Convex nonnegative matrix factorization with missing data

Ronan Hamon, Valentin Emiya, Cédric Févotte

2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY

CONVEX NONNEGATIVE MATRIX FACTORIZATION WITH MISSING DATA

*Ronan Hamon, Valentin Emiya**

Aix Marseille Univ, CNRS, LIF, Marseille, France

Cédric Févotte

CNRS & IRIT, Toulouse, France

ABSTRACT

Convex nonnegative matrix factorization (CNMF) is a variant of nonnegative matrix factorization (NMF) in which the components are a convex combination of atoms of a known dictionary. In this contribution, we propose to extend CNMF to the case where the data matrix and the dictionary have missing entries. After a formulation of the problem in this context of missing data, we propose a majorization-minimization algorithm for the solving of the optimization problem incurred. Experimental results with synthetic data and audio spectrograms highlight an improvement of the performance of reconstruction with respect to standard NMF. The performance gap is particularly significant when the task of reconstruction becomes arduous, e.g. when the ratio of missing data is high, the noise is steep, or the complexity of data is high.

Index Terms— matrix factorization, nonnegativity, low-rankness, matrix completion, spectrogram inpainting

1. INTRODUCTION

Convex NMF (CNMF) [1] is a special case of nonnegative matrix factorization (NMF) [2], in which the matrix of components is constrained to be a linear combination of atoms of a known dictionary. The term "convex" refers to the constraint of the linear combination, where the combination coefficients forming each component are nonnegative and sum to 1. Compared to the fully unsupervised NMF setting, the use of known atoms is a source of supervision that may guide learning based on this additional data: in particular, an interesting case of CNMF consists in auto-encoding the data themselves, by defining the atoms as the data matrix. CNMF has been of interest in a number of contexts, such as clustering, data analysis, face recognition, or music transcription [1, 3]. It is also related to the *self-expressive* dictionary-based representation proposed in [4].

An issue that has not yet been addressed is when the data matrix has missing coefficients. Such an extension of CNMF is worth being considered, as it opens the way to data-reconstruction settings with nonnegative low-rank constraints, which covers several relevant applications. One example concerns the field of image or audio inpainting [5, 6, 7, 8], where CNMF may improve the current reconstruction techniques. In inpainting of audio spectrograms for example, setting up the dictionary to be a comprehensive collection of notes from a specific instrument may guide the factorization toward a realistic and meaningful decomposition, increasing the quality of the reconstruction of the missing data. In this contribution, we also consider the case where the dictionary may have missing coefficients itself.

The paper is organized as follows. Section 2 formulates CNMF in the presence of missing entries in the data matrix and in the dictionary. Section 3 describes the proposed majorization-minimization (MM) algorithm. Sections 4 and 5 report experimental results with synthetic data and audio spectrograms.

2. CONVEX NONNEGATIVE MATRIX FACTORIZATION WITH MISSING DATA

2.1. Notations and definitions

For any integer N, the integer set $\{1, 2, \dots, N\}$ is denoted by $[N]$. The coefficients of a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ are denoted by either a_{mn} or $[\mathbf{A}]_{mn}$. The element-wise matrix product, matrix division and matrix power are denoted by $\mathbf{A}.\mathbf{B}$, $\frac{\mathbf{A}}{\mathbf{B}}$ and $\mathbf{A}^{.\gamma}$, respectively where $\mathbf{A}$ and $\mathbf{B}$ are matrices with same dimensions and γ is a scalar. $\mathbf{0}$ and $\mathbf{1}$ denote vectors or matrices composed of zeros and ones, respectively, with dimensions that can be deduced from the context. The element-wise negation of a binary matrix $\mathbf{M}$ is denoted by $\bar{\mathbf{M}} \triangleq \mathbf{1} - \mathbf{M}$.

2.2. NMF and Convex NMF

NMF consists in approximating a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ as the product $\mathbf{WH}$ of two nonnegative matrices $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$. Often, $K < \min(F, N)$, such that $\mathbf{WH}$ is a low-rank approximation of $\mathbf{V}$. Every sample $\mathbf{v}_n$, the n-th column of $\mathbf{V}$, is thus decomposed as a linear combination of K elementary *components* or *patterns* $\mathbf{w}_1, \dots, \mathbf{w}_K \in \mathbb{R}_+^F$, the columns of $\mathbf{W}$. The coefficients of the linear combination are given by the n-th column $\mathbf{h}_n$ of $\mathbf{H}$.

In [9] and [10], algorithms have been proposed for the unsupervised estimation of $\mathbf{W}$ and $\mathbf{H}$ from $\mathbf{V}$, by minimization

*This work was supported by ANR JCJC program MAD (ANR- 14-CE27-0002).

978-1-5090-0746-2/16/$31.00

of the cost function $D_\beta\left(\mathbf{V}|\mathbf{WH}\right)=\sum_{fn}d_\beta\left(v_{fn}|[\mathbf{WH}]_{fn}\right)$, where $d_\beta\left(x|y\right)$ is the β-divergence defined as:

$$d_\beta\left(x|y\right)\triangleq\begin{cases}\frac{1}{\beta(\beta-1)}\left(x^\beta+(\beta-1)\,y^\beta-\beta xy^{\beta-1}\right) & \\ \qquad\qquad\qquad\qquad\text{for }\beta\in\mathbb{R}\backslash\{0,1\} & \\ x\log\frac{x}{y}-x+y\qquad\qquad\text{for }\beta=1 & \\ \frac{x}{y}-\log\frac{x}{y}-1\qquad\qquad\text{for }\beta=0\end{cases}\tag{1}$$

When ill-defined, we set by convention $d_\beta(0|0)=0$.

CNMF is a variant of NMF in which $\mathbf{W}=\mathbf{SL}$. $\mathbf{S}=[\mathbf{s}_1,\ldots,\mathbf{s}_P]\in\mathbb{R}_+^{F\times P}$ is a nonnegative matrix of *atoms* and $\mathbf{L}=[\mathbf{l}_1,\ldots,\mathbf{l}_K]\in\mathbb{R}_+^{P\times K}$ is the so-called *labeling* matrix. Each dictionary element $\mathbf{w}_k$ is thus equal to $\mathbf{Sl}_k$, with usually $P>>K$, and the data is in the end decomposed as $\mathbf{V}=\mathbf{SLH}$. The scale indeterminacy between $\mathbf{L}$ and $\mathbf{H}$ may be lifted by imposing $\|\mathbf{l}_k\|_1=1$, in which case $\mathbf{w}_k$ is precisely a convex combination of the elements of the subspace $\mathbf{S}$. CNMF can be related to the so-called archetypal analysis [11], but without considering any nonnegativity constraint.

The use of known examples in $\mathbf{S}$ can then be seen as a source of supervision that guides learning. A special case of CNMF is obtained by setting $\mathbf{S}=\mathbf{V}$, thus auto-encoding the data as $\mathbf{VLH}$. This particular case is studied in depth in [1]. In this paper, we consider the general case for $\mathbf{S}$, with or without missing data.

2.3. Convex NMF with missing data

We assume that some coefficients in $\mathbf{V}$ and $\mathbf{S}$ may be missing. Let $\mathcal{V}\subset[F]\times[N]$ be a set of pairs of indices that locates the observed coefficients in $\mathbf{V}$: $(f,n)\in\mathcal{V}$ iff v_{fn} is known. Similarly, let $\mathcal{S}\subset[F]\times[P]$ be a set of pairs of indices that locates the observed coefficients in $\mathbf{S}$. The use of sets $\mathcal{V}$ and $\mathcal{S}$ may be reformulated equivalently by defining masking matrices $\mathbf{M}_\mathcal{V}\in\{0,1\}^{F\times N}$ and $\mathbf{M}_\mathcal{S}\in\{0,1\}^{F\times P}$ from $\mathcal{V}$ and $\mathcal{S}$ as

$$[\mathbf{M}_\mathcal{V}]_{fn}\triangleq\begin{cases}1\text{ if }(f,n)\in\mathcal{V}\\0\text{ otherwise}\end{cases}\qquad\forall(f,n)\in[F]\times[N]\tag{2}$$

$$[\mathbf{M}_\mathcal{S}]_{fp}\triangleq\begin{cases}1\text{ if }(f,p)\in\mathcal{S}\\0\text{ otherwise}\end{cases}\qquad\forall(f,p)\in[F]\times[P]\tag{3}$$

A major goal in this paper is to estimate $\mathbf{L}$, $\mathbf{H}$ and the missing entries in $\mathbf{S}$, given the partially observed data matrix $\mathbf{V}$. Denoting by $\mathbf{S}^o$ the set of observed/known dictionary matrix coefficients, our aim is to minimize the objective function

$$C\left(\mathbf{S},\mathbf{L},\mathbf{H}\right)\triangleq D_\beta\left(\mathbf{M}_\mathcal{V}.\mathbf{V}|\mathbf{M}_\mathcal{V}.\mathbf{SLH}\right)\tag{4}$$

subject to $\mathbf{S}\in\mathbb{R}_+^{F\times P}$, $\mathbf{L}\in\mathbb{R}_+^{P\times K}$, $\mathbf{H}\in\mathbb{R}_+^{K\times N}$, and $\mathbf{M}_\mathcal{S}.\mathbf{S}=\mathbf{M}_\mathcal{S}.\mathbf{S}^o$. The particular case where the dictionary equals the data matrix itself is obtained by setting $(\mathbf{M}_\mathcal{S},\mathbf{S}^o)\triangleq(\mathbf{M}_\mathcal{V},\mathbf{V})$.

Algorithm 1 CNMF with missing data

Require: $\mathbf{V}$, $\mathbf{S}^o$, $\mathbf{M}_\mathcal{V}$, $\mathbf{M}_\mathcal{S}$, β

Initialize $\mathbf{S}$, $\mathbf{L}$, $\mathbf{H}$ with random nonnegative values

loop

Update $\mathbf{S}$:

$$\mathbf{S}\leftarrow\mathbf{M}_\mathcal{S}.\mathbf{S}^o+\overline{\mathbf{M}}_\mathcal{S}.\mathbf{S}.\left(\frac{\left(\mathbf{M}_\mathcal{V}.\left(\mathbf{SLH}\right)^{.(\beta-2)}.\mathbf{V}\right)\left(\mathbf{LH}\right)^T}{\left(\mathbf{M}_\mathcal{V}.\left(\mathbf{SLH}\right)^{.(\beta-1)}\right)\left(\mathbf{LH}\right)^T}\right)^{.\gamma(\beta)}\tag{5}$$

Update $\mathbf{L}$:

$$\mathbf{L}\leftarrow\mathbf{L}.\left(\frac{\mathbf{S}^T\left(\mathbf{M}_\mathcal{V}.\left(\mathbf{SLH}\right)^{.(\beta-2)}.\mathbf{V}\right)\mathbf{H}^T}{\mathbf{S}^T\left(\mathbf{M}_\mathcal{V}.\left(\mathbf{SLH}\right)^{.(\beta-1)}\right)\mathbf{H}^T}\right)^{.\gamma(\beta)}\tag{6}$$

Update $\mathbf{H}$:

$$\mathbf{H}\leftarrow\mathbf{H}.\left(\frac{\left(\mathbf{SL}\right)^T\left(\mathbf{M}_\mathcal{V}.\left(\mathbf{SLH}\right)^{.(\beta-2)}.\mathbf{V}\right)}{\left(\mathbf{SL}\right)^T\left(\mathbf{M}_\mathcal{V}.\left(\mathbf{SLH}\right)^{.(\beta-1)}\right)}\right)^{.\gamma(\beta)}\tag{7}$$

Rescale $\mathbf{L}$ and $\mathbf{H}$:

$$\forall k\in[K],\mathbf{h}_k\leftarrow\|\mathbf{l}_k\|_1\times\mathbf{h}_k\tag{8}$$

$$\mathbf{l}_k\leftarrow\frac{\mathbf{l}_k}{\|\mathbf{l}_k\|_1}\tag{9}$$

end loop

return $\mathbf{S}$, $\mathbf{L}$, $\mathbf{H}$

3. PROPOSED ALGORITHM

3.1. General description of the algorithm

Algorithm 1 extends the algorithm proposed in [9] for complete CNMF with the β-divergence to the case of missing entries in $\mathbf{V}$ or $\mathbf{S}$. The algorithm is a block-coordinate descent procedure in which each block is one the three matrix factors. The updates of each block/factor is obtained via majorization-minimization (MM), a classic procedure that consists in iteratively minimizing a tight upper bound (called auxiliary function) of the objective function. In the present setting, the MM procedure leads to multiplicative updates, characteristic of many NMF algorithms, that automatically preserve nonnegativity given positive initialization.

3.2. Detailed updates

We consider the optimization of $C\left(\mathbf{S},\mathbf{L},\mathbf{H}\right)$ with respect to each of its three arguments individually, using MM. Current updates are denoted with a tilde, i.e., $\widetilde{\mathbf{S}}$, $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{H}}$. We start by recalling the definition of an auxiliary function:

Definition 1 (Auxiliary function)**.** The mapping $G\left(\mathbf{A}|\widetilde{\mathbf{A}}\right)$:

$\mathbb{R}_+^{I\times J} \times \mathbb{R}_+^{I\times J} \mapsto \mathbb{R}_+$ is an auxiliary function to $C(\mathbf{A})$ iff

$$\begin{cases} \forall \mathbf{A} \in \mathbb{R}_+^{I\times J}, & C(\mathbf{A}) = G(\mathbf{A}|\mathbf{A}) \\ \forall \mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{R}_+^{I\times J}, & C(\mathbf{A}) \leqslant G\left(\mathbf{A}|\widetilde{\mathbf{A}}\right). \end{cases} \tag{10}$$

The iterative minimization of $G(\mathbf{A}|\mathbf{A})$ with respect to $\mathbf{A}$, with replacement of $\widetilde{\mathbf{A}}$ at every iteration, monotonically decreases the objective $C(\mathbf{A})$ until convergence. As explained in detail in [9], the β-divergence may be decomposed into the sum of a convex term $\breve{d}_\beta(.|.)$, a concave term $\hat{d}_\beta(.|.)$ and a constant term cst. The first two terms can be majorized using routine Jensen and tangent inequalities, respectively, leading to tractable updates. The auxiliary functions used to derive Algorithm 1 are given by the three following propositions [1] and the monotonicity of the algorithm follows by construction.

