Bayesian Non-Negative Matrix Factorization

Pierre Alquier, Benjamin Guedj

Bayesian Non-Negative Matrix Factorization

Pierre Alquier* & Benjamin Guedj†

January 6, 2016

Abstract

The aim of this paper is to provide some theoretical understanding of Bayesian non-negative matrix factorization methods, along with practical implementations. We provide a sharp oracle inequality for a quasi-Bayesian estimator, also known as the exponentially weighted aggregate (Dalalyan and Tsybakov, 2008). This result holds for a very general class of prior distributions and shows how the prior affects the rate of convergence. We then discuss possible algorithms. A natural choice in Bayesian statistics is the Gibbs sampler, used for example in Salakhutdinov and Mnih (2008). This algorithm is asymptotically exact, yet it suffers from the fact that the convergence might be very slow on large datasets. When faced with massive datasets, a more efficient path is to use approximate methods based on optimisation algorithms: we here describe a blockwise gradient descent which is a Bayesian version of the algorithm in Xu et al. (2012). Here again, the general form of the algorithm helps to understand the role of the prior, and some priors will clearly lead to more efficient (*i.e.*, faster) implementations. We end the paper with a short simulation study and an application to finance. These numerical studies support our claim that the reconstruction of the matrix is usually not very sensitive to the choice of the hyperparameters whereas rank identification is.

*CREST, ENSAE, Université Paris Saclay, pierre.alquier@ensae.fr. This author gratefully acknowledges support from the research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*.

†Modal project-team, Inria, benjamin.guedj@inria.fr.

1 Introduction

Non-negative matrix factorization (NMF) is a set of algorithms in high-dimensional data analysis which aims at factorizing a large matrix Y, say $m_1 \times m_2$, with nonnegative entries, as a product of two matrices of smaller dimension: $Y \simeq UV^T$ 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 non-negative entries. Interpreting the columns $Y_{\cdot,j}$ of Y as (non-negative) signals, NMF amounts to decompose each signal as a combination of the "elementary" signals $U_{\cdot,1}, \ldots, U_{\cdot,K}$:

$$Y_{\cdot,j} \simeq \sum_{\ell=1}^{K} V_{j,\ell} U_{\cdot,\ell}.$$

Since the seminal paper from Lee and Seung (1999), NMF was successfully applied to such various fields as image processing and face classification (Guillamet and Vitria, 2002), separation of sources in audio and video processing (Ozerov and Févotte, 2010), collaborative filtering and recommender systems on the web (Koren et al., 2009), document clustering (Xu et al., 2003; Shahnaz et al., 2006), medical image processing (Allen et al., 2014) or topic extraction in texts (Paisley et al., 2015). Many algorithms were proposed, e.g. in Lee and Seung (2001) and Arora et al. (2012). Among the most popular family of algorithms, let us mention variants of gradient descent (Lin, 2007; Guan et al., 2012), linear programming (Recht et al., 2012), alternative direction algorithm (Xu et al., 2012) or block coordinate descent (Xu and Yin, 2013). While all these algorithm usually provide an interpretable set of elementary signals U, the non-unicity of the decomposition $Y \simeq UV^T$ makes it hard to understand from a theoretical point of view. Some progress in this direction was made by Donoho and Stodden (2003).

From a statistical perspective, most of the algorithms mentioned above can be interpreted as the minimization of a penalized least-square criterion. Yet Bayesian methods for (general) matrix factorization were recently proposed (Corander and Villani, 2004; Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Lawrence and Urtasun, 2009; Zhou et al., 2010). Theoretical guarantees on the statistical performance of these methods were proved by Alquier (2013), Mai and Alquier (2015) and Suzuki (2015). The algorithms used are essentially the Gibbs sampler (Salakhutdinov and Mnih, 2008) and variational Bayes methods (Lim and Teh, 2007): we refer the reader to Alquier et al. (2014) for a comprehensive survey and a simulation study. Finally, one of the first attempts in Bayesian NMF can be found in Paisley et al. (2015).

The aim of this paper is to provide some theoretical and experimental facts that should help understand the performance of Bayesian NMF. First, we

prove a sharp oracle inequality on the reconstruction of the matrix Y by a quasi-Bayesian estimator, in the sense that the likelihood is not necessarily assumed to be true. The use of Bayesian estimation in this context is advocated in the Bayesian literature (Bissiri et al., 2013). It is also well known in machine learning theory as the PAC-Bayesian approach (Shawe-Taylor and Williamson, 1997; McAllester, 1998; Catoni, 2003, 2004, 2007; Guedj and Alquier, 2013; Guedj and Robbiano, 2015; Giulini, 2015, among many other references). We actually use PAC-Bayesian theoretical tools from Dalalyan and Tsybakov (2008) to prove our oracle inequality.

We then review different approaches to implement the quasi-Bayesian estimators: the Gibbs sampler and optimization algorithms. The Gibbs sampler was used successfully for quasi-Bayesian estimators by Guedj and Alquier (2013) in high-dimensional regression and Guedj and Robbiano (2015) for the ranking problem. Optimization topics in the Bayesian literature include variational Bayes algorithms (Jordan et al., 1999; MacKay, 2002; Bishop, 2006) and were actually used for Bayesian NMF in Paisley et al. (2015). It is worth noting that the PAC-Bayesian approach provides sound theoretical foundations for this family of algorithms (Alquier et al., 2015). A different perspective is at work in this paper, we use efficient optimization algorithms to approximate the maximum *a posteriori* (MAP). The MAP is theoretically very efficient in some high-dimensional problems (see Abramovich and Lahav, 2015, in the context of additive regression). Its main computational advantage in the Bayesian NMF framework is its tractable form which allows to use for example alternating direction method of multipliers (ADMM) as in Boyd et al. (2011) or block coordinate descent (Bertsekas, 1999). We implement a blockwise gradient descent which can be interpreted as a Bayesian version of the algorithm presented in the aforecited Xu et al. (2012). Finally, we present numerical experiments on synthetic data, which are not meant to be exhaustive but rather serve to assess the behavior of Bayesian NMF and more specifically the impact of hyperparameters in the prior on the reconstruction of Y.

The paper is organized as follows. Notation for the Bayesian NMF model and the definition of our quasi-Bayesian estimator are given in Section 2. The oracle inequality is given in Section 3, its proof is postponed to Appendix A. We then discuss two algorithms: the Gibbs sampler and the blockwise gradient descent in Section 4. Section 5 contains our numerical experiments on synthetic data, and an application to assets prices is presented in Section 6.

2 Notation

For any $p \times q$ matrix A we denote by $A_{i,j}$ its (i,j)-th entry, $A_{i,\cdot}$ its i-th row and $A_{\cdot,j}$ its j-th 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 $\|A\|_F^2 = \langle A, A \rangle_F$. We let $A_{-i,\cdot}$ denote the matrix A where the i-th 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 i-th coordinate removed. We 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

$$Y = M + \mathcal{E}, \tag{1}$$

and we assume that $\mathcal{E}$ is random with $\mathbb{E}(\mathcal{E}) = 0$. The objective is to factorize M under the assumption that is has non-negative entries and can be reasonnably well approximated 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 both U and V have non-negative entries. Our theoretical analysis 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*

$$m(x) = \mathbb{E}\left[\varepsilon_{i,j} \mathbf{1}_{(\varepsilon_{i,j} \le x)}\right] \quad \text{and} \quad F(x) = \mathbb{P}(\varepsilon_{i,j} \le x),$$

assume that there exists a non-negative and bounded function g such that

$$\int_u^v m(x)\mathrm{d}x = \int_u^v g(x)\mathrm{d}F(x).$$

We put $\sigma^2 = \|g\|_\infty$.

