Algorithms for Nonnegative Tucker problems

Valentin Leplat, Jeremy E. Cohen, ??

1 Paper content

- intro - parcimonie - beta= 2 HALS - optimi dure
todo: - codes synthé > Valentin - examples > jérémy

2 Litterature on NTD

2.1 Model properties

- Zhou Cichocki 2014 Uniqueness
- Cohen 2017, Skau 2022 Compression (nnCANDELIND)
- (nice paper!, first paper on full nonnegative NTD and algorithms) Morup 2008 Sparse NTD (MU) + interpretability
- Semi-supervised NTD (graph regularized, label propagation) (Qiu 2021, Li 2017)
-

2.2 Applications

- 3D skeleton classification (Li 2021)
- Brain source separation (Morup 2008)
- Music factorization and segmentation (Smith 2019, Marmoret 2020)
- Blind Unmixing (Sun 2021), HSI superresolution (Zare, Kazemi 2021), HSI compression (Li 2017)
- Neuroimaging EEG (Phan 2010, Dao 2018, Rostakova 2020)
- patient Phenotyping (Yang 2019)

2.3 Algorithms from NMF and sparse/orthogonal NMF

- Paatero 1998
- Lee and Song 1999, first MU, 2000 proof convergence
- Choi 2008 Orthogonal NMF
- Fevotte xxx (majorante jointe + papier avec Jerome)
- Gillis xxx (2011 accHALS, 2012 “sparse and unique” sparse HALS with max=1 normalization, 2018 HER HALS)
- (help Valentin!) Sparse things
  - Normalization (LeRoux xx)
  - Mixed sparsity objective (Hoyer 2002 2004)
  - l1 et max=1 (Gillis 2012 sparse and unique)
  -

2.4 Algorithms for NTD and CPD

- Xu Yin 2016 APG
- ANLS HALS (Cichocki 2011) > better version with nicolas' HALS but not paper on it > cover beta=2 more specifically as well ! Already used in nn-fac toolbox but never explained.
- Sparse TD parallel Kaya, Ucar 2015
- Tucker Sketching Tropp 2019
- Zdunek 2011 and Junjun Pan 2021 Orthogonal NTD
- (NTD !!) Kim Choi 2007 (also below)
- Incremental (Online) NTD Zdunek 2022

2.5 Non-euclidean losses in Tensor Decompositions

- Hong, Kolda GCP
- Vandecapelle, Vervliet 2019, 2020
- (NTD !!) Kim Choi 2007
- Pu, Xiao Fu 2021 stochastic mirror descent
- Ghalamkari, Sugiyama 2021 Mean-Field (not nonnegative)

3 Some insights on sparsity and normalization

Consider the toy problem

$$\underset{X\in\mathbb{R}^{m\times r}, Y\in\mathbb{R}^{r\times n}}{\operatorname{argmin}} f(XY) + \mu_x\|X\|_1 + \frac{1}{2}\mu_y\|Y\|_F^2 \tag{1}$$

where f is some cost function, and $\mu_{x,y}$ are nonnegative regularization parameters. We can be more general than this using arbitrary regularizations, but let's keep it simple for now. Let us denote $\phi(X,Y)$ the cost function, and we surcharge the notation with $\phi(X,Y,\mu_x,\mu_y)$ abusively when hyperparameters may vary.

We make a few observations:

- If $\mu_y = 0$, then because $f(XY)$ is invariant to a scaling of columns (resp. rows) of X (resp. Y), we have for any couple X, Y and any $\lambda < 1$

  $$\phi(\lambda X, \frac{1}{\lambda}Y) = f(XY) + \mu_x\lambda\|X\|_1 < \phi(X,Y)\ . \tag{2}$$

  This means that the cost can always be decreased by scaling down the columns of X, which implies that a global solution must be $X = 0$ which is however impossible. In other words this problem is degenerate and does not admit good NMF solutions, let alone sparse solutions. In practice the regularization term is will grow extremely small and the problem becomes simply $\operatorname{argmin}_{X,Y} f(XY)$.

