An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization

Pierre Alquier, Benjamin Guedj

An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization

Pierre Alquier* & Benjamin Guedj†

August 23, 2016

Abstract

The aim of this paper is to provide some theoretical understanding of Bayesian non-negative matrix factorization methods. We derive an oracle inequality for a quasi-Bayesian estimator. This result holds for a very general class of prior distributions and shows how the prior affects the rate of convergence. We illustrate our theoretical results with a short numerical study along with a discussion on existing implementations.

1 Introduction

Non-negative matrix factorization (NMF) is a set of algorithms in high-dimensional data analysis which aims at factorizing a large matrix M with non-negative 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 non-negative entries. Interpreting the columns $M_{\cdot,j}$ of M as (non-negative) signals, NMF amounts to decompose (exactly, or approximately) each signal as a combination of the "elementary" signals $U_{\cdot,1}, \ldots, U_{\cdot,K}$:

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

*CREST, ENSAE, Université Paris Saclay, pierre.alquier@ensae.fr. This 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).

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

Since the seminal paper from Lee and Seung (1999), NMF was successfully applied to various fields such 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 topics extraction in texts (Paisley et al., 2015). In all these applications, it has been pointed out that NMF provides a decomposition which is usually interpretable. Donoho and Stodden (2003) have given a theoretical foundation to this interpretatibility by exhibiting conditions under which the decomposition $M \simeq UV^T$ is unique. 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, a Bayesian strategy seems appropriate. Bayesian tools have extensively been used for (general) matrix factorization (Corander and Villani, 2004; Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Lawrence and Urtasun, 2009; Zhou et al., 2010) and have been adapted for the Bayesian NMF problem (Moussaoui et al., 2006; Cemgil, 2009; Févotte et al., 2009; Schmidt et al., 2009; Tan and Févotte, 2009; Zhong and Girolami, 2009, among others).

The aim of this paper is to provide some theoretical analysis on the performance of Bayesian NMF. We propose a general Bayesian estimator for which we derive an oracle inequality. The message of this theoretical bound is that our procedure is able to adapt to the unknown rank of M. This result holds even for noisy observations, with no parametric assumption on the noise. That is, the likelihood used to build the Bayesian estimator does not have to be well-specified (it is usually referred to as a quasi-likelihood). To this regard, our procedure may be called quasi-Bayesian. The use of quasi-likelihoods in Bayesian estimation is advocated by Bissiri et al. (2013) using decision-theoretic arguments. This methodology is also popular in machine learning, and various authors developed a theoretical framework to analyze it (Shawe-Taylor and Williamson, 1997; McAllester, 1998; Catoni, 2003, 2004, 2007; Dalalyan and Tsybakov, 2008, this is known as the PAC-Bayesian theory).

The paper is organized as follows. Notation for the Bayesian 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 Appendix A. We illustrate this theoretical result by a short numerical study. To that matter, we discuss two

implementations (MCMC and optimization) of the quasi-Bayesian estimator in Section 4. Section 5 contains our numerical experiments on synthetic and real data. The main message of these experiments is that the choice of the prior appear to have limited impact on the accuracy of the reconstruction of M. However, a fine tuning of the prior may be crucial if the goal is to enforce shrinkage of many terms in (1) to 0.

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 $\|A\|_F^2 = \langle A, A \rangle_F$. 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. 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 non-negative entries. Note that, under (2), depending on the distribution of $\mathcal{E}$, Y might have some negative entries (the non-negativity 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 non-negative and bounded function g (put $\sigma^2 := \|g\|_\infty$) such that*

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

The introduction of this rather involved 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 $\sigma^2 := s^2$) or uniform ($\varepsilon_{i,j} \sim \mathcal{U}[-b, b]$ and $\sigma^2 := b^2/2$).

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 5 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 for 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

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

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

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

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

$$\pi(U, V, \gamma) = \prod_{\ell=1}^{K} \left(\prod_{i=1}^{m_1} g_{\gamma_\ell}(U_{i,\ell}) \right) \left(\prod_{j=1}^{m_2} g_{\gamma_\ell}(V_{j,\ell}) \right) h(\gamma_\ell) \tag{3}$$

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

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

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 (3) encompasses as special cases almost all the priors used in the papers mentioned in the introduction. We end this subsection with classical examples of functions f and h.