The introduction of this rather technical condition is due to the technical analysis of our estimator which is based on Theorem 2 in Appendix A. Theorem 2 has first been proved by Dalalyan and Tsybakov (2007) using Stein's formula with a Gaussian noise. However, Dalalyan and Tsybakov (2008) have shown that C1 is actually sufficient to prove Theorem 2. For the sake of understanding, note that C1 is fulfilled when the noise is Gaussian ($\varepsilon_{i,j} \sim \mathcal{N}(0, s^2)$ and $\|g\|_\infty = s^2$) or uniform ($\varepsilon_{i,j} \sim \mathcal{U}[-b, b]$ and $\|g\|_\infty = b^2/2$).

2.2 Prior

We remind the idea of the factorization:

$$\underbrace{M}_{m_1\times m_2} = \underbrace{U}_{m_1\times K}\,\underbrace{V^\top}_{K\times m_2}.$$

We actually define a prior $\pi(U,V)$ for a fixed K. However, Section 5 suggests that our quasi-Bayesian estimator is adaptive, in the sense that if K is chosen much greater 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 for setting a large K prior to the analysis, say $K = m_1 \wedge m_2$. Yet this choice could be ruinous from a computational perspective (see the discussion Section 4). Anyhow, the following theoretical analysis only requires$1 \le K \le m_1 \wedge m_2$.

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

$$S_f := \int_0^\infty x^2 f(x)\mathrm{d}x < +\infty.$$

For any $\alpha, x > 0$, let

$$g_\alpha(x) := \frac{1}{\alpha} f\left(\frac{x}{\alpha}\right).$$

We define the prior on U and V by

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

where

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

With the notation $\gamma = (\gamma_1, \dots, \gamma_K)$, let us define the prior π by

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

and

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

The idea behind this prior is that under h, many γ_ℓ should be small and lead to non-significant 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. This is the meaning of the following assumption.

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

$$\int_0^{\varepsilon} h(x)\mathrm{d}x \geq \alpha \varepsilon^{\beta}.$$

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

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

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

C3 is crucial to understand the behavior of Bayesian NMF: the heavier the tails of $\tilde{f}(x)$, the better the performance of BNMF. However, efficient algorithms for ultra high-dimensional data usually require tractable forms for the prior that might not be compatible with this constraint.

We end this subsection with examples of functions f and h to keep in mind.

1. Exponential prior $f(x) = \exp(-x)$ with $\tilde{f} = f$, $\mathcal{C}_f = 1$ and $S_f = 2$.
2. Truncated Gaussian prior $f(x) \propto \exp(2ax - x^2)$ with $a \in \mathbb{R}$.
3. More generally, any prior from the exponential family.
4. Heavy-tailed prior $f(x) \propto \frac{1}{(1+x)^{\zeta}}$ with $\zeta > 1$.

A trade-off is at work here between statistical and computational properties. The heavy-tailed prior will lead to very nice theoretical results. Contrary to the exponential and truncated Gaussian prior which lead to estimators that are much easier to approximate through optimization algorithms.

For h, the inverse gamma prior $h(x) = \frac{b^a}{\Gamma(a)} \frac{1}{x^{a+1}} \exp\left(-\frac{b}{x}\right)$ is classical in the literature (see for example Salakhutdinov and Mnih, 2008; Alquier et al., 2014). However, other choices are possible, such as a gamma prior with appropriately tuned parameters. Note that h is less crucial than f from an algorithmic perspective. This point will be further discussed in Section 4.

2.3 Quasi-posterior and estimator

We define the quasi-likelihood as

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

for some fixed parameter $\lambda > 0$. Note that under the assumption that $\varepsilon_{i,j} \sim \mathcal{N}(0, 1/2\lambda)$, this would actually be the (true) likelihood up to a multiplicative

constant. As we pointed out, the use of quasi-likelihood to define quasi-posteriors is becoming rather popular in Bayesian statistics and machine learning literatures. We thus define the quasi-posterior

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

where

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

is a normalization constant. We are now ready to define our quasi-Bayesian estimator.

Definition 1. *The estimator is defined as*

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

In the sequel, we will study the theoretical (Section 3) and algorithmic (Section 4) properties of this estimator.

3 A sharp 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 shrinked to 0. The following set of matrices is introduced to formalize this idea.

Definition 2. *For any $r \in \{1,\dots,K\}$, let $\mathcal{M}_r$ be the set of pairs of matrices (U^0,V^0) with nonnegative entries such that*

$$U^0 = \begin{pmatrix} U^0_{11} & \dots & U^0_{1r} & 0 & \dots & 0\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ U^0_{m_1 1} & \dots & U^0_{m_1 r} & 0 & \dots & 0\end{pmatrix}$$

and

$$V^0 = \begin{pmatrix} V^0_{11} & \dots & V^0_{1r} & 0 & \dots & 0\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ V^0_{m_2 1} & \dots & V^0_{m_2 r} & 0 & \dots & 0\end{pmatrix}.$$

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), $U^0_{i,\ell}, V^0_{j,\ell} \le L$.

We are now in a position to state our main theorem. Its main message is that $\widehat{M}_\lambda$ is as close to M as would be an estimator designed with the actual knowledge of its rank, up to a small remainder. Note that the formula for the remainder is actually cumbersome, simplified versions are discussed below for the sake of clarity.

Theorem 1. *Fix $\lambda = \frac{1}{4\sigma^2}$ and assume that $\sigma^2 \leq 2K^{\frac{3}{2}}$. Under assumptions C1, C2 and C3,*

$$\begin{aligned}
\mathbb{E}\left(\|\widehat{M}_\lambda - M\|_F^2\right) \leq \inf_{1\leq r\leq K} \inf_{(U^0,V^0)\in\mathcal{M}_r} \Bigg\{ & \|U^0V^{0\top} - M\|_F^2 \\
& + 8\sigma^2(m_1\vee m_2)r\log\left(\sqrt{\frac{2(m_1\vee m_2)}{r}}\frac{\left(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\right)^2}{\sigma^2\mathcal{C}_f}\right) \\
& + 4\sigma^2 \sum_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\tilde{f}(U^0_{i\ell}+1)}\right) + 4\sigma^2 \sum_{\substack{1\leq j\leq m_2\\ 1\leq \ell\leq r}} \log\left(\frac{1}{\tilde{f}(V^0_{j\ell}+1)}\right) \\
& + 4\sigma^2\beta K\log\left(\frac{2S_f\sqrt{m_1m_2}\left(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr}\right)^2}{r\sigma^2}\right) \\
& + 8\sigma^2 r + 4\sigma^2 K\log\left(\frac{1}{\alpha}\right) + 4\sigma^2\log(4)\Bigg\}.
\end{aligned}$$

We remind the reader that the proof is given in Appendix A. It is based on a PAC-Bayesian theorem of Dalalyan and Tsybakov (2008). In order to explicit this result, we provide a weaker version, where we only compare $\widehat{M}_\lambda$ to the best factorization in $\mathcal{M}_r(L)$.