- Because of the scaling ambiguity, it was shown in [][Roald 2021 todo] that we can fix $\mu_x = \mu_y$ without loss of generality. Indeed, for any positive u,

  $$\phi(X,Y,\mu,\mu) = f(XY) + \mu\|X\|_1 + \frac{\mu}{2}\|Y\|_F^2 = f(\frac{X}{u}uY) + \mu u\|\frac{X}{u}\|_1 + \frac{\mu}{2u^2}\|uY\|_F^2 = \phi(\tilde{X},\tilde{Y},\mu u,\frac{\mu}{u^2}) \tag{3}$$

  where $\tilde{X}$ and $\tilde{Y}$ are new, scaled variable. We then see that

  $$\min_{X,Y}\ \phi(X,Y,\mu,\mu) = \min_{X,Y}\ \phi(\tilde{X},\tilde{Y},\mu u,\frac{\mu}{u^2}) = \min_{\tilde{X},\tilde{Y}}\ \phi(\tilde{X},\tilde{Y},\mu u,\frac{\mu}{u^2}) = \min_{X,Y}\ \phi(X,Y,\mu u,\frac{\mu}{u^2})\ . \tag{4}$$

  Because any couple (μ_x,μ_y) can be written as $(\mu u, \frac{\mu}{u^2})$, setting $\mu_x = \mu_y = \mu$ does not change the minimum of the cost (but we do change the argmin). Alternatively, we can fix one of the two regularization parameters arbitrarily.

- When $\mu_x = \mu_y = \mu$, it can be shown that the solution (X^*, Y^*) must satisfy $\|X^*\|_1 = \|Y^*\|_F^2$. Indeed, fix $a = \mu\|X^*\|_1, b = \mu\|Y^*\|_F^2$, and minimize the one-dimensional cost $\phi(\lambda) := f(\frac{X}{\lambda}\lambda Y) + a\lambda + \frac{b}{\lambda^2}$ for positive lambda. It can be shown easily that $\lambda^* = (\frac{b}{a})^{1/3}$. However, if λ^* is not equal to one, it means that $\phi(X, Y)$ can be reduced by scaling and thus X^* and Y^* are not solutions. By contraposition we must have $a = b$ which yields $\|X^*\|_1 = \|Y^*\|_F^2$. This is very interesting because it means that we can balance the terms in the decomposition using penalisations on each term. In fact using the same reasoning with columnwise scaling ambiguity yields the stronger result $\|X_i^*\|_1 = \|Y_i^*\|_2^2$ for any column/row index i. This also means that we should in fact not use $\mu_x = \mu_y$ because we will not be able to decrease the norm of X without also decreasing Y which may heavily biais the fitting term $f(XY)$. Rather, a sound strategy is to fix $\mu_y = 1$ and only tune μ_x. We will then have at optimality $\mu_x\|X_i^*\|_1 = \|Y_i^*\|_2^2$ which allows to scale the ℓ_1 norm of columns of X as desired.

4 Algorithms for Sparse beta-divergence NTD

In this section we propose various algorithms able to compute a candidate solution to approximate Nonnegative Tucker Decomposition (NTD) with β-divergence as a loss function and mixed-sparsity regularizations on the factors and the core tensor:

$$\underset{W\geq 0, H\geq 0, Q\geq 0, \mathcal{G}\geq 0}{\operatorname{argmin}} D_\beta(\mathcal{X}|\mathcal{G} \times_1 W \times_2 H \times_3 Q) + \mu_{\mathcal{G}}\|\mathcal{G}\|_1 + \frac{1}{2}\left(\mu_W\|W\|_F^2 + \mu_H\|H\|_F^2 + \mu_Q\|Q\|_F^2\right) \tag{5}$$

with $D_\beta(.|.)$ the element-wise β-divergence between two tensors, and $\mu_{\mathcal{G}}$, μ_W, μ_H and μ_Q are positive scalars denoting the penalty weights associated to each sparsity regularization. For the later, those will be further referred to as penalty functions.

The objective function in (5) is non-convex jointly in $\{W, H, Q, \mathcal{G}\}$. Moreover, computing a global solution to NTD is NP-Hard since NTD generalizes NMF [4,6]. Hence most algorithms developed to solve NMF and therefore NTD optimization problems are based on iterative local optimization schemes converging to local solutions. For such optimization problems, it is usually easier to optimize over one factor (or one mode for NTD) given the others terms are known and fixed. Indeed, the obtained subproblems when fixing all but one mode is convex as long as $\beta \in [1, 2]$. For this reason many of the algorithms developed to tackle (5) or its NMF variants rely on Block Coordinate Descent (BCD) schemes and we adopt this strategy in this paper.

Furthermore, this work extends previous works that have been carried out to tackle NTD with β-divergence as loss function. Indeed, the seminal paper by Lee and Seung [2] proposed the first alternating algorithm for NMF with β-divergence but without penalty functions. These algorithms have been later revisited by [1] and [3] (this one integrates penalty functions and equality constraints) and finally extended to non-penalized β-divergence NTD in [4]. One of the core ideas in [4] is the fact that the NTD model can be rewritten using tensor matricization along the different modes. For instance, along the first mode we have:

$$\begin{aligned} &\mathcal{X} = \mathcal{G} \times_1 W \times_2 H \times_3 Q \\ &\Leftrightarrow \mathcal{X}_{(1)} = W\mathcal{G}_{(1)}\left(H \boxtimes Q\right)^T \end{aligned} \tag{6}$$

where $\mathcal{X}_{(i)}$ is the matricization of the tensor $\mathcal{X}$ along the mode i and $\boxtimes$ denotes the Kronecker product. The matricizations are analogous for factors H and Q. Hence, Equation (6) can be interpreted as an NMF of $\mathcal{X}_{(i)}$ with factors W and $\mathcal{G}_{(1)}\left(H \boxtimes Q\right)^T$. This observation led the authors in [4] to develop Multiplicative Updates based algorithms to derive the updates of each factor of NTD with β-divergence loss function. In this paper, we follow similar approaches.

The two next sections respectively present the updates for the factors W, H and Q and the core tensor $\mathcal{G}$ to tackle (5).

4.1 Updates of the factors

In this section we derive updates for factors W, H, Q. Since all the subproblems have similar structure, we will only focus on the subproblem in W without loss of generality.

For clarity, let us start by considering the matricization of the tensor $\mathcal{X}$ along the first mode given by Equation (6) and pose $V := \mathcal{X}_{(1)} \in \mathbb{R}_+^{F\times N}$ and $U := \mathcal{G}_{(1)}\left(H \boxtimes Q\right)^T \in \mathbb{R}_+^{K\times N}$. Note that $\mathcal{G}_{(1)}\left(H \boxtimes Q\right)^T$ can be computed based on the following identity: $\mathcal{G}_{(1)}\left(H \boxtimes Q\right)^T = \left(\mathcal{G} \times_2 H \times_3 Q\right)_{(1)}$. The subproblem in W is defined as follows:

$$\underset{W \geq 0}{\operatorname{argmin}} \; D_\beta(V|WU) + \frac{1}{2}\mu_W \|W\|_F^2 \tag{7}$$

The objective function in Problem (7) is separable with respect to the rows of W, that is $D_\beta(V|WU) + \frac{1}{2}\mu_W\|W\|_F^2 = \sum_f^F \left[D_\beta(v_f|w_f U) + \frac{1}{2}\mu_W\|w_f\|_F^2\right]$, hence can be minimized over the F rows of W independently. For the following, we focus on the minimization over one particular row f of W leading to the following optimization problem:

$$\underset{w \geq 0}{\operatorname{argmin}} \; D_\beta(v|wU) + \frac{1}{2}\mu_W \|w\|_F^2 \tag{8}$$

where the selected subscript f has been dropped for clarity purposes and $D_\beta(v|wU) = \sum_n d_\beta(v_n|\,[wU]_n)$ with the discrete β-divergence denoted $d_\beta(x|y)$ and equal to

$$\begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta x y^{\beta-1}\right) \text{ for } \beta \neq 0, 1, \\ x\log\frac{x}{y} - x + y \text{ for } \beta = 1, \\ \frac{x}{y} - \log\frac{x}{y} - 1 \text{ for } \beta = 0. \end{cases}$$

For $\beta = 2$, this the standard squared Euclidean distance, that is, the squared Frobenius norm $||V - WU||_F^2$. For $\beta = 1$ and $\beta = 0$, the β-divergence corresponds to the Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence, respectively.

In order to tackle Problem (8), we follow the Majorization-Minimization (MM) framework. Let us briefly recall the high-level ideas to derive updates via MM. Let us consider the general problem

$$\min_{w \in \mathcal{W}} f(w).$$

Given an initial iterate $\widetilde{w} \in \mathcal{W}$, MM generates a new iterate $\hat{w} \in \mathcal{W}$ that is guaranteed to decrease the objective function, that is, $f(\hat{w}) \leq f(\widetilde{w})$. To do so, it uses the following two steps:

- Majorization: find a function that is an upper approximation of the objective and is tight at the current iterate, which is referred to as a majorizer. More precisely find a function $g(w|\widetilde{w})$ such that

$$(i)\ g(\widetilde{w}|\widetilde{w}) = f(\widetilde{w}) \quad \text{and} \quad (ii)\ g(w|\widetilde{w}) \geq f(w) \text{ for all } w \in \mathcal{W}.$$

- Minimization: minimize the majorizer, that is, solve $\min_{w\in\mathcal{W}} g(w|\widetilde{w})$ approximately or exactly, to obtain the next iterate $\hat{w} \in \mathcal{W}$ which is such that (iii) $g(\hat{w}|\widetilde{w}) \leq g(\widetilde{w}|\widetilde{w})$. This guarantees the objective function to decrease at each step of this iterative process since

$$f(\hat{w}) \underset{(ii)}{\leq} g(\hat{w}|\widetilde{w}) \underset{(iii)}{\leq} g(\widetilde{w}|\widetilde{w}) \underset{(i)}{=} f(\widetilde{w}).$$

The updates are obtained using MM where the majorizer g is chosen separable, that is, $g(w|\widetilde{w}) = \sum_{i=1} g_i(w_i|\widetilde{w}_i)$ for some well chosen univariate functions g_i's; discussed later in the next section. This choice typically makes the minimization of g admit a closed-form solution.

In Problem (8), the second term is separable w.r.t. each entry of w. For the data fitting term, we use the majorizer proposed in [1]. For the sake of completeness, we briefly recall it in the following. It consists in majorizing the convex part of the β-divergence using Jensen's inequality and majorizing the concave part by its tangent (first-order Taylor approximation). We have

$$d_\beta(x|y) = \check{d}_\beta(x|y) + \hat{d}_\beta(x|y) + \bar{d}_\beta(x|y), \tag{9}$$

where $\check{d}$ is convex function of y, $\hat{d}$ is a concave function of y and $\bar{d}$ is a constant of y; see Table 1.

Table 1: Differentiable convex-concave-constant decomposition of the β-divergence under the form (9) [1].

	$\check{d}(x\|y)$	$\hat{d}(x\|y)$	$\bar{d}(x)$
$\beta = 0$	xy^{-1}	$\log(y)$	$x(\log(x) - 1)$
$\beta \in [1, 2]$	$d_\beta(x\|y)$	0	0

Lemma 1 ([1]) *Let $\tilde{v} = \tilde{w}U$ and $\tilde{w}$ be such that $\tilde{v_n} > 0$ for all n and $\tilde{w_k} > 0$ for all k. Then the function*

$$\begin{aligned} G(w|\tilde{w}) = \sum_n \left[\sum_k \frac{\tilde{w_k} u_{kn}}{\tilde{v_n}} \check{d}(v_n|\tilde{v_n}\frac{w_k}{\tilde{w_k}}) \right] + \bar{d}(v_n) \\ + \left[\check{d}'(v_n|\tilde{v_n}) \sum_k (w_k - \tilde{w_k}) u_{kn} + \hat{d}(v_n|\tilde{v_n}) \right] \end{aligned} \tag{10}$$

is a majorizer for $\sum_n d(v_n| [wU]_n)$ at $\tilde{w}$.

Finally the problem we need to solve has the form:

$$\underset{w \geq 0}{\text{argmin}}\ G(w|\tilde{w}) + \frac{1}{2}\mu_W \|w\|_F^2 \tag{11}$$

Optimization problem (11) is a particular instance of problems covered by framework proposed in [3] in the case no additional equality constraints are required. More formally, assuming $\mu_W > 0$, Problem (11) satisfies:

- Assumption 1 from [3] since the penalty function $\Phi(w) = \frac{1}{2}\|w\|_F^2$ is lower bounded on the feasible set and admits a particular upper approximation at any current iterate $\widetilde{w}$:

$$\Phi(w) \leq \Phi(\widetilde{w}) + \langle \nabla\Phi(\widetilde{w}), w - \widetilde{w} \rangle + \sum_k \frac{L_k}{2}(w_k - \widetilde{w}_k)^2 \tag{12}$$

  assuming $L_k \geq 1 > 0$ for all k.

- first setting covered by Proposition 1 [3], namely coefficients $a_k = \mu_W \frac{L_k}{2}$ are strictly positive for all k.

Therefore, by Proposition 1 from [3], there exists a unique minimizer of (11) in $(0, \infty)$. In the following, we illustrate this theoretical result by computing the closed form expression of this minimizer. Note that the closed form expression can be derived if $\beta \in \{0, 1, 3/2, 2\}$ according to [3]. Outside this set of values for β, an iterative scheme such as Newton-Raphson is required to compute the minimizer. Indeed, we will see further that the task of computing the minimizer is equivalent of finding the positive real roots of a set of monovariate polynomial equations in each entry w_k of w.

To compute the positive minimizer of (11), we are looking for $w \in \mathbb{R}^K$ that cancels the gradient of objective function $G(w|\tilde{w}) + \frac{1}{2}\mu_W\|w\|_F^2$. Since the objective function is separable w.r.t. each entry w_k, we focus on solving:

$$\text{find } \hat{w_k} \text{ such that } \nabla_{w_k}\left[G(w|\tilde{w}) + \frac{1}{2}\mu_W \|w\|_F^2 \right] = 0 \tag{13}$$

As mentioned previously, the next steps depend on the value chosen for β. In the following, we consider the particular cases $\beta \in \{1, 3/2, 2\}$ for which a closed form of the minimizer $\hat{w}$ can be derived, with a particular emphasis on the case $\beta = 1$. Similar approach can be followed for the case $\beta = 0$ and is detailed in Appendix 4.2.

Based on Lemma 1 and Table 1 for $\beta \in [1, 2]$, we have:

$$\begin{aligned} \nabla_{w_k} G(w|\tilde{w}) &= \sum_n \frac{\tilde{w_k} u_{kn}}{\tilde{v_n}} \nabla_{w_k} d_\beta(v_n|\tilde{v_n}\frac{w_k}{\tilde{w_k}}) \\ &= \sum_n u_{kn} \left[\left(\tilde{v_n}\frac{w_k}{\tilde{w_k}}\right)^{\beta-1} - v_n \left(\tilde{v_n}\frac{w_k}{\tilde{w_k}}\right)^{\beta-2} \right] \end{aligned} \tag{14}$$

since $\nabla_y d_\beta(x|y) = y^{\beta-1} - xy^{\beta-2}$. Equation (13) becomes:

$$\begin{aligned} &\nabla_{w_k}\left[G(w|\tilde{w}) + \frac{1}{2}\mu_W \|w\|_F^2 \right] = 0 \\ &\Leftrightarrow a w_k + b w_k^{\beta-1} - c w_k^{\beta-2} = 0 \\ &\Leftrightarrow a w_k^{3-\beta} + b w_k - c = 0 \text{ since } \hat{w_k} \in (0, \infty). \end{aligned} \tag{15}$$

where:

- $a = \mu_W > 0$ by hypothesis,
- $b = \sum_n u_{kn} \left(\frac{\tilde{v_n}}{\tilde{w_k}}\right)^{\beta-1} \geq 0$ given that $\tilde{v_n}$, $\tilde{w_k}$ and u_{kn} are nonnegative for all n and k.
- $c = \sum_n u_{kn} v_n \left(\frac{\tilde{v_n}}{\tilde{w_k}}\right)^{\beta-2} \geq 0$ given that v_n, $\tilde{v_n}$, $\tilde{w_k}$ and u_{kn} are nonnegative for all n and k.

Therefore, computing the k-th entry of the minimizer $\hat{w}$ is equivalent to finding the root of a monovariate polynomial equation of degree $3-\beta$ (for $\beta \in [1,2]$) in w_k. Further, we focus on the particular case $\beta = 1$.

For $\beta = 1$, Equation (15) becomes $aw_k^2 + bw_k - c = 0$. The positive (real) root is computed as follows:

$$\hat{w_k} = \frac{\sqrt{(\sum_n u_{kn})^2 + 4\mu_W \tilde{w_k} \sum_n u_{kn} \frac{v_n}{\tilde{v_n}}} - \sum_n u_{kn}}{2\mu_W} \tag{16}$$

with $\mu_W > 0$. Note that although the closed-form expression in Equation (16) has a negative term in the numerator of the right-hand side, it can be easily checked that it always remains nonnegative given v_n, u_{kn} and $\tilde{w_k}$ are nonnegative for all k, n. Equation (16) can be expressed in matrix form as follows:

$$\hat{W} = \frac{\left[C^{.2} + S\right]^{.\frac{1}{2}} - C}{2\mu_W} \tag{17}$$

where $C = eU^T$ with e is a all-one matrix of size F-by-N and $S = 4\mu_W \tilde{W} \odot \left(\frac{[V]}{[\tilde{W}U]} U^T\right)$ with $A \odot B$ (resp. $\frac{[A]}{[B]}$) is the Hadamard product (resp. division) between A and B and $A^{(.\alpha)}$ is the element-wise α exponent of A.

Some important insights are discussed here-under:

- Computing the limit $\lim_{\mu_W \to 0} \hat{W}(\mu_W)$ makes Equation (17) tends to the original Multiplicative Updates introduced by Lee and Seung [2]. Indeed let us compute this limit in the scalar case from Equation (16) and pose $\alpha = \sum_n u_{kn}$ and $\eta = \tilde{w_k} \sum_n u_{kn} \frac{v_n}{\tilde{v_n}}$ for convenience:

$$\begin{aligned} \lim_{\mu_W \to 0} \frac{\sqrt{\alpha^2 + 4\mu_W \eta} - \alpha}{2\mu_W} &= \frac{\text{"0"}}{\text{"0"}} \\ &\underset{\mathrm{H}}{=} \lim_{\mu_W \to 0} \frac{\frac{\partial}{\partial_{\mu_W}} \sqrt{\alpha^2 + 4\mu_W \eta} - \alpha}{\frac{\partial}{\partial_{\mu_W}} 2\mu_W} \\ &= \lim_{\mu_W \to 0} (\alpha^2 + 4\mu_W \eta)^{-\frac{1}{2}} \eta = \frac{\eta}{\alpha} = \tilde{w_k} \frac{\sum_n u_{kn} \frac{v_n}{\tilde{v_n}}}{\sum_n u_{kn}} \end{aligned} \tag{18}$$

  In matrix form we have:

$$\lim_{\mu_W \to 0} \hat{W}(\mu_W) = \tilde{W} \odot \frac{\left[\frac{[V]}{[\tilde{W}U]} U^T\right]}{[eU^T]} \tag{19}$$

  which are the MU proposed by Lee and Seung for $\beta = 1$.

- Regarding the analysis of monovariate polynomial equation given in (15), interestingly the existence of a unique positive real root could have been also established by using Descartes rules of sign. Indeed we can count the number of real positive roots that $p(w_k) = aw_k^{3-\beta} + bw_k - c$ has (for $\beta \in [1,2]$). More specifically, let v be the number of variations in the sign of the coefficients a, b, c (ignoring coefficients that are zero). Let n_p be the number of real positive roots. Then:

  1. $n_p \leq v$,
  2. $v - n_p$ is an even integer.

  Let us consider the case $\beta = 1$ as an example: given the polynomial $p(w_k) = aw_k^2 + bw_k - c$, assuming that c is positive. Then $v = 1$, so n_p is either 0 or 1 by rule 1. But by rule 2, $v - n_p$ must be even, hence $n_p = 1$. Similar conclusion can be made for any $\beta \in [1,2]$.

As mentioned earlier, similar rationale can be followed to derive updates in closed-form for $\beta = 3/2$ or $\beta = 2$. The derivation of the updates in the case $\beta = 0$ is detailed in Appendix A since the polynomial equation differs from the one given in Equation (15). For other values of β, a numerical scheme will be necessary to compute the real positive root of the obtained polynomial equations. We can cite the Newton's method, the Muller's method and the the procedure developed in [5] which is based on the explicit calculation of the intermediary root of a canonical form of cubic. This procedure is suited for providing highly accurate numerical results in the case of badly conditioned polynomials.

In the next section we detail the updates for the core tensor $\mathcal{G}$.

4.2 Update of the core tensor

For the core tensor, we start by using the vectorization property defined as follows:

$$\operatorname{vec}(\mathcal{X}) = (W \boxtimes H \boxtimes Q) \operatorname{vec}(\mathcal{G}) \tag{20}$$

The updates for the core tensor follows the approach detailed in Section 4.1 for factors. Let us pose $U := (W \boxtimes H \boxtimes Q) \in \mathbb{R}^{F \times K}$, $v := \operatorname{vec}(\mathcal{X}) \in \mathbb{R}^F$ and $g := \operatorname{vec}(\mathcal{G}) \in \mathbb{R}^K$. The subproblem in g is defined as follows:

$$\operatorname*{argmin}_{g \geq 0} D_\beta(v|Ug) + \mu_{\mathcal{G}} \|g\|_1 \tag{21}$$

Again we follow the MM framework, the final problem we need to solve has the form:

$$\operatorname*{argmin}_{g \geq 0} G(g|\tilde{g}) + \mu_{\mathcal{G}} \|g\|_1 \tag{22}$$

where $\|g\|_1 = \sum_k g_k$ since $g \geq 0$. Again Problem (22) is a particular instance of problems covered by framework proposed in [3] in the case no additional equality constraints are required. According to Proposition 1 from [3], there exists a unique minimizer of (22) in $(0, \infty)$ as soon as $\beta < 2$. Indeed, we will illustrate later that a positive real minimizer is not guaranteed to exist when $\beta \geq 2$. Moreover, according to [3], it is possible to derive closed-form expressions for the minimizer of Problem (22) for the following values of β: $\beta \in (-\infty, 1] \cup \{5/4, 4/3, 3/2\}$. Outside these values for β, an iterative scheme is required to numerically compute the minimizer. To be compliant with the developments presented in Section 4.1, we will consider the interval $\beta \in [1, 2)$ and more particularly the case $\beta = 1$. For this setting, we can show that computing the minimizer of Problem (22) