015), pp. 1273–1282.
45. V. Y. Tan and C. Févotte, "Automatic Relevance Determination in Nonnegative Matrix Factorization", in *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations* (2009).
46. W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Nonnegative Matrix Factorization", in *Proc. 26th Annual Internat. ACM SIGIR Conf. on Research and Development in Inform. Retrieval* (ACM, 2003), pp. 267–273.
47. Y. Xu and W. Yin, "A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion", SIAM J. on Imaging Sci. **6** (3), 1758–1789 (2013).
48. Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An Alternating Direction Algorithm for Matrix Completion with Nonnegative Factors", Frontiers of Mathematics in China **7** (2), 365–384 (2012).
49. M. Zhong and M. Girolami, "Reversible Jump MCMC for Nonnegative Matrix Factorization", in *Internat. Conf. Artificial Intelligence and Statist.* (2009), pp. 663–670.
50. M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin, *Nonparametric Bayesian Matrix Completion*, in *Proc. IEEE SAM* (2010).