Corollary 1. *Fix $\lambda = \frac{1}{4\sigma^2}$ and assume that $\sigma^2 \leq 2K^{\frac{3}{2}}$. Under assumptions C1, C2 and C3,*

$$\begin{aligned}
\mathbb{E}\left(\|\widehat{M}_\lambda - M\|_F^2\right) \leq \inf_{1\leq r\leq K} \inf_{L>0} \inf_{(U^0,V^0)\in\mathcal{M}_r(L)} \Bigg\{ & \|U^0V^{0\top} - M\|_F^2 \\
& + 8\sigma^2(m_1\vee m_2)r\log\left(\frac{(2L+1)^2K\sqrt{2(m_1\vee m_2)^3}}{\tilde{f}(L+1)\sigma^2\mathcal{C}_f\sqrt{r}}\right) \\
& + 4\sigma^2\beta K\log\left(\frac{2(2L+1)^2S_f\sqrt{m_1m_2}(m_1\vee m_2)}{\sigma^2}\right) \\
& + 8\sigma^2 r + 4\sigma^2 K\log\left(\frac{1}{\alpha}\right) + 4\sigma^2\log(4)\Bigg\}.
\end{aligned}$$

First, note that when L is fixed, up to log terms, the order of the magnitude of the error bound is

$$\sigma^2(m_1 \vee m_2)r,$$

which is roughly the variance multiplied by the number of parameters to be estimated in any $(U^0, V^0) \in \mathcal{M}_r(L)$. If we know that M can actually be written $M = U_0 V_0^T$ for $(U^0, V^0) \in \mathcal{M}_r(L)$ with a small L, this rate is very nice. Alternatively, whenever L is large, the log parts in

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

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

$$8\sigma^2(m_1 \vee m_2)rL^2$$

which is terrible if L is large. On the contrary, with the heavy-tailed prior $f(x) \propto (1+x)^{-\zeta}$ (as in Dalalyan and Tsybakov, 2008), the leading term is

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

which is way more satisfactory. Although this comes at the price of more demanding computations, as highlighted in the next section.

4 Gibbs sampler and optimization algorithms

This section is devoted to two implementations of our quasi-Bayesian estimator, both with their advantages. We remind the quasi-posterior formula

$$\begin{aligned}\widehat{\rho}_\lambda(U,V,\gamma) &= \frac{1}{Z}\widehat{L}(U,V)\pi(U,V,\gamma) \\ &= \frac{1}{Z}\exp\left(-\lambda\|Y - UV^\top\|_{\mathrm{F}}^2\right)\prod_{\ell=1}^{K}\left[h(\gamma_\ell)\prod_{i=1}^{m_1} g_{\gamma_\ell}(U_{i\ell})\prod_{j=1}^{m_2} g_{\gamma_\ell}(V_{j\ell})\right].\end{aligned}$$

We first derive in Section 4.1 a general form for the conditional posteriors $\widehat{\rho}_\lambda(U_{i,\cdot}|U_{-i,\cdot}, V, \gamma, Y)$, $\widehat{\rho}_\lambda(V_{j,\cdot}|U, V_{-j,\cdot}, \gamma, Y)$ and $\widehat{\rho}_\lambda(\gamma_\ell|U, V, \gamma_{-\ell}, Y)$ that are needed to implement the Gibbs sampler. These conditional posteriors are explicitly given on two examples in Section 4.2 and Section 4.3. Secondly, we propose an alternative implementation in Section 4.4 using a block coordinate descent approach.

4.1 General form of the conditional posteriors

As a function of $U_{i,\cdot}$,

$$\begin{aligned}\widehat{L}(U,V)\pi(U,V,\gamma) &\propto \exp\left(-\lambda\|Y-UV^\top\|_{\mathrm{F}}^2\right)\prod_{\ell=1}^{K} g_{\gamma_\ell}(U_{i,\ell}) \\ &\propto \exp\left(-\lambda\|Y_{i,\cdot}-U_{i,\cdot}V^\top\|^2\right)\prod_{\ell=1}^{K} g_{\gamma_\ell}(U_{i,\ell}).\end{aligned}$$

Let $\widehat{U}_i = Y_{i,\cdot}V(V^TV)^{-1}$ and $\Sigma_U = (V^TV)^{-1}$. This yields

$$\begin{aligned}\widehat{\rho}_\lambda(U_{i,\cdot}|U_{-i,\cdot},V,\gamma,Y) &= \widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y) \\ &\propto \exp\left(-\lambda(\widehat{U}_i-U_{i,\cdot})(\Sigma_U)^{-1}(\widehat{U}_i-U_{i,\cdot})^T\right)\prod_{\ell=1}^{K} g_{\gamma_\ell}(U_{i,\ell}).\end{aligned}$$

In the same way, we define $\widehat{V}_j = Y_{\cdot,j}^T U(U^TU)^{-1}$ and $\Sigma_V = (U^TU)^{-1}$ and we have

$$\begin{aligned}\widehat{\rho}_\lambda(V_{j,\cdot}|V_{-j,\cdot},U,\gamma,Y) &= \widehat{\rho}_\lambda(V_{j,\cdot}|U,\gamma,Y) \\ &\propto \exp\left(-\lambda(\widehat{V}_j-V_{j,\cdot})(\Sigma_V)^{-1}(\widehat{V}_j-V_{j,\cdot})^T\right)\prod_{\ell=1}^{K} g_{\gamma_\ell}(V_{j,\ell})\end{aligned}$$

Finally,

$$\begin{aligned}\widehat{\rho}_\lambda(\gamma_\ell|U,V,\gamma_{-\ell},Y) &= \widehat{\rho}_\lambda(\gamma_\ell|U,V,Y) \\ &\propto h(\gamma_\ell)\prod_{i=1}^{m_1} g_{\gamma_\ell}(U_{i\ell})\prod_{j=1}^{m_2} g_{\gamma_\ell}(V_{j\ell}).\end{aligned}$$

The Gibbs sampler (described in its general form in Bishop, 2006, 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 all generality, sampling from $\widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y)$ might require considerable effort. For example, for the heavy-tailed prior $f(x)=\frac{1}{(1+x)^\zeta}$,

$$\widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y)\propto \exp\left(-\lambda(\widehat{U}_i-U_{i,\cdot})(\Sigma_U)^{-1}(\widehat{U}_i-U_{i,\cdot})^T\right)\frac{1}{\prod_{\ell=1}^K(1+U_{i,\ell})^\zeta},$$

which is not a standard distribution. In such cases, a Metropolis-within-Gibbs approach is often the best choice (as advocated in Guedj and Alquier, 2013; Guedj and Robbiano, 2015) and exhibits nice properties, sadly at the cost of quite substantial computational power. In the ultra high-dimensional context of NMF, this choice appeared unrealistic to us and we promote in Section 4.2 and Section 4.3 the idea that choosing specific priors yields known distributions as quasi-posteriors.