Proposition 1 (Auxiliary function for $\mathbf{S}$)**.**
Let $\widetilde{\mathbf{S}} \in \mathbb{R}_+^{F\times P}$ be such that $\forall(f,n) \in [F]\times[N], \tilde{v}_{fn} > 0$, and $\forall(f,p) \in [F]\times[P], \tilde{s}_{fp} > 0$, where $\widetilde{\mathbf{V}} \triangleq \widetilde{\mathbf{S}}\widetilde{\mathbf{L}}\widetilde{\mathbf{H}}$. Then the function

$$G_S\left(\mathbf{S}|\widetilde{\mathbf{S}}\right) \triangleq \sum_{fn}[\mathbf{M}_\mathcal{V}]_{fn}\left[\breve{G}_{fn}\left(\mathbf{S}|\widetilde{\mathbf{S}}\right) + \hat{G}_{fn}\left(\mathbf{S}|\widetilde{\mathbf{S}}\right)\right] + cst$$

$$\text{where } \breve{G}_{fn}\left(\mathbf{S}|\widetilde{\mathbf{S}}\right) \triangleq \sum_p \frac{[\mathbf{L}\mathbf{H}]_{pn}\,\tilde{s}_{fp}}{\widetilde{v}_{fn}}\breve{d}_\beta\left(v_{fn}|\widetilde{v}_{fn}\frac{s_{fp}}{\tilde{s}_{fp}}\right)$$

$$\hat{G}_{fn}\left(\mathbf{S}|\widetilde{\mathbf{S}}\right) \triangleq \hat{d}_\beta(v_{fn}|\widetilde{v}_{fn}) + \hat{d}'_\beta(v_{fn}|\widetilde{v}_{fn})\sum_{fp}[\mathbf{L}\mathbf{H}]_{pn}(s_{fp}-\tilde{s}_{fp})$$

is an auxiliary function to $C(\mathbf{S},\mathbf{L},\mathbf{H})$ with respect to $\mathbf{S}$ and its minimum is given by equation (5)*. The auxiliary function decouples with respect to the individual coefficients of $\mathbf{S}$ and as such, the constraint $\mathbf{M}_\mathcal{S}.\mathbf{S} = \mathbf{M}_\mathcal{S}.\mathbf{S}^o$ is directly imposed by only updating the coefficients of $\mathbf{S}$ with indices in $\bar{\mathcal{S}}$.*

Proposition 2 (Auxiliary function for $\mathbf{L}$)**.**
Let $\widetilde{\mathbf{L}} \in \mathbb{R}_+^{P\times K}$ be such that $\forall(f,n) \in [F]\times[N], \tilde{v}_{fn} > 0$ and $\forall(p,k) \in [P]\times[K], \tilde{l}_{pk} > 0$, where $\widetilde{\mathbf{V}} \triangleq \mathbf{S}\widetilde{\mathbf{L}}\mathbf{H}$. Then the function

$$G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \triangleq \sum_{fn}[\mathbf{M}_\mathcal{V}]_{fn}\left[\breve{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) + \hat{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right)\right] + cst$$

$$\text{where } \breve{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \triangleq \sum_{pk} \frac{s_{fp}\tilde{l}_{pk}h_{kn}}{\widetilde{v}_{fn}}\breve{d}_\beta\left(v_{fn}|\widetilde{v}_{fn}\frac{l_{pk}}{\tilde{l}_{pk}}\right)$$

$$\hat{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \triangleq \hat{d}_\beta(v_{fn}|\widetilde{v}_{fn}) + \hat{d}'_\beta(v_{fn}|\widetilde{v}_{fn})\sum_{pk} s_{fp}h_{kn}\left(l_{pk}-\tilde{l}_{pk}\right)$$

[1] The proof of these propositions are available in the extended version at https://hal-amu.archives-ouvertes.fr/hal-01346492.

is an auxiliary function to $C(\mathbf{S},\mathbf{L},\mathbf{H})$ with respect to $\mathbf{L}$ and its minimum subject to $\mathbf{M}_\mathcal{S}.\mathbf{S} = \mathbf{M}_\mathcal{S}.\mathbf{S}^o$ for $\mathbf{M}_\mathcal{S} \in \{0,1\}^{F\times P}$ and $\mathbf{S}^o \in \mathbb{R}_+^{F\times P}$ is given by equation (6)*.*

Proposition 3 (Auxiliary function for $\mathbf{H}$)**.**
Let us define $\mathbf{W} \triangleq \mathbf{S}\mathbf{L}$ and let $\widetilde{\mathbf{H}} \in \mathbb{R}_+^{K\times N}$ be such that $\forall(f,n) \in [F]\times[N], \tilde{v}_{fn} > 0$ and $\forall(k,n) \in [K]\times[N], \tilde{h}_{kn} > 0$, where $\widetilde{\mathbf{V}} \triangleq \mathbf{W}\widetilde{\mathbf{H}}$. Then the function

$$G_H\left(\mathbf{H}|\widetilde{\mathbf{H}}\right) \triangleq \sum_{fn}[\mathbf{M}_\mathcal{V}]_{fn}\left[\breve{G}_{fn}\left(\mathbf{H}|\widetilde{\mathbf{H}}\right) + \hat{G}_{fn}\left(\mathbf{H}|\widetilde{\mathbf{H}}\right)\right] + cst$$

$$\text{where } \breve{G}_{fn}\left(\mathbf{H}|\widetilde{\mathbf{H}}\right) \triangleq \sum_k \frac{w_{fk}\tilde{h}_{kn}}{\widetilde{v}_{fn}}\breve{d}_\beta\left(v_{fn}|\widetilde{v}_{fn}\frac{h_{kn}}{\tilde{h}_{kn}}\right)$$

$$\hat{G}_{fn}\left(\mathbf{H}|\widetilde{\mathbf{H}}\right) \triangleq \hat{d}_\beta(v_{fn}|\widetilde{v}_{fn}) + \hat{d}'_\beta(v_{fn}|\widetilde{v}_{fn})\sum_k w_{fk}\left(h_{kn}-\tilde{h}_{kn}\right)$$

is an auxiliary function to $C(\mathbf{S},\mathbf{L},\mathbf{H})$ with respect to $\mathbf{H}$ and its minimum is given by equation (7)*.*

4. EXPERIMENT ON SYNTHETIC DATA

4.1. Experimental setting

The objective of this experiment is to analyze the performance of CNMF for reconstructing missing data, by comparing it with the regular NMF. We consider a data matrix $\mathbf{V}^*$ of rank K^* synthesized under the CNMF model $\mathbf{V}^* = \mathbf{S}^*\mathbf{L}^*\mathbf{H}^*$, where the matrix of atoms $\mathbf{S}^*$ and the ground truth factors $\mathbf{L}^*$ and $\mathbf{H}^*$ are generated as the absolute values of Gaussian noise. It is worth noting that $\mathbf{V}^*$ is also consistent with a NMF model by defining $\mathbf{W}^* = \mathbf{S}^*\mathbf{L}^*$. A perturbed data matrix $\mathbf{V}$ is obtained by considering a multiplicative noise, obtained using a Gamma distribution with mean 1 and variance $\frac{1}{\alpha}$. Hence the parameter α controls the importance of the perturbation. The mask $\mathbf{M}_\mathcal{V}$ of known elements in $\mathbf{V}$ is derived by considering missing coefficients randomly and uniformly distributed over the matrix, such that the ratio of missing values is equal to $\sigma_\mathcal{V}$. Generation of data is repeated 3 times, as well as the generation of the masks. Results are averaged over these repetitions.