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 Schmidt et al. (2009). A generalization of the exponential prior is the gamma prior used in Cemgil (2009).

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 $\zeta > 1$.

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, 2013). Alquier et al. (2014) chose h as the density of the $\Gamma(m_1 + m_2 - 1/2, b)$ distribution for $b > 0$. Both lead to explicit conditional posteriors for γ.

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 be the actual likelihood up to a multiplicative constant. As we pointed out, the use of quasi-likelihoods to define quasi-posteriors

is becoming rather popular in Bayesian statistics and machine learning literatures. Here, the Frobenius norm is to be seen as a fitting criterion rather than as a ground truth. Note that other criterion were used in the literature: the Poisson likelihood (Lee and Seung, 1999), or the Itakura-Saito divergence (Févotte et al., 2009).

Definition 1. *We define the quasi-posterior as*

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

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. The posterior mean will be denoted by

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

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 shrinked to 0. The following set of matrices is introduced to formalize this idea. For any $r \in \{1,\dots,K\}$, let $\mathcal{M}_r$ be the set of pairs of matrices (U^0,V^0) with non-negative entries such that

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

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, in the form of the following sharp oracle inequality.

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

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

We remind the reader that the proof is given in Appendix A. 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 explicit the rate of convergence, we now provide a weaker version, where we only compare $\widehat{M}_\lambda$ to the best factorization in $\mathcal{M}_r(L)$.

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

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

First, note that when $L^2 = \mathcal{O}(\sigma)$, up to log terms, 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 $(U^0,V^0)\in\mathcal{M}_r(L)$. Alternatively, when $M\in\mathcal{M}_r(L)$ only for huge L, the log term in

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

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

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

which is terrible for large L. 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. Still, this prior has not received much attention from practitioners, since its implementation seems less straightforward.

4 Algorithms for Bayesian NMF

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 Salakhutdinov and Mnih (2008), Alquier et al. (2014) for the general case and Moussaoui et al. (2006), Schmidt et al. (2009) and Zhong and Girolami (2009) for NMF.

However, it is well known by practitioners that the computational cost of MCMC-based methods becomes prohibitive in very high dimensional models. Indeed, the optimization algorithms used in non-Bayesian NMF are much faster to converge in practice. Many of these algorithms share the characteristic to minimize iteratively the criterion $\|Y - UV^\top\|_F^2$ with respect to U, then V, an approach known as block coordinate descent (Bertsekas, 1999). This minimization may be achieved by the original multiplicative algorithm Lee and Seung (1999, 2001) or projected gradient descent (Lin, 2007; Guan et al., 2012). This approach is studied in full generality in Xu and Yin (2013). Other methods include second order schemes (Kim et al., 2008), linear progamming (Bittorf et al., 2012) or ADMM (alternative direction method of multipliers Boyd et al., 2011; Xu et al., 2012).

In order to enjoy the desirable computational properties of the aforementioned algorithms in Bayesian statistics, some authors proposed to use optimization tools to compute an approximation of the posterior. This method is known as Variational Bayes (Jordan et al., 1999; MacKay, 2002; Bishop, 2006). The theoretical properties of this approximation are studied in Alquier et al. (2015). It was used for Bayesian matrix factorization (Lim and Teh, 2007; Alquier et al., 2014) and more recently in Bayesian NMF (Paisley et al., 2015).

Finally, another option is to use optimization algorithms to compute the mode of the posterior, also known as maximum a posteriori (MAP). While this estimator is in general different from the posterior mean, it still leads to optimal rates of convergence in some complex models (see Abramovich and Lahav, 2015, in the context of additive regression).

Still, we would like to point out that the function to be maximized is not concave with respect to (U,V), which makes the optimization problem hard to solve. In particular, note that no rates of convergence for optimization algorithms are known for NMF in general. We do not intend to solve this problem here as it goes far beyond the scope of this paper.

In the sequel, we give a pseudo-code for the Gibbs sampler, and block coordinate descent to compute the MAP.