4.2 Gibbs Sampler with an exponential prior f

Here $f(x)=\exp(-x)$ and $g_\alpha(x)=\frac{1}{\alpha}\exp\left(-\frac{x}{\alpha}\right)$. So

$$\begin{aligned}
&\widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y)\\
&\quad\propto \exp\left(-\lambda(\widehat{U}_i-U_{i,\cdot})(\Sigma_U)^{-1}(\widehat{U}_i-U_{i,\cdot})^T-\sum_{\ell=1}^K\frac{U_{i,\ell}}{\gamma_\ell}\right)\\
&\quad=\exp\left(-\lambda(\widehat{U}_i-U_{i,\cdot})(\Sigma_U)^{-1}(\widehat{U}_i-U_{i,\cdot})^T-U_i\gamma^{-1}\right)\\
&\quad=\exp\left(-\lambda\left[\left(\widehat{U}_i-\frac{1}{2\lambda}\Sigma_U\gamma^{-1}\right)-U_{i,\cdot}\right](\Sigma_U)^{-1}\left[\left(\widehat{U}_i-\frac{1}{2\lambda}\Sigma_U\gamma^{-1}\right)-U_{i,\cdot}\right]^T\right),
\end{aligned}$$

where we use the abusive notation $\gamma^{-1}=(1/\gamma_1,\dots,1/\gamma_K)$. So $\widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y)$ amounts to a truncated Gaussian distribution

$$\mathcal{N}\left(\widehat{U}_i-\frac{1}{2\lambda}\Sigma_U\gamma^{-1},\frac{2}{\lambda}\Sigma_U\right)\mathbf{1}_{\mathbb{R}_+^K}$$

restricted to vectors with non-negative entries. Sampling from it can be done using the R package *tmvtnorm* from Wilhelm (2015) (as in Mai and Alquier, 2015). Computation of $\widehat{\rho}_\lambda(V_{j,\cdot}|U,\gamma,Y)$ is similar.

Note that

$$\begin{aligned}
\widehat{\rho}_\lambda(\gamma_\ell|U,V,\gamma_{-\ell},Y)&\propto h(\gamma_\ell)\prod_{i=1}^{m_1}g_{\gamma_\ell}(U_{i\ell})\prod_{j=1}^{m_2}g_{\gamma_\ell}(V_{j\ell})\\
&=h(\gamma_\ell)\gamma_\ell^{-(m_1+m_2)}\exp\left(-\frac{\sum_{i=1}^{m_1}U_{i,\ell}+\sum_{j=1}^{m_2}V_{j,\ell}}{\gamma_\ell}\right),
\end{aligned}$$

providing an incentive to consider the inverse gamma prior $\mathcal{IG}(a,b)$, *i.e.*, $h(x) = \frac{b^a}{\Gamma(a)} x^{-a+1}\exp(-b/x)$. This leads to the conditional quasi-posterior

$$\mathcal{IG}\left(a + m_1 + m_2, b + \sum_{i=1}^{m_1} U_{i,\ell} + \sum_{j=1}^{m_2} V_{j,\ell}\right),$$

which is a classical choice in the Bayesian literature (see Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Lawrence and Urtasun, 2009; Zhou et al., 2010; Alquier et al., 2014). However, as pointed out in Alquier et al. (2014), another conjugate choice is the gamma prior $\Gamma(a,b)$ for $a = m_1 + m_2 - 1/2$. Actually, when $h(x) = \frac{b^a}{\Gamma(a)} x^{a-1}\exp(-bx)$,

$$\widehat{\rho}_\lambda(\gamma_\ell | U, V, \gamma_{-\ell}, Y) \propto \gamma_\ell^{-(m_1+m_2)+a-1} \exp\left(-\frac{\sum_{i=1}^{m_1} U_{i,\ell} + \sum_{j=1}^{m_2} V_{j,\ell}}{\gamma_\ell} - b\gamma_\ell\right).$$

Thus, choosing the prior $\Gamma(m_1 + m_2 - 1/2, b)$ yields the conditional quasi-posterior

$$\mathcal{IG}\left(\sqrt{\frac{\sum_{i=1}^{m_1} U_{i,\ell} + \sum_{j=1}^{m_2} V_{j,\ell}}{b}}, 2\left(\sum_{i=1}^{m_1} U_{i,\ell} + \sum_{j=1}^{m_2} V_{j,\ell}\right)\right),$$

where $\mathcal{IG}(\mu,\nu)$ denotes the inverse Gaussian distribution, whose density is proportional to $x^{-3/2}\exp\left(-\frac{\nu}{2}\left(\frac{x}{\mu^2} + \frac{1}{x}\right)\right)$. Alquier et al. (2014) contains numerical experiments to assess that this prior is less sensitive than the inverse gamma prior to a misspecification of K.

4.3 Gibbs sampler with a truncated Gaussian prior f

Here, $f(x) \propto \exp(2ax - x^2)$. So

$$\begin{aligned}
&\widehat{\rho}_\lambda(U_{i,\cdot} | V, \gamma, Y) \\
&\quad \propto \exp\left(-\lambda(\widehat{U}_i - U_{i,\cdot})(\Sigma_U)^{-1}(\widehat{U}_i - U_{i,\cdot})^T + 2a\sum_{\ell=1}^{K}\frac{U_{i,\ell}}{\gamma_\ell} - \sum_{\ell=1}^{K}\frac{U_{i,\ell}^2}{\gamma_\ell^2}\right) \\
&\quad = \exp\left(-\lambda(\widehat{U}_i - U_{i,\cdot})(\Sigma_U)^{-1}(\widehat{U}_i - U_{i,\cdot})^T + 2aU_{i,\cdot}\gamma^{-1} - U_{i,\cdot}\mathrm{Diag}(\gamma)^{-2}U_{i,\cdot}^T\right),
\end{aligned}$$

which is the density of the truncated Gaussian distribution

$$\mathcal{N}\left(\left(\frac{1}{\lambda}\Sigma_U + \mathrm{Diag}(\gamma)^2\right)\left(a\gamma^{-1} + \lambda\Sigma_U^{-1}\widehat{U}_i^\top\right), 2\left(\frac{1}{\lambda}\Sigma_U + \mathrm{Diag}(\gamma)^2\right)\right)\mathbf{1}_{\mathbb{R}_+^K}.$$

Computation of $\widehat{\rho}_\lambda(V_{j,\cdot} | U, \gamma, Y)$ is similar.

Next,

$$\widehat{\rho}_\lambda(\gamma_\ell|U,V,Y) \propto h(\gamma_\ell)\gamma_\ell^{-(m_1+m_2)}\exp\left(2a\frac{\sum_{i=1}^{m_1}U_{i,\ell}+\sum_{j=1}^{m_2}V_{j,\ell}}{\gamma_\ell}-\frac{\sum_{i=1}^{m_1}U_{i,\ell}^2+\sum_{j=1}^{m_2}V_{j,\ell}^2}{\gamma_\ell^2}\right).$$

Clearly, in all generality we cannot hope to recover an inverse gamma nor an inverse Gaussian distribution. However, when $a=0$,

$$\widehat{\rho}_\lambda(\gamma_\ell|U,V,Y)\propto h(\gamma_\ell)\left(\gamma_\ell^2\right)^{-\frac{m_1+m_2}{2}}\exp\left(-\frac{\sum_{i=1}^{m_1}U_{i,\ell}^2+\sum_{j=1}^{m_2}V_{j,\ell}^2}{\gamma_\ell^2}\right),$$

and in that case, considering a more convenient prior $\mathcal{IG}(a,b)$ for γ_ℓ^2 instead of γ_ℓ, we have

$$\gamma_\ell^2|U,V,Y\sim\mathcal{IG}\left(a+\frac{m_1+m_2}{2},\sum_{i=1}^{m_1}U_{i,\ell}^2+\sum_{j=1}^{m_2}V_{j,\ell}^2\right).$$

Alternatively, with the prior $\gamma_\ell^2\sim\Gamma((m_1+m_2-1)/2,b)$,

$$\gamma_\ell^2|U,V,Y\sim\mathcal{IG}\left(\sqrt{\frac{\sum_{i=1}^{m_1}U_{i,\ell}^2+\sum_{j=1}^{m_2}V_{j,\ell}^2}{b}},2\left[\sum_{i=1}^{m_1}U_{i,\ell}^2+\sum_{j=1}^{m_2}V_{j,\ell}^2\right]\right),$$

allowing for efficient sampling of our quasi-Bayesian estimator.