From a matrix $\mathbf{V}$ with missing entries, NMF and CNMF with missing values are applied using K components. Only the case where $\beta = 2$ has been considered in this experiment. In both algorithms, the convergence is reached when the relative difference of the cost function between two iterations is below 10^{-5}. 3 repetitions are performed using different random initialization, and the best instance (i.e., the instance which minimizes the cost function) is retained. The reconstructed data matrix is obtained as $\widetilde{\mathbf{V}} = \mathbf{S}\mathbf{L}\mathbf{H}$.

The reconstruction error is obtained by computing the β-divergence between the noiseless data matrix $\mathbf{V}^*$, and the reconstructed matrix $\widetilde{\mathbf{V}}$; the error is computed on and averaged

along the missing coefficients only, as

$$e_{\text{test}} = \frac{1}{\sum_{ij} [\bar{\mathbf{M}}_{\mathcal{V}}]_{ij}} d_\beta(\bar{\mathbf{M}}_{\mathcal{V}}.\mathbf{V}^*, \bar{\mathbf{M}}_{\mathcal{V}}\tilde{\mathbf{V}}) \qquad (11)$$

where $\bar{\mathbf{M}}_{\mathcal{V}}$ is the mask of unknown elements in $\mathcal{V}$. In the case of CNMF, we consider two choices for $\mathbf{S}$: the data matrix $\mathbf{V}$ with missing values, and the ground truth matrix of atoms $\mathbf{S}^*$, considered here without missing values.

The following parameters are fixed: $F = 100$, $N = 250$, $P = 50$, $K^* = 10$, $\beta = 2$. We particularly investigate the influence of four factors: the number of estimated components $K \in [2, 14]$; the ratio of missing data $\sigma_{\mathcal{V}} \in [0.1, 0.9]$ in $\mathbf{V}$, i.e., of zeros in $\mathbf{M}_{\mathcal{V}}$; the choice of the matrix $\mathbf{S} \in \{\mathbf{S}^*, \mathbf{V}\}$ for the CNMF; the noise level $\alpha \in [10, 5000]$ in $\mathbf{V}$ (α is inversely proportional to the variance of the noise).

4.2. Results

We first focus on the influence of the number of estimated components k for the case where the true dictionary is fully known. Figure 1 displays the test error with respect to the number of estimated components K, for two levels of noise. Performance of reconstruction obtained by NMF and CNMF with $\mathbf{S} = \mathbf{S}^*$ are plotted, for different values of ratio of missing values in $\mathcal{V}$.

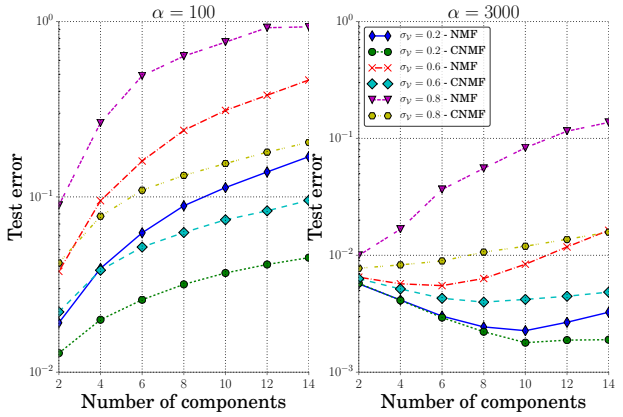

Fig. 1. Test error vs. number of components. Two levels of noise are displayed: high noise level ($\alpha = 100$, left) and low noise level ($\alpha = 3000$, right).

These results show that the noise level has a high influence on the best number of estimated components. As expected, a high noise requires a strong regularization, obtained here by selecting a low value of K. On the contrary, when the noise is low, the best choice of K is closer to the true value $K^* = 10$. Similarly, when the number of missing data is low ($\sigma_{\mathbf{V}} = 0.2$), one should set K to $K^* = 10$, either for CNMF or for NMF. When it gets higher, the optimal K gets lower in order to limit the effect of overfitting. In this case, the best number of components drops down to $K = 2$, either for CNMF or NMF, the former still performing better than the latter. These first results outline the difference between NMF and CNMF, emphasized in the next figures.

This comparison is augmented by considering the case where $\mathbf{S} = \mathbf{V}$, with respect to the ratio $\sigma_{\mathcal{V}}$ of missing values in $\mathbf{V}$. Figure 2 displays the performance of NMF and CNMF as a function of $\sigma_{\mathcal{V}}$, for two noise levels. CNMF with the true atoms ($\mathbf{S} = \mathbf{S}^*$) gives the best results on the full range missing data ratio. When there are very few missing data and a low noise, NMF performs almost as well as CNMF. However, the NMF error increases much faster than the CNMF error as the number of missing data grows, or as the noise in data becomes important. This higher sensitivity of NMF to missing data may be explained by overfitting since the number of free parameters in NMF is higher than in CNMF. In the case of CNMF with $\mathbf{S} = \mathbf{V}$, the model cannot fit the data as well as CNMF with $\mathbf{S} = \mathbf{S}^*$ or as NMF. Consequently, the resulting modeling error is observed when there is few missing data, and when comparing $\mathbf{S} = \mathbf{V}$ and $\mathbf{S} = \mathbf{S}^*$ on all values. However, it performs better than NMF at high values of $\sigma_{\mathbf{V}}$ since the constraint $\mathbf{S} = \mathbf{V}$ can be seen as a regularization.

We finally investigate the robustness of methods by looking at the influence of the multiplicative noise, controlled by the parameter α, on the performance. Figure 3 shows the test error for the NMF and the CNMF with $\mathbf{S} = \mathbf{S}^*$ with respect to α and for some values of $\sigma_{\mathcal{V}}$. As expected, the test error decreases according to the variance of the noise, inversely proportional to α. If a low value of α disrupts abruptly the performance of reconstruction ($\alpha < 10^3$), the test error is slightly decreasing for $\alpha > 10^3$. When the variance of the noise is close to zero ($\alpha = 5000$), the performance of NMF and CNMF are almost the same. The performance of reconstruction differs when the variance of the noise increases, as well as the ratio of missing values in $\mathbf{V}$.

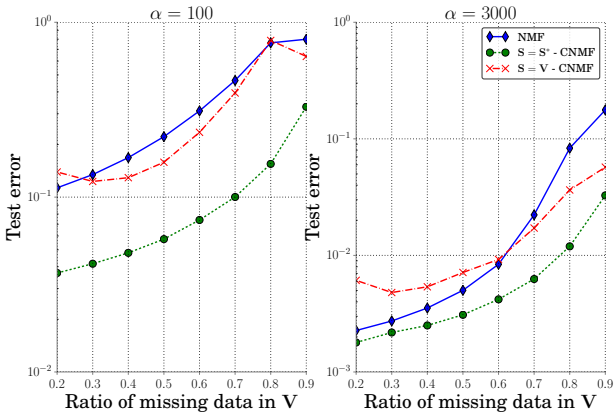

Fig. 2. Test error vs. ratio of missing data in $\mathbf{V}$, with $K = K^*$. Two levels of noise are displayed: high noise level ($\alpha = 100$, left) and low noise level ($\alpha = 3000$, right).

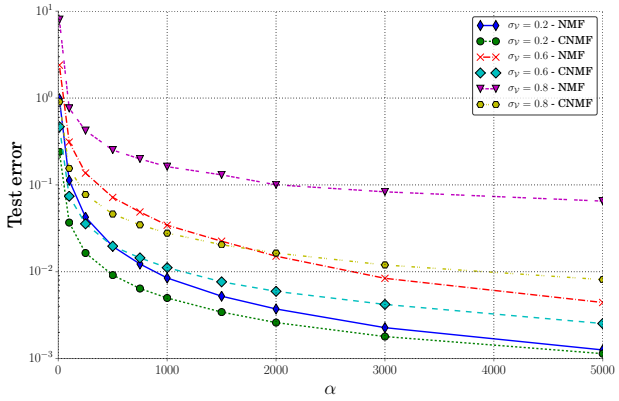

Fig. 3. Test error vs. noise level α. Each curve describes the performance of reconstruction of missing data in $\mathbf{V}$, with $K = K^*$, according to the method and the ratio of missing data $\sigma_{\mathcal{V}}$.

5. APPLICATION TO SPECTROGRAM INPAINTING

In order to illustrate the performance of the proposed algorithm on real data, we consider spectrograms of piano music signals, which are known to be well modeled by NMF methods [12]. Indeed, NMF may provide a note-level decomposition of a full music piece, each NMF component being the estimated spectrum and activation of a single note. This approximation has proved successful and is also limited in terms of modelling error and of lack of supervision to guide the NMF algorithm. In such condition, we have designed an experiment with missing data to compare regular NMF against two CNMF variants: in the first one, we set $\mathbf{S} = \mathbf{V}$; in the second one, $\mathbf{S}$ contains examples of all possible isolated note spectra from another piano.

5.1. Experimental setting

We consider 17 piano pieces from the MAPS dataset [13]. For each recording, the magnitude spectrogram is computed using a 46-ms sine window with 50%-overlap and 4096 frequency bins, the sampling frequency being 44.1kHz. Matrix $\mathbf{V}$ is created from the resulting spectrogram by selecting the $F = 500$ lower-frequency part and the first five seconds, i.e., the $N = 214$ first time frames. Missing data in $\mathbf{V}$ are artificially created by removing coefficients uniformly at random.

Three systems are compared, based on the *test error* defined as the β-divergence computed on the estimation of missing data. In all of them, the number of component K is set to the true number of notes available from the dataset annotation and we set $\beta = 1$. The first system is the regular NMF, randomly initialized. The second system is the proposed CNMF with $(\mathbf{M}_{\mathcal{S}}, \mathbf{S}^o) \triangleq (\mathbf{M}_{\mathcal{V}}, \mathbf{V})$. The third system is the proposed CNMF with $\mathbf{S} = \mathbf{D}$ set as a specific matrix $\mathbf{D}$ of $P = 61$ atoms. Each atom is a single-note spectrum extracted from the recording of another piano instrument from the MAPS dataset, from C3 to C8 [2] .

5.2. Results

Figure 4 displays the test error with respect to the ratio of missing data in $\mathbf{V}$, averaged over the 17 piano pieces. It clearly shows that the CNMF with the specific dictionary $\mathbf{S} = \mathbf{D}$ is much more robust to missing data than the other two systems. When less than 40% of data are missing, NMF performs slightly better; however, the NMF test error dramatically increases when more data are missing, by a factor 5.10^3 when more than 80% data are missing. This must be due to overfitting since NMF has a large number of free parameters to be estimated from very few observations when data are missing. The performance of the CNMF system with $\mathbf{S} = \mathbf{V}$ probably suffers from modelling error when very few data are missing – since the columns of $\mathbf{V}$ may not be able to combine into K components in a convex way. In the range $50 - 70\%$, its performance is similar to that of NMF. Beyond this range, it seems to be less prone to overfitting than NMF, probably due to less free parameters or to a regularization effect provided by $\mathbf{S} = \mathbf{V}$.

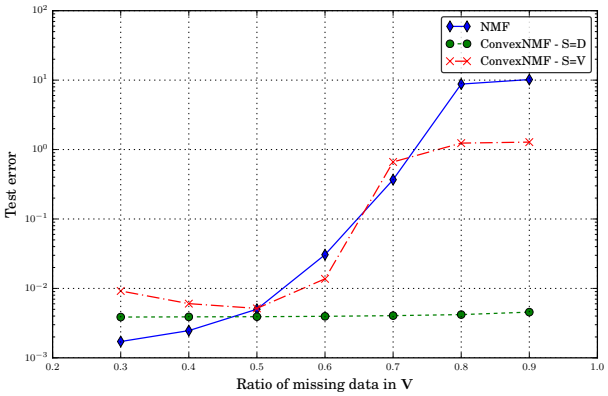

Fig. 4. Test error vs. missing data ratio in audio spectrograms.

We now investigate the influence of the "complexity" of the audio signal on the test error when the ratio of missing data is set to the high value 80%. Figure 5 displays, for each music recording, the test error with respect to the number of different pitches for all notes in the piano piece, which also equals the number of component K used by each system. CNMF with $\mathbf{S} = \mathbf{D}$ performs better than NMF whatever the number of notes and the error increases by a small factor along the represented range. NMF performs about five times worse for "easy" pieces, i.e., pieces composed of notes with about 6 different pitches and it performs about 10^4 times worse when the number of pitches is larger than 25. CNMF

[2] The code of the experiments is available on the webpage of the MAD project `http://mad.lif.univ-mrs.fr/`.

with $\mathbf{S} = \mathbf{V}$ performs slightly better than NMF. Since the number of components K increases equals the number of note pitches, those results confirm that NMF may highly suffer from overtraining while CNMF may not, being robust to missing data even for large values of K.

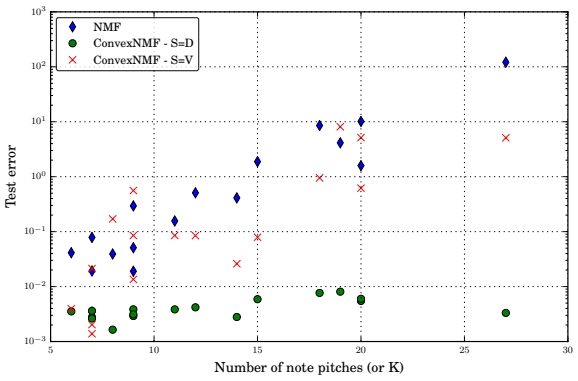

Fig. 5. Test error vs. number of different note pitches for 80% missing data (each dot represents one piece of music).

6. CONCLUSION

In this paper, we have proposed an extension of convex nonnegative matrix factorization in the case where missing data, as it has been previously presented for regular NMF. The proposed method can deal with missing values both in the data matrix $\mathbf{V}$ and in the dictionary $\mathbf{S}$, which is particularly useful in the case $\mathbf{S} = \mathbf{V}$ where the data is autoencoded. In this framework, an algorithm has been provided and analyzed using a Majorization-Minimization (MM) scheme to guarantee the convergence to a local minimum. A large set of experiments on synthetic data showed promising results for this variant of NMF for the task of reconstruction of missing data, and validated the value of this approach. In many situations, CNMF outperforms NMF, especially when the ratio of missing values is high and when the matrix data $\mathbf{V}$ is noisy. This trend has been confirmed on real audio spectrograms of piano music. In particular, we have shown how the use of a generic set of isolated piano notes as atoms could dramatically enhance the robustness to missing data.

This preliminary study indicates that it is worthy of further investigation, beyond the proposed settings where missing values are uniformly distributed over the matrix. Furthermore, the influence of missing values in the dictionary has not been completely assessed, as only the case where $\mathbf{S} = \mathbf{V}$ has been taken into account. On the application side, this approach could give new insight in many problems dealing with estimation of missing data.

7. REFERENCES

[1] C. Ding, Tao Li, and M.I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

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

[3] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[4] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[5] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc of SIGGRAPH*. ACM, 2000, pp. 417–424.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for spectral audio signals," in *Proc. of MLSP*, Grenoble, France, Sept. 2009.

[7] J. Le Roux, H. Kameoka, N. Ono, A. De Cheveigne, and S. Sagayama, "Computational auditory induction as a missing-data model-fitting problem with bregman divergence," *Speech Communication*, vol. 53, no. 5, pp. 658–676, 2011.

[8] D.L. Sun and R. Mazumder, "Non-negative matrix completion for bandwidth extension: A convex optimization approach," in *Proc. of MLSP*, Sept. 2013, pp. 1–6.

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

[10] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *Proc. of MLSP*, vol. 10, pp. 1, 2010.

[11] A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[12] C. Févotte, N. Bertin, and J-L Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[13] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1643–1654, 2010.

A. PROOFS

We detail the proofs of Propositions 1 and 2. The proof of Proposition 3 is straithforward using the same methodology.

We first recall preliminary elementss from [9]. The β-divergence $d_\beta(x|y)$ can be decomposed as a sum of a convex term, a concave term and a constant term with respect to its second variable y as

$$d_\beta(x|y) = \breve{d}_\beta(x|y) + \hat{d}_\beta(x|y) + \overline{d}_\beta(x) \tag{12}$$

This decomposition is not unique. We will use decomposition given in [9, Table 1], for which we have the following derivatives w.r.t. variable y:

$$\breve{d}'_\beta(x|y) \triangleq \begin{cases} -xy^{\beta-2} & \text{if } \beta < 1 \\ y^{\beta-2}(y-x) & \text{if } 1 \leqslant \beta \leqslant 2 \\ y^{\beta-1} & \text{if } \beta > 2 \end{cases} \tag{13}$$

$$\hat{d}'_\beta(x|y) \triangleq \begin{cases} y^{\beta-1} & \text{if } \beta < 1 \\ 0 & \text{if } 1 \leqslant \beta \leqslant 2 \\ -xy^{\beta-2} & \text{if } \beta > 2 \end{cases} \tag{14}$$

A.1. Update of S (Proof of Proposition 1)

We prove Proposition 1 by first constructing the auxiliary function (Proposition 4 below) and then focussing on its minimum (Proposition 5 below). Due to separability, the update of $\mathbf{S} \in \mathbb{R}_+^{F\times P}$ relies on the update of each of its columns. Hence, we only derive the update for a vector $\mathbf{s} \in \mathbb{R}_+^P$.

Definition 2 (Objective function $C_S(\mathbf{s})$)**.** For $\mathbf{v} \in \mathbb{R}_+^N, \mathbf{L} \in \mathbb{R}_+^{P\times K}, \mathbf{H} \in \mathbb{R}_+^{K\times N}, \mathbf{m} \in \{0,1\}^N, \mathbf{m}^o \in \{0,1\}^P, \mathbf{s} \in \mathbb{R}_+^P$, let us define

$$C_S(\mathbf{s}) \triangleq \sum_n m_n \left[\breve{C}_n(\mathbf{s}) + \hat{C}_n(\mathbf{s})\right] + \overline{C} \tag{15}$$

0where

$$\breve{C}_n(\mathbf{s}) \triangleq \breve{d}_\beta\left(v_n|\left[\mathbf{s}^T\mathbf{LH}\right]_n\right), \hat{C}_n(\mathbf{s}) \triangleq \hat{d}_\beta\left(v_n|\left[\mathbf{s}^T\mathbf{LH}\right]_n\right) \text{ and } \overline{C} \triangleq \overline{d}_\beta(\mathbf{m}.\mathbf{v}) + \lambda\overline{d}_\beta(\mathbf{m}^o.\mathbf{s}^o) \tag{16}$$

Proposition 4 (Auxiliary function $G_S(\mathbf{s}|\widetilde{\mathbf{s}})$ for $C_S(\mathbf{s})$)**.** *Let $\widetilde{\mathbf{s}} \in \mathbb{R}_+^P$ be such that $\forall n, \widetilde{v}_n > 0$ and $\forall p, \widetilde{s}_p > 0$, where $\widetilde{\mathbf{v}} \triangleq \left[\mathbf{s}^T\mathbf{LH}\right]^T$. Then the function $G_S(\mathbf{s}|\widetilde{\mathbf{s}})$ defined by*

$$G_S(\mathbf{s}|\widetilde{\mathbf{s}}) \triangleq \sum_n m_n \left[\breve{G}_n(\mathbf{s}|\widetilde{\mathbf{s}}) + \hat{G}_n(\mathbf{s}|\widetilde{\mathbf{s}})\right] + \overline{C} \tag{17}$$

where

$$\breve{G}_n(\mathbf{s}|\widetilde{\mathbf{s}}) \triangleq \sum_p \frac{[\mathbf{LH}]_{pn}\widetilde{s}_p}{\widetilde{v}_n}\breve{d}_\beta\left(v_n|\widetilde{v}_n\frac{s_p}{\widetilde{s}_p}\right) \text{ and } \hat{G}_n(\mathbf{s}|\widetilde{\mathbf{s}}) \triangleq \hat{d}_\beta(v_n|\widetilde{v}_n) + \hat{d}'_\beta(v_n|\widetilde{v}_n)\sum_p [\mathbf{LH}]_{pn}(s_p - \widetilde{s}_p). \tag{18}$$

is an auxiliary function for $C_S(\mathbf{s})$.

Proof. We trivially have $G_S(\mathbf{s}|\mathbf{s}) = C_S(\mathbf{s})$. We use the separability in n and in p in order to upper bound each convex term $\breve{C}_n(\mathbf{s})$ and each concave term $\hat{C}_n(\mathbf{s})$.

Convex term $\breve{C}_n(\mathbf{s})$. Let us prove that $\breve{G}_n(\mathbf{s}|\widetilde{\mathbf{s}}) \geqslant \breve{C}_n(\mathbf{s})$. Let $\mathcal{P}$ be the set of indices such that $[\mathbf{LH}]_{pn} \neq 0$ and define, for $p \in \mathcal{P}$,

$$\widetilde{\lambda}_{pn} \triangleq \frac{[\mathbf{LH}]_{pn}\widetilde{s}_p}{\widetilde{v}_n} = \frac{[\mathbf{LH}]_{pn}\widetilde{s}_p}{\sum_{p'\in\mathcal{P}}[\mathbf{LH}]_{p'n}\widetilde{s}_{p'}}. \tag{19}$$

We have $\sum_{p\in\mathcal{P}} \widetilde{\lambda}_{pn} = 1$ and

$$\breve{G}_n\left(\mathbf{s}|\widetilde{\mathbf{s}}\right) = \sum_{p\in\mathcal{P}} \widetilde{\lambda}_{pn} \breve{d}_\beta\left(v_n|\frac{[\mathbf{LH}]_{pn}\, s_p}{\widetilde{\lambda}_{pn}}\right) \geqslant \breve{d}_\beta\left(v_n|\sum_{p\in\mathcal{P}} \widetilde{\lambda}_{pn}\frac{[\mathbf{LH}]_{pn}\, s_p}{\widetilde{\lambda}_{pn}}\right) = \breve{d}_\beta\left(v_n|\sum_{p=1}^{P}[\mathbf{LH}]_{pn}\, s_p\right) = \breve{C}_n\left(\mathbf{s}\right) \tag{20}$$

Concave term $\widehat{C}_n\left(\mathbf{s}\right)$. We have $\widehat{G}_n\left(\mathbf{s}|\widetilde{\mathbf{s}}\right) \geqslant \widehat{C}_n\left(\mathbf{s}\right)$ since $\widehat{C}_n\left(\mathbf{s}\right)$ is concave and $\mathbf{s}\mapsto\widehat{G}_n\left(\mathbf{s}|\widetilde{\mathbf{s}}\right)$ is a tangent plane to $\widehat{C}_n\left(\mathbf{s}\right)$ in $\widetilde{\mathbf{s}}$:

$$\widehat{G}_n\left(\mathbf{s}|\widetilde{\mathbf{s}}\right) = \widehat{d}_\beta\left(v_n|\widetilde{v}_n\right) + \sum_p \widehat{d}'_\beta\left(v_n|\sum_{p'}[\mathbf{LH}]_{p'n}\,\widetilde{s}_{p'}\right)[\mathbf{LH}]_{pn}\left(s_p-\widetilde{s}_p\right) = \widehat{C}_n\left(\widetilde{\mathbf{s}}\right) + \left\langle\nabla\widehat{C}_n\left(\widetilde{\mathbf{s}}\right), \mathbf{s}-\widetilde{\mathbf{s}}\right\rangle \tag{21}$$

□

Proposition 5 (Minimum of $G_S\left(\mathbf{s}|\widetilde{\mathbf{s}}\right)$)**.** *The minimum of* $\mathbf{s}\mapsto G_S\left(\mathbf{s}|\widetilde{\mathbf{s}}\right)$ *subject to the constraint* $\mathbf{m}^o.\mathbf{s} = \mathbf{m}^o.\mathbf{s}^o$ *is reached at* $\mathbf{s}^{MM}$ *with*

$$\forall p, s_p^{MM} \triangleq \begin{cases} \widetilde{s}_p\left(\frac{\sum_n m_n \widetilde{v}_n^{\beta-2} v_n [\mathbf{LH}]_{pn}}{\sum_n m_n \widetilde{v}_n^{\beta-1}[\mathbf{LH}]_{pn}}\right)^{\gamma(\beta)} & \text{if } m_p^o = 0 \\ s_p^o & \text{if } m_p^o = 1. \end{cases} \tag{22}$$

Proof. Since variable s_p is fixed for p such that $m_p^o = 1$, we only consider variables s_p for p such that $m_p^o = 0$. The related penalty term in $G_S\left(\mathbf{s}|\widetilde{\mathbf{s}}\right)$ vanishes when $m_p^o = 0$. Using (13) and (14), the minimum is obtained by cancelling the gradient

$$\nabla_{s_p} G_S\left(\mathbf{s}|\widetilde{\mathbf{s}}\right) = \sum_n m_n \mathbf{LH}_{pn}\left[\breve{d}'_\beta\left(v_n|\widetilde{v}_n\frac{s_p}{\widetilde{s}_p}\right) + \widehat{d}'_\beta\left(v_n|\widetilde{v}_n\right)\right] \tag{23}$$

and by considering that the Hessian matrix is diagonal with nonnegative entries since $\breve{d}_\beta\left(x|y\right)$ is convex:

$$\nabla^2_{s_p} G_S\left(\mathbf{s}|\widetilde{\mathbf{s}}\right) = \sum_n m_n \widetilde{v}_n \frac{\mathbf{LH}_{pn}}{\widetilde{s}_p}\breve{d}''_\beta\left(v_n|\widetilde{v}_n\frac{s_p}{\widetilde{s}_p}\right) \geqslant 0. \tag{24}$$

□

A.2. Update of L (Proof of Proposition 2)

We proove Proposition 2 by first constructing the auxiliary function (Proposition 6 below) and then focussing on its minimum (Proposition 7 below). As opposed to the update of $\mathbf{S}$, no separability is considered here.

Definition 3 (Objective function $C_L\left(\mathbf{l}\right)$)**.** For $\mathbf{v}\in\mathbb{R}_+^F, \mathbf{S}\in\mathbb{R}_+^{F\times P}, \mathbf{H}\in\mathbb{R}_+^{K\times N}, \mathbf{M}\in\{0,1\}^{F\times N}, \mathbf{L}\in\mathbb{R}_+^{P\times K}$, let us define

$$C_L\left(\mathbf{L}\right) \triangleq \sum_{fn} m_{fn}\left[\breve{C}_{fn}\left(\mathbf{L}\right) + \widehat{C}_{fn}\left(\mathbf{L}\right)\right] + \overline{C} \tag{25}$$

where

$$\breve{C}_{fn}\left(\mathbf{L}\right) \triangleq \breve{d}_\beta\left(v_{fn}|\left[\mathbf{SLH}\right]_{fn}\right), \widehat{C}_{fn}\left(\mathbf{L}\right) \triangleq \widehat{d}_\beta\left(v_{fn}|\left[\mathbf{SLH}\right]_{fn}\right) \text{ and } \overline{C} \triangleq \overline{d}_\beta\left(\mathbf{M}.\mathbf{V}\right). \tag{26}$$

Proposition 6 (Auxiliary function $G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right)$ for $C_L\left(\mathbf{L}\right)$)**.** *Let* $\widetilde{\mathbf{L}}\in\mathbb{R}_+^{P\times K}$ *be such that* $\forall f,n, \widetilde{\mathbf{V}}_{fn} > 0$ *and* $\forall p,k, \widetilde{\mathbf{L}}_{pk} > 0$*, where* $\widetilde{\mathbf{V}} \triangleq \mathbf{S}\widetilde{\mathbf{L}}\mathbf{H}$*. Then the function* $G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right)$ *defined by*

$$G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \triangleq \sum_{fn} m_{fn}\left[\breve{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) + \widehat{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right)\right] + \overline{C} \tag{27}$$

where

$$\breve{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \triangleq \sum_{pk} \frac{s_{fp}\widetilde{l}_{pk}h_{kn}}{\widetilde{v}_{fn}} \breve{d}_{\beta}\left(v_{fn}|\widetilde{v}_{fn}\frac{l_{pk}}{\widetilde{l}_{pk}}\right) \tag{28}$$

$$\widehat{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \triangleq \left[\widehat{d}_{\beta}\left(v_{fn}|\widetilde{v}_{fn}\right) + \widehat{d}'_{\beta}\left(v_{fn}|\widetilde{v}_{fn}\right)\sum_{pk} s_{fp}h_{kn}\left(l_{pk}-\widetilde{l}_{pk}\right)\right] \tag{29}$$

is an auxiliary function for $C_L\left(\mathbf{L}\right)$.

Proof. We trivially have $G_L\left(\mathbf{L}|\mathbf{L}\right) = C_L\left(\mathbf{L}\right)$. In order to prove that $G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \geqslant C_L\left(\mathbf{L}\right)$, we use the separability in f and n and we upper bound the convex terms $\breve{C}_{fn}\left(\mathbf{L}\right)$ and the concave terms $\widehat{C}_{fn}\left(\mathbf{L}\right)$.
Convex term $\breve{C}_{fn}\left(\mathbf{L}\right)$. Let us prove that $\breve{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \geqslant \breve{C}_{fn}\left(\mathbf{L}\right)$. Let $\mathcal{P}$ be the set of indices such that $s_{fp} \neq 0$, $\mathcal{K}$ be the set of indices such that $h_{kn} \neq 0$ and define, for $(p,k) \in \mathcal{P} \times \mathcal{K}$,

$$\widetilde{\lambda}_{fpkn} \triangleq \frac{s_{fp}\widetilde{l}_{pk}h_{kn}}{\widetilde{v}_{fn}} = \frac{s_{fp}\widetilde{l}_{pk}h_{kn}}{\sum_{(p',k')\in\mathcal{P}\times\mathcal{K}} s_{fp'}\widetilde{l}_{p'k'}h_{k'n}}. \tag{30}$$

We have $\sum_{(p,k)\in\mathcal{P}\times\mathcal{K}} \widetilde{\lambda}_{fpkn} = 1$ and

$$\breve{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) = \sum_{(p,k)\in\mathcal{P}\times\mathcal{K}} \widetilde{\lambda}_{fpkn}\breve{d}_{\beta}\left(v_{fn}|\frac{s_{fp}l_{pk}h_{kn}}{\widetilde{\lambda}_{fpkn}}\right) \tag{31}$$

$$\geqslant \breve{d}_{\beta}\left(v_{fn}|\sum_{(p,k)\in\mathcal{P}\times\mathcal{K}} \widetilde{\lambda}_{fpkn}\frac{s_{fp}l_{pk}h_{kn}}{\widetilde{\lambda}_{fpkn}}\right) = \breve{d}_{\beta}\left(v_{fn}|\sum_{p=1}^{P}\sum_{k=1}^{K} s_{fp}l_{pk}h_{kn}\right) = \breve{C}_{fn}\left(\mathbf{L}\right) \tag{32}$$

Concave term $\widehat{C}_{fn}\left(\mathbf{L}\right)$. We have $\widehat{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) \geqslant \widehat{C}_{fn}\left(\mathbf{L}\right)$ since $\widehat{C}_{fn}\left(\mathbf{L}\right)$ is concave and $\mathbf{L} \mapsto \widehat{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right)$ is a tangent plane to $\widehat{C}_{fn}\left(\mathbf{L}\right)$ in $\widetilde{\mathbf{L}}$:

$$\widehat{G}_{fn}\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) = \widehat{d}_{\beta}\left(v_{fn}|\widetilde{v}_{fn}\right) + \sum_{pk} \widehat{d}'_{\beta}\left(v_{fn}|\sum_{p'k'} s_{fp'}\widetilde{l}_{p'k'}h_{k'n}\right) s_{fp}h_{kn}\left(l_{pk}-\widetilde{l}_{pk}\right) \tag{33}$$

$$= \widehat{C}_{fn}\left(\widetilde{\mathbf{L}}\right) + \left\langle \nabla\widehat{C}_{fn}\left(\widetilde{\mathbf{L}}\right), \mathbf{L}-\widetilde{\mathbf{L}}\right\rangle \tag{34}$$

□

Proposition 7 (Minimum of $G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right)$)**.** *The minimum of* $\mathbf{L} \mapsto G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right)$ *is reached at* $\mathbf{L}^{MM}$ *with*

$$\forall p,k, l_{p,k}^{MM} \triangleq \begin{cases} l_{p,k}\left(\frac{\sum_{fn} s_{fp}m_{fn}\widetilde{v}_{fn}^{\beta-2}v_{fn}h_{kn}}{\sum_{fn} s_{fp}m_{fn}\widetilde{v}_{fn}^{\beta-1}h_{kn}}\right)^{\gamma(\beta)} & \text{if } \mathbf{M} \neq \mathbf{0} \\ l_{p,k} \text{ otherwise.} \end{cases} \tag{35}$$

Proof. Using (13) and (14), the minimum is obtained by cancelling the gradient

$$\nabla_{l_{pk}} G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) = \sum_{fn} m_{fn}s_{fp}h_{kn}\left[\breve{d}'_{\beta}\left(v_{fn}|\widetilde{v}_{fn}\frac{l_{pk}}{\widetilde{l}_{pk}}\right) + \widehat{d}'_{\beta}\left(v_{fn}|\widetilde{v}_{fn}\right)\right] \tag{36}$$

and by considering that the Hessian matrix is diagonal with nonnegative entries since $\breve{d}_{\beta}\left(x|y\right)$ is convex:

$$\nabla^2_{l_{pk}} G_L\left(\mathbf{L}|\widetilde{\mathbf{L}}\right) = \sum_{fn} m_{fn}\widetilde{v}_{fn}\frac{s_{fp}h_{kn}}{\widetilde{l}_{pk}}\breve{d}''_{\beta}\left(v_{fn}|\widetilde{v}_{fn}\frac{l_{pk}}{\widetilde{h}_{pk}}\right) \geqslant 0. \tag{37}$$

□