4.1 General form of the conditional posteriors for the Gibbs sampler

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 $k = 1,\dots,N$:

For $i = 1,\dots,m_1$: draw $U_{i,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(U_{i,\cdot}|V^{(k-1)},\gamma^{(k-1)},Y)$.

For $j = 1,\dots,m_2$: draw $V_{j,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(V_{j,\cdot}|U^{(k)},\gamma^{(k-1)},Y)$.

For $\ell = 1,\dots,K$: draw $\gamma_\ell^{(k)} \sim \widehat{\rho}_\lambda(\gamma_\ell|U^{(k)},V^{(k)},Y)$.

We now explicit the conditional posteriors. We first remind the quasi-posterior formula

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

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

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

In all generality, sampling from $\widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y)$ might require considerable effort. In such cases, a Metropolis-within-Gibbs approach is often the best choice, sadly at the cost of quite substantial computational power. In the high-dimensional context of NMF, this choice appeared unrealistic to us. However, it appears that when the prior f is exponential or truncated Gaussian, sampling from $\widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y)$ becomes straightforward. The detailed algorithms are provided in Appendix B.

4.2 Optimization through block coordinate descent

In this section, we discuss an algorithm for the implementation of the MAP estimator

$$(\widetilde{U}_\lambda,\widetilde{V}_\lambda,\widetilde{\gamma}_\lambda) = \arg\max_{U,V,\gamma}\ \widehat{\rho}_\lambda(U,V,\gamma)$$

$$= \arg\min_{U,V,\gamma} \left\{ \lambda \|Y - UV^\top\|_F^2 - \log \pi(U,V,\gamma) \right.$$
$$\left. - \sum_{i=1}^{m_1} \sum_{\ell=1}^{K} \log\left(g_{\gamma_\ell}(U_{i,\ell})\right) - \sum_{j=1}^{m_2} \sum_{\ell=1}^{K} \log\left(g_{\gamma_\ell}(V_{j,\ell})\right) - \sum_{\ell=1}^{K} \log\left(h(\gamma_\ell)\right) \right\}.$$

The block coordinate descent approach is used in practice with reasonnable results. This algorithm seems to be relatively standard in NMF as discussed above 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, $k := k+1$:

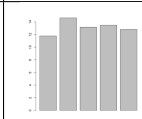
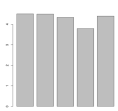
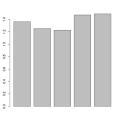
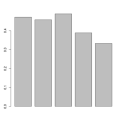
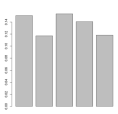
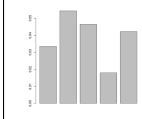
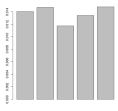
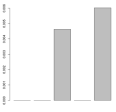
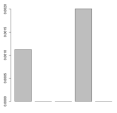
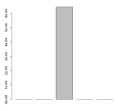
$$U^{(k)} := \arg\min_{U} \left\{ \lambda \|Y - U(V^{(k-1)})^\top\|_F^2 - \sum_{i=1}^{m_1} \sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(U_{i,\ell})] \right\}$$
$$V^{(k)} := \arg\min_{V} \left\{ \lambda \|Y - U^{(k)}V^\top\|_F^2 - \sum_{j=1}^{m_2} \sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(V_{j,\ell})] \right\}$$
$$\gamma^{(k)} := \arg\min_{\gamma} \sum_{\ell=1}^{K} \left\{ -\sum_{i=1}^{m_1} \log[g_{\gamma_\ell}(U_{i,\ell}^{(k)})] - \sum_{j=1}^{m_2} \log[g_{\gamma_\ell}(V_{j,\ell}^{(k)})] - \log[h(\gamma_\ell)] \right\}$$

Note that when the functions f and h are not conveniently chosen, this optimization problem can be very cumbersome, if tractable at all. In the examples mentioned in Section 2, that is, when f is exponential or truncated Gaussian, the optimization problems in U and V are quadratic problems with a non-negativity constraint, that can be solved thanks to the various approaches mentioned above. We derive explicit forms of Algorithm 2 using a projected gradient algorithm when the priors are exponential or truncated Gaussian in Appendix C.