4.4 Optimization through block coordinate descent

In this section, we discuss the implementation of the MAP estimator

$$\begin{aligned}(\widetilde{U}_\lambda,\widetilde{V}_\lambda,\widetilde{\gamma}_\lambda) &= \arg\max_{U,V,\gamma}\ \widehat{\rho}_\lambda(U,V,\gamma)\\ &= \arg\min_{U,V,\gamma}\Bigg\{\lambda\|Y-UV^\top\|_F^2-\log\pi(U,V,\gamma)\\ &\qquad -\sum_{i=1}^{m_1}\sum_{\ell=1}^{K}\log\left(g_{\gamma_\ell}(U_{i,\ell})\right)-\sum_{j=1}^{m_2}\sum_{\ell=1}^{K}\log\left(g_{\gamma_\ell}(V_{j,\ell})\right)-\sum_{\ell=1}^{K}\log\left(h(\gamma_\ell)\right)\Bigg\}.\end{aligned}$$

Note that the minimization problem

$$\min_{U,V}\ \|Y-UV^\top\|_F^2$$

is already known as a tough one: while it is individually convex in U and V, it is not convex as a function of the pair (U,V). Such a task is know as a biconvex problem (see Boyd et al., 2011, section 9.2), for which an iterative algorithm may get stuck in a local minimum. However, the block coordinate descent approach (Bertsekas, 1999) is used in practice with reasonnable results. This algorithm seems to be relatively standard in (non-Bayesian) NMF and is described in Algorithm 2.

Algorithm 2 Pseudo-algorithm for block coordinate descent.

Input Y, λ.

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

While not converged, $k := k+1$:

$$U^{(k)} := \arg\min_{U} \left\{ \lambda \|Y - U(V^{(k-1)})^\top\|_F^2 - \sum_{i=1}^{m_1}\sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(U_{i,\ell})] \right\}$$

$$V^{(k)} := \arg\min_{V} \left\{ \lambda \|Y - U^{(k)}V^\top\|_F^2 - \sum_{j=1}^{m_2}\sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(V_{j,\ell})] \right\}$$

$$\gamma^{(k)} := \arg\min_{\gamma} \sum_{\ell=1}^{K} \left\{ -\sum_{i=1}^{m_1} \log[g_{\gamma_\ell}(U_{i,\ell}^{(k)})] - \sum_{j=1}^{m_2} \log[g_{\gamma_\ell}(V_{j,\ell}^{(k)})] - \log[h(\gamma_\ell)] \right\}$$

Several remarks are in order here. In this case, another popular algorithm, the Alternating Direction Method of Multipliers (ADMM) (reviewed in Boyd et al., 2011), takes a similar form. When the functions f and h are not conveniently chosen, this optimization problem can be very cumbersome, if tractable at all.

If one picks f and h as in Section 4.2 or Section 4.3, the optimization problems in U and V are quadratic problems with a non-negativity constraint. Some algorithms are available for this task in the linear regression setting (Bro and De Jong, 1997, for example). For NMF, we can either replace each optimization step by a single gradient descent step (as in Lin, 2007) or by a full gradient descent. In each case, after the gradient step, we have to project U and V on the set of matrices with non-negative entries. Let P denote this projection, which replaces any negative entry by 0.

In the sequel, we derive explicit forms of Algorithm 2 when the priors are chosen as in Section 4.2 and Section 4.3. We optimize with respect to U and V by projected gradient descent. Since the modes of $\mathcal{IG}(\alpha,\beta)$ and of $\mathcal{IG}(\mu,\nu)$

are known to be respectively

$$\frac{\beta}{\alpha+1} \quad \text{and} \quad \mu\left(\left(1+\frac{9\mu^2}{4\nu^2}\right)^{1/2}-\frac{3\mu}{2\nu}\right),$$

we provide an explicit optimization with respect to γ. The algorithm for the exponential prior is detailed in Algorithm 3, whereas Algorithm 4 is adapted to the truncated Gaussian prior. This ends this section devoted to practical implementation of our quasi-Bayesian estimator.

Algorithm 3 Block coordinate descent - exponential prior for U, V.

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$ and a decreasing sequence (α_ℓ).

While not converged, $k = k+1$:

$\ell := 0$ and $U^{(k,0)} := U^{(k-1)}$

While not converged, $\ell = \ell+1$:

$$U^{(k,\ell)} := P\left(U^{(k,\ell-1)} + \alpha_\ell\left(2\lambda[Y - U^{(k,\ell-1)}(V^{(k-1)})^\top]V^{(k-1)} - (\gamma^{(k-1)})^{-1}\right)\right)$$

$U^{(k)} := U^{(k,\ell)}$

$\ell := 0$ and $V^{(k,0)} := V^{(k-1)}$

While not converged, $\ell = \ell+1$:

$$V^{(k,\ell)} := P\left(V^{(k,\ell-1)} + \alpha_\ell\left(2\lambda[Y^\top - V^{(k,\ell-1)}U^{(k)\top}]U^{(k)} - (\gamma^{(k-1)})^{-1}\right)\right)$$

$V^{(k)} := V^{(k,\ell)}$

For $\ell = 1, \ldots, K$:

If Inverse gamma prior $\mathcal{IG}(a,b)$, $\gamma_\ell^{(k)} := \frac{b+\sum_{i=1}^{m_1} U_{i,\ell}^{(k)} + \sum_{j=1}^{m_2} V_{j,\ell}^{(k)}}{a+m_1+m_2+1}$

If Gamma prior $\Gamma(m_1+m_2-1/2, b)$, $\gamma_\ell^{(k)} := \sqrt{\frac{\sum_{i=1}^{m_1} U_{i,\ell}^{(k)} + \sum_{j=1}^{m_2} V_{j,\ell}^{(k)}}{b}} \times$
$$\left\{\sqrt{1+\frac{9}{16b\left(\sum_{i=1}^{m_1} U_{i,\ell}^{(k)} + \sum_{j=1}^{m_2} V_{j,\ell}^{(k)}\right)}} - \frac{3}{4\sqrt{b\left(\sum_{i=1}^{m_1} U_{i,\ell}^{(k)} + \sum_{j=1}^{m_2} V_{j,\ell}^{(k)}\right)}}\right\}$$

Algorithm 4 Block coordinate descent - truncated Gaussian prior for U, V.

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$ and a decreasing sequence (α_ℓ).

While not converged, $k = k+1$:

$\ell := 0$ and $U^{(k,0)} := U^{(k-1)}$

While not converged, $\ell = \ell + 1$:

$$U^{(k,\ell)} := P\left(U^{(k,\ell-1)} + \alpha_k \left(2\lambda[Y - U^{(k,\ell-1)}(V^{(k-1)})^\top]V^{(k)} \right.\right.$$
$$\left.\left. +2aU^{(k,\ell-1)}(\gamma^{(k)})^{-1} - U^{(k,\ell-1)}\mathrm{Diag}\left(\gamma^{(k)}\right)^{-2} U^{(k,\ell-1)\top}\right)\right)$$

$U^{(k)} := U^{(k,\ell)}$

$\ell := 0$ and $V^{(k,0)} := V^{(k-1)}$

While not converged, $\ell = \ell + 1$:

$$V^{(k,\ell)} := P\left(U^{(k,\ell-1)} + \alpha_k \left(2\lambda[Y^\top - V^{(k,\ell-1)}U^{(k)\top}]U^{(k)} \right.\right.$$
$$\left.\left. +2aV^{(k,\ell-1)}(\gamma^{(k)})^{-1} - V^{(k,\ell-1)}\mathrm{Diag}\left(\gamma^{(k)}\right)^{-2} V^{(k,\ell-1)\top}\right)\right)$$

$V^{(k)} := V^{(k,\ell)}$

For $\ell = 1, \ldots, K$:

If Inverse gamma prior $\mathcal{IG}(a,b)$,

$$\gamma_\ell^{(k)} := \frac{b + \sum_{i=1}^{m_1}\left(U_{i,\ell}^{(k)}\right)^2 + \sum_{j=1}^{m_2}\left(V_{j,\ell}^{(k)}\right)^2}{a + \frac{m_1+m_2}{2} + 1}$$

If Gamma prior $\Gamma(m_1+m_2-1/2, b)$, $\gamma_\ell^{(k)} := \sqrt{\frac{\sum_{i=1}^{m_1}\left(U_{i,\ell}^{(k)}\right)^2 + \sum_{j=1}^{m_2}\left(V_{j,\ell}^{(k)}\right)^2}{b}} \times$
$$\left\{\sqrt{1 + \frac{9}{16b\left(\sum_{i=1}^{m_1}\left(U_{i,\ell}^{(k)}\right)^2 + \sum_{j=1}^{m_2}\left(V_{j,\ell}^{(k)}\right)^2\right)}} - \frac{3}{4\sqrt{b\left(\sum_{i=1}^{m_1}\left(U_{i,\ell}^{(k)}\right)^2 + \sum_{j=1}^{m_2}\left(V_{j,\ell}^{(k)}\right)^2\right)}}\right\}$$

5 Numerical experiments

NMF has attracted a great deal of interest and the number of available algorithms is massive. Therefore, instead of embarking on lengthy comparisons with a subjective set of competitors, we focus in this section on the behavior of Bayesian NMF on synthetic data, especially the influence of the choice of the prior hyperparameters. To do so, let us consider the Bayesian MAP estimator coupled with the exponential prior $\mathcal{E}(1)$ on (U,V) and with the gamma prior $\Gamma(m_1+m_1-1/2,b)$ on the coefficients γ_j.

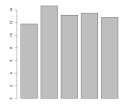
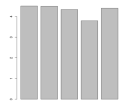
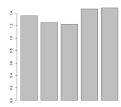
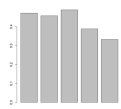
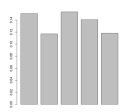
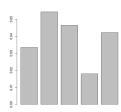
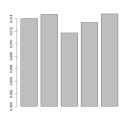
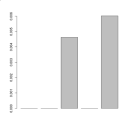
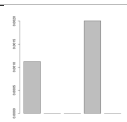
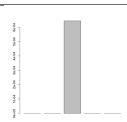
To assess the impact of the hyperparameter b on the quality of the reconstruction of the matrix M, let us consider the following exponential grid: $b \in \{10^0, 10^1, \ldots, 10^9\}$. We report in Figure 1 the performance of the reconstruction of M in terms of the mean square error: $\text{MSE} = \frac{1}{m_1 m_2}\|M - \widehat{M}_\lambda\|_F^2$. It is also worth mentioning the possibility to infer the rank of M. In theory, note that our estimator will report a full rank estimator $\widehat{M}_\lambda$. However in practice, we can expect that many $\widehat{\gamma}_\ell$'s will be close enough to 0 so that thresholding their values will not affect the performance of the reconstruction. As highlighted below, this will typically occur for large values of b.

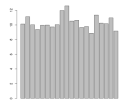
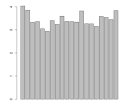
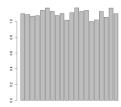
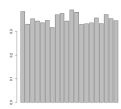
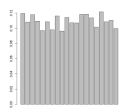
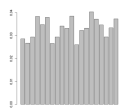
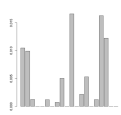
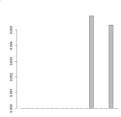
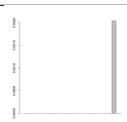
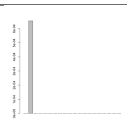
In a first experiment, we simulate $Y = M + \mathcal{E} = UV^T + \mathcal{E}$ with $m_1 = m_2 = 100$ and U, V two 100×2 matrices with entries drawn independently from a $\mathcal{U}([0,3])$ distribution. We also choose $K = 5$, which allows a representation of γ for each experiment. We simulate the entries $\mathcal{E}_{i,j}$ of $\mathcal{E}$ independenly from a $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 = 0.01$. The results are reported in Figure 1 (left column).

Clearly, for any value of b between 1 and 10^6 the MSE's are of similar magnitude (between 0.0007 and 0.002). While this is satisfactory, our method fails to identify a rank 2 matrix and delivers a full rank (5) estimate. For larger b, the MSE tends to increase yet a rank 2 matrix is identified. This remark is in accordance with the fact that model identification and estimation objectives are often incompatible in high-dimensional statistics (as pointed out Yang, 2005, among others). The overall good results are obviously related to the rather strong prior knowledge $K = 5$: we repeated the same design with the more relaxed assumption $K = 20$ (reported in Figure 1, right column).

From these results, we gather that too small values for b possibly lead to overfitting. The optimal value in terms of MSE seem around $b = 10^6$. About the estimation of the actual rank, the vector γ is sparse for $b = 10^6$ but we only identify the actual rank when $b = 10^7$. The rank identification problem seems more sensitive to a proper tuning of b than the estimation problem, *i.e.*, minimization of the MSE.

Figure 1: Results of the first (left) and second (right) experiments with $K = 5$ and $K = 20$, respectively.

b	MSE	$\widehat{\gamma}$
1	0.0007617591	
10	0.0027504950	
10^2	0.0010887216	
10^3	0.0007118330	
10^4	0.0007736950	
10^5	0.0007109982	
10^6	0.0024563021	
10^7	0.0099288813	
10^8	0.1465389771	
10^9	0.6329240955	

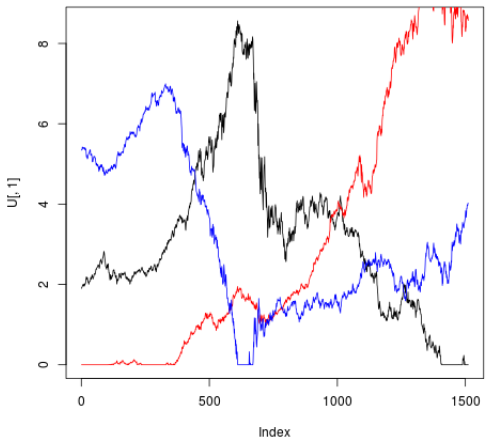
b	MSE	$\widehat{\gamma}$
1	0.013176016	
10	0.010169835	
10^2	0.009516530	
10^3	0.009934154	
10^4	0.006730481	
10^5	0.006475302	
10^6	0.001571762	
10^7	0.007313480	
10^8	0.607692862	
10^9	0.640797126	

6 Application to assets prices

We finally test our method on a real dataset, consisting in the daily prices of 993 financial assets between 2006-01-03 and 2011-12-30. It is available in the R package *pprobedata*[1].

What can be expected at first sight is that assets related to similar markets will behave similarly. It might thus be possible to derive some common trends for these assets. We expect that NMF will extract underlying trends $U_{\cdot,\ell}$ and approximate the price of each asset as a mixture of these trends. We use the MAP estimator with $b = 10^7$, $K = 3$. Figure 2 presents these underlying trends $U_{\cdot,1}$, $U_{\cdot,2}$ and $U_{\cdot,3}$, with a clear separation in three regimes, where each of the asset is dominant. As an example, we illustrate the (good) reconstruction of the prices of assets 1 and 4 in Figure 3.

Figure 2: The three trends $U_{\cdot,1}$, $U_{\cdot,2}$ and $U_{\cdot,3}$ obtained on the assets prices dataset.

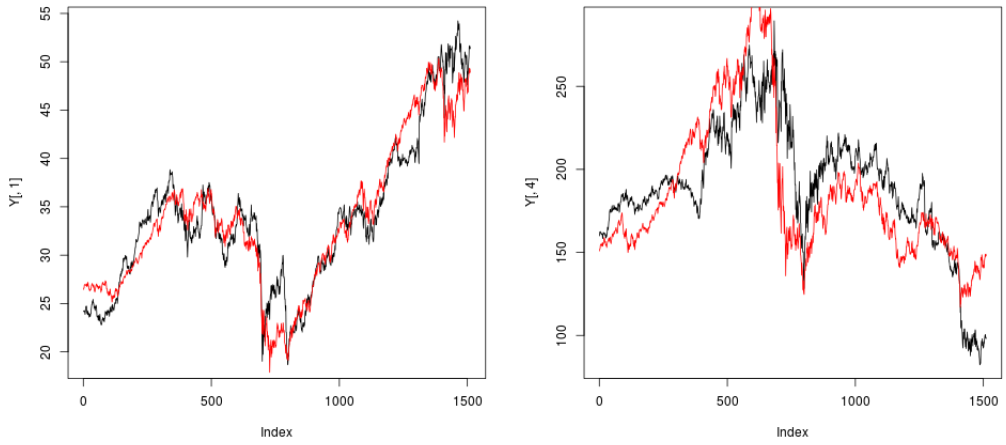


Acknowledgements

The authors would like to thank Jialia Mei and Yohann de Castro (Université Paris-Sud) for insightful discussions and for providing many references on NMF.

[1] http://www.portfolioprobe.com/R

Figure 3: Prices of assets 1 (left, black) and 4 (right, black), and reconstruction of this series as a linear combination of $U_{\cdot,1}$, $U_{\cdot,2}$ and $U_{\cdot,3}$ (red).



References

F. Abramovich and T. Lahav. Sparse additive regression on a regular lattice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):443–459, 2015. 3

G. I. Allen, L. Grosenick, and J. Taylor. A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505): 145–159, 2014. doi: 10.1080/01621459.2013.852978. 2

P. Alquier. Bayesian methods for low-rank matrix estimation: short survey and theoretical study. In *Algorithmic Learning Theory 2013*, pages 309–323. Springer, 2013. 2

P. Alquier, V. Cottet, N. Chopin, and J. Rousseau. Bayesian matrix completion: prior specification. Preprint arXiv:1406.1440, 2014. 2, 6, 12

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. Preprint arXiv:1506.04091, 2015. 3

S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012. 2

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999. 3, 14

C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 10. Springer, 2006. 3, 10

P. Bissiri, C. Holmes, and S. Walker. A general framework for updating belief distributions. Preprint arXiv:1306.6430, 2013. 3

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. 3, 14

R. Bro and S. De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, 11(5):393–401, 1997. 14

O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint Laboratoire de Probabilités et Modèles Aléatoires, PMA-840, 2003. 3, 24

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004. 3, 24

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007. 3, 24

J. Corander and M. Villani. Bayesian assessment of dimensionality in reduced rank regression. *Statistica Neerlandica*, 58:255–270, 2004. 2

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008. 1, 3, 4, 8, 9, 24, 25

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In N. Bshouty and C. Gentile, editors, *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 97–111. Springer Berlin Heidelberg, 2007. 4

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, 2003. 2

I. Giulini. PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. Preprint arXiv:1511.06263, 2015. 3

N. Guan, D. Tao, Z. Luo, and B. Yuan. NeNMF: an optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012. 2

B. Guedj and P. Alquier. PAC-Bayesian Estimation and Prevision in Sparse Additive Models. *Electronic Journal of Statistics*, 7:264–291, 2013. 3, 11

B. Guedj and S. Robbiano. PAC-Bayesian High Dimensional Bipartite Ranking. Preprint arXiv:1511.02729, 2015. 3, 11

D. Guillamet and J. Vitria. Classifying faces with nonnegative matrix factorization. In *Proc. 5th Catalan conference for artificial intelligence*, pages 24–31, 2002. 2

M. I. Jordan, Z. Ghahrapani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233, 1999. 3

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. 2

N. D. Lawrence and R. Urtasun. Non-linear matrix factorization with gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 601–608. ACM, 2009. 2, 12

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 2

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001. 2

G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006. 24

Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, volume 7, pages 15–21, 2007. 2, 12

C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007. 2, 14

D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2002. 3

T. T. Mai and P. Alquier. A Bayesian approach for matrix completion: optimal rates under general sampling distributions. *Electronic Journal of Statistics*, 9:823–841, 2015. 2, 11

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York, 1998. ACM. 3, 24

A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010. 2

J. Paisley, D. Blei, and M. I. Jordan. *Bayesian nonnegative matrix factorization with stochastic variational inference*, volume Handbook of Mixed Membership Models and Their Applications, chapter 11. Chapman and Hall/CRC, 2015. 2, 3

B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pages 1214–1222, 2012. 2

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008. 1, 2, 6, 12

F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006. 2

J. Shawe-Taylor and R. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, New York, 1997. ACM. 3, 24

T. Suzuki. Convergence rate of Bayesian tensor estimatior and its minimax optimality. In *Proceedings of the 32nd International Conference on Machine Learning (Lille, 2015)*, pages 1273–1282, 2015. 2

S. Wilhelm. *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*, 2015. URL `http://CRAN.R-project.org/package=tmvtnorm`. R package version 1.4-10. 11

W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*, pages 267–273. ACM, 2003. 2

Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013. 2

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. 1, 2, 3

Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005. 17

M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin. Nonparametric Bayesian Matrix Completion. In *Proc. IEEE SAM*, 2010. 2, 12

A Proofs

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

A.1 A PAC-Bayesian bound from Dalalyan and Tsybakov (2008)

The analysis of quasi-Bayesian estimators with PAC bounds started with Shawe-Taylor and Williamson (1997). McAllester improved on the initial method and introduced the name "PAC-Bayesian bounds" (McAllester, 1998). Catoni also improved these results to derive sharp oracle inequalities (Catoni, 2003, 2004, 2007). Dalalyan and Tsybakov (2008) proved a different PAC-Bayesian bound based on the idea of unbiased risk estimation (see Leung and Barron, 2006). We first recall its form in the context of matrix factorization.

Theorem 2. *Under C1, as soon as* $\lambda \leq \frac{1}{4\sigma^2}$,

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

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 Dalalyan and Tsybakov (2008), stated for vectors, is still valid for matrices.

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

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

This is done in Section A.2. We simply upper bound the infimum over all ρ by the infimum over this parametric family. So, we only have to calculate, or upper-bound

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

and

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

This is done in two lemmas in Section A.3 and Section A.4 respectively. We finally gather all the pieces together in Section A.5, and optimize with respect to c.

A.2 A parametric family of factorizations

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

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

A.3 Upper bound for the integral part

Lemma A.1.

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

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

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

$$= c\left(\|U^0\|_F + \|V^0\|_F + c\right).$$

Now, let Π be the orthogonal projection on the set

$$\left\{M^0 \colon \|M^0 - U^0 V^{0\top}\|_F \le c\left(\|U^0\|_F + \|V^0\|_F + c\right)\right\}$$

with respect to the Frobenius norm. Note that

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

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

A.4 Upper bound for the Kullback-Leibler divergence

Lemma A.2. *Under C2 and C3,*

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

Proof. By definition

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

Then, note that

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

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

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

Then

$$\int \left(\int \mathbf{1}_{\{\|U-U^0\|_F \le c\}} \pi(U|\gamma) \mathrm{d}U \right) \pi(\gamma) \mathrm{d}\gamma \ge \int_E \underbrace{\left(\int \mathbf{1}_{\{\|U-U^0\|_F \le c\}} \pi(U|\gamma) \mathrm{d}U \right)}_{=:I_3} \pi(\gamma) \mathrm{d}\gamma$$

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

$$\begin{aligned}
I_3 &= \pi\left(\sum_{\substack{1\le i\le m_1 \\ 1\le \ell \le K}} (U_{i,\ell} - U^0_{i,\ell})^2 \le c^2 \middle| \gamma \right) \\
&= \pi\left(\sum_{\substack{1\le i\le m_1 \\ 1\le \ell \le r}} (U_{i,\ell} - U^0_{i,\ell})^2 + \sum_{\substack{1\le i\le m_1 \\ r+1\le \ell \le K}} U^2_{i,\ell} \le c^2 \middle| \gamma \right) \\
&\ge \pi\left(\sum_{\substack{1\le i\le m_1 \\ r+1\le \ell \le K}} U^2_{i,\ell} \le \frac{c^2}{2} \middle| \gamma \right) \pi\left(\sum_{\substack{1\le i\le m_1 \\ 1\le \ell \le r}} (U_{i,\ell} - U^0_{i,\ell})^2 \le \frac{c^2}{2} \middle| \gamma \right) \\
&\ge \underbrace{\pi\left(\sum_{\substack{1\le i\le m_1 \\ r+1\le \ell \le K}} U^2_{i,\ell} \le \frac{c^2}{2} \middle| \gamma \right)}_{=:I_4} \prod_{\substack{1\le i\le m_1 \\ 1\le \ell \le r}} \pi\left((U_{i,\ell} - U^0_{i,\ell})^2 \le \frac{c^2}{2m_1 r} \middle| \gamma \right).
\end{aligned}$$

Now, using Markov's inequality,

$$\begin{aligned}
1 - I_4 &= \pi\left(\sum_{\substack{1\le i\le m_1 \\ r+1\le \ell \le K}} U^2_{i,\ell} \ge \frac{c^2}{2} \middle| \gamma \right) \\
&\le 2 \frac{\mathbb{E}_\pi\left(\sum_{\substack{1\le i\le m_1 \\ r+1\le \ell \le K}} U^2_{i,\ell} \middle| \gamma \right)}{c^2} \\
&= 2 \frac{\sum_{\substack{1\le i\le m_1 \\ r+1\le \ell \le K}} \gamma_j^2 S_f^2}{c^2} \\
&\le \frac{1}{2},
\end{aligned}$$

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

$$\begin{aligned}\pi\left(\left(U_{i,\ell}-U_{i,\ell}^0\right)^2 \leq \frac{c^2}{2m_1 r}\middle|\gamma\right) &\geq \int_{U_{i,\ell}^0}^{U_{i,\ell}^0+\frac{c}{\sqrt{2m_1 r}}} \frac{1}{\gamma_j} f\left(\frac{u}{\gamma_j}\right)\mathrm{d}u \\ &\geq \int_{U_{i,\ell}^0}^{U_{i,\ell}^0+\frac{c}{\sqrt{2m_1 r}}} \frac{\mathcal{C}_f}{\gamma_j}\tilde{f}\left(\frac{u}{\gamma_j}\right)\mathrm{d}u.\end{aligned}$$

Elementary calculus shows that, as $\tilde{f}$ is nonnegative and nonincreasing, $\gamma_j \mapsto \tilde{f}\left(u/\gamma_j\right)/\gamma_j$ is nonincreasing. As such, when $\gamma \in E$ and $j \leq r$, $\gamma_j \leq 1$,

$$\begin{aligned}\pi\left(\left(U_{i,\ell}-U_{i,\ell}^0\right)^2 \leq \frac{c^2}{2m_1 r}\middle|\gamma\right) &\geq \frac{c\mathcal{C}_f}{\sqrt{2m_1 r}}\tilde{f}\left(U_{i,\ell}^0+\frac{c}{\sqrt{2m_1 r}}\right) \\ &\geq \frac{c\mathcal{C}_f}{\sqrt{2m_1 r}}\tilde{f}\left(U_{i,\ell}^0+1\right)\end{aligned}$$

as $c \leq \sqrt{Kr} \leq \sqrt{m_1 r}$. We plug this result and the lower-bound $I_4 \geq 1/2$ into the expression of I_3 to get

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

So

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

$$\geq \frac{1}{2}\left(\frac{c\mathcal{C}_f}{\sqrt{2m_1 r}}\right)^{m_1 r}\left[\prod_{\substack{1\leq i\leq m_1\\ 1\leq \ell\leq r}} \tilde{f}\left(U^0_{i,\ell}+1\right)\right]\alpha^K\left(\frac{c}{2S_f\sqrt{2Km_1}}\right)^{\beta K},$$

using C2. Proceeding exactly in the same way,

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

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

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

This ends the proof of the lemma. □

A.5 Conclusion

We now plug Lemma A.1 and Lemma A.2 into Theorem 2. We obtain, under C1, C2 and C3,

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

Remind that we fixed $\lambda = \frac{1}{4\sigma^2}$. We finally (approximately) optimize with respect to $c \in \left(0, \sqrt{Kr}\right]$: this would lead to

$$\begin{aligned} c &= \frac{(m_1 \vee m_2) r}{2\lambda(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr})^2} \\ &= \frac{2\sigma^2(m_1 \vee m_2) r}{(\|U^0\|_F + \|V^0\|_F + \sqrt{Kr})^2} \leq \frac{2\sigma^2(m_1 \vee m_2)}{K} \leq \sqrt{Kr} \end{aligned}$$

as soon as $\sigma^2 \leq \frac{2K^{\frac{3}{2}}}{(m_1 \vee m_2)}$. As this might be restrictive, we choose a much smaller c that will only have impact on the logarithmic terms in the oracle inequality:

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

with the condition that $\sigma^2 \leq 2K^{\frac{3}{2}}$. The inequality becomes

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

which ends the proof.