5 Numerical experiments

Note that NMF has attracted a great deal of interest and the number of available algorithms is massive. The objective of this section is not to convince the reader that Bayesian NMF would uniformly be the best possible

Figure 1: Results of the simulations with $K = 5$: MSEs and vectors γ obtained for each value of b.

b	10^0	10^1	10^2	10^3	10^4
MSE	0.000762	0.002750	0.001089	0.000712	0.000774
γ					
b	10^5	10^6	10^7	10^8	10^9
MSE	0.000711	0.002456	0.009929	0.146539	0.632924
γ					

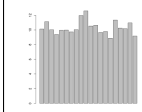
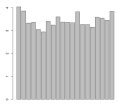
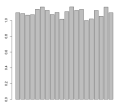
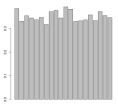
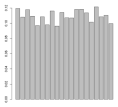
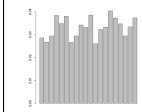
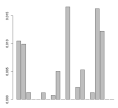
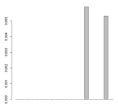
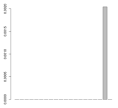
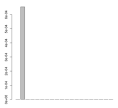
method. Instead, our main objective is to illustrate the influence of the choice of the prior hyperparameters. To do so, we performed a numerical study of the performances of the Bayesian MAP estimator coupled with the exponential prior with parameter 1 on (U,V) and with the gamma prior $\Gamma(m_1+m_1-1/2,b)$ on the coefficients γ_j. We approximated this estimator thanks to Algorithm 2.

To assess the impact of the hyperparameter b on the quality of the factorization, let us consider the following exponential grid: $b \in \{10^0, 10^1, \ldots, 10^9\}$. We define 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 an optimal size for the dictionary U. In theory, note that our estimator will report a dictionary U with maximal size K. However in practice, we can expect that many γ_ℓs will be shrunk to 0 so that thresholding their values will not affect the performance of the reconstruction, thus setting many $U_{\cdot,\ell}$s to 0. 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 uniform $\mathcal{U}([0,3])$ distribution. In a first time choose $K = 5$. We simulate the entries $\mathcal{E}_{i,j}$ of $\mathcal{E}$ independenly from a Gaussian $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 = 0.01$. We report in Figure 1 the MSEs and vectors γ obtained for each value of b.

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

Figure 2: Results of the simulations = with $K = 20$: MSEs and vectors γ obtained for each value of b.

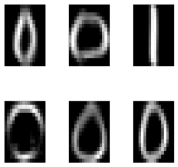
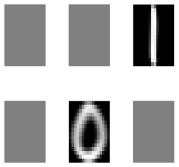
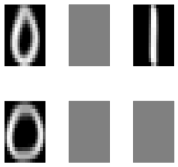
b	10^0	10^1	10^2	10^3	10^4
MSE	0.013176	0.010170	0.009517	0.009934	0.006730
γ					
b	10^5	10^6	10^7	10^8	10^9
MSE	0.006475	0.001572	0.007313	0.607693	0.640797
γ					

to identify the minimal support of γ. For larger b, a minimal support of γ 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 by 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 2).

The optimal value in terms of MSE seems to lie around $b = 10^6$. About the estimation of the size of the dictionary, the vector γ is sparse for $b = 10^6$ but we only identify the true sparsity for $b = 10^7$. The rank identification problem seems more sensitive to a proper tuning of b than the estimation problem, *i.e.*, minimization of the MSE.

These findings are confirmed by an additional experiment on the famous USPS database (Le Cun et al., 1990). Each image is a 16×16 pixels and thus can be represented by a vector in $\mathbb{R}^{256}$. We store all the images of zeros and ones in a 2199×256 matrix Y. We then run our algorithm with $K = 6$. We still choose f as the exponential prior and h as the gamma prior with different values of b. We plot the images corresponding to all the $U_{\cdot,\ell}$ for $\ell \in \{1,\dots,6\}$ in Figure 3. For small values of b, we identify various shapes of zeros. When b increases, we shrink the dictionary, and this leads to a smaller set of shapes of zeros. A huge value for b leads to a uniformly null decomposition.

Figure 3: Experiments on the USPS dataset with f as the exponential prior and h as the gamma prior with parameter b. We represent the images $U_{\cdot,\ell}$ obtained for various values of b. A uniformly grey image means that $U_{\cdot,\ell} = 0$.

b	U	b	U
10^3		10^9	
10^8		10^{10}	

Acknowledgements

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

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. 9

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. 5, 19

P. Alquier, V. Cottet, N. Chopin, and J. Rousseau. Bayesian matrix completion: prior specification. Preprint arXiv:1406.1440, 2014. 5, 8, 26

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. Preprint arXiv:1506.04091, 2015. 8

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999. 8

C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 10. Springer, 2006. 8, 9

P. Bissiri, C. Holmes, and S. Walker. A general framework for updating belief distributions. Preprint arXiv:1306.6430, to appear in the Journal of the Royal Statistical Society Series B, 2013. 2

V. Bittorf, B. Recht, C. Re, and J. Tropp. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pages 1214–1222, 2012. 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. 8

O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint Laboratoire de Probabilités et Modèles Aléatoires, PMA-840, 2003. 2, 19

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. 2, 19

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. 2, 19

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2009. 2, 5

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. 2, 4, 8, 19, 20

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

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. 2, 6

I. Giulini. PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. Preprint arXiv:1511.06263, 2015. 19

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. 8

B. Guedj and P. Alquier. PAC-Bayesian Estimation and Prevision in Sparse Additive Models. *Electronic Journal of Statistics*, 7:264–291, 2013. 19

B. Guedj and S. Robbiano. PAC-Bayesian High Dimensional Bipartite Ranking. Preprint arXiv:1511.02729, 2015. 19

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. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999. 8

D. Kim, S. Sra, and I. S. Dhillon. Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Statistical Analysis and Data Mining*, 1(1):38–51, 2008. 8

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, 26

Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*. Citeseer, 1990. 13

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 2, 6, 8

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001. 8

G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006. 19

L. Li, B. Guedj, and S. Loustau. PAC-Bayesian online clustering. *arXiv preprint arXiv:1602.00522*, 2016. 19

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, 8, 26

C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007. 8

D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2002. 8

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. 19, 26

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York, 1998. ACM. 2, 19

S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. *IEEE Transactions on Signal Processing*, 54(11): 4133–4145, 2006. 2, 8

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, 8

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. 2, 5, 8, 26

M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation*, pages 540–547. Springer, 2009. 2, 5, 8

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. 2, 19

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. 19

V. Y. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009. 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. 26

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. 8

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. 8

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. 13

M. Zhong and M. Girolami. Reversible jump MCMC for non-negative matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 663–670, 2009. 2, 8

M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin. Nonparametric Bayesian Matrix Completion. In *Proc. IEEE SAM*, 2010. 2, 26

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). This methods were used in various complex models of statistical learning (Guedj and Alquier, 2013; Alquier, 2013; Suzuki, 2015; Mai and Alquier, 2015; Guedj and Robbiano, 2015; Giulini, 2015; Li et al., 2016). 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 of Theorem 1 is organized as follows. First, we define in Section A.2 a parametric family of probability distributions ρ:

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

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

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

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 $0 < c \le \sqrt{\sigma K r}$, the density

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

A.3 Upper bound for the integral part

Lemma A.1.

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

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 $c \le \sqrt{\sigma K r}$ to get the result. □

A.4 Upper bound for the Kullback-Leibler divergence

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

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

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

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

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

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

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

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

So

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

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

using C2. Proceeding exactly in the same way,

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

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

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

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,

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

Remind that we fixed $\lambda = \frac{1}{4\sigma^2}$. We finally choose

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

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

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

which ends the proof.

B Explicit formulas for the Gibbs sampler with exponential and Gaussian priors

B.1 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 $\gamma^{-1} = (1/\gamma_1, \ldots, 1/\gamma_K)$. So $\widehat{\rho}_\lambda(U_{i,\cdot}|V,\gamma,Y)$ amounts to a truncated Gaussian distribution

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

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.

B.2 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,

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

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.

C Explicit optimization algorithms with exponential and Gaussian priors

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 $\beta/(\alpha+1)$ and $\mu[\sqrt{1+(9\mu^2)/(4\nu^2)}-(3\mu)(2\nu)]$, 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.

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 $\ell = 1,\dots,K$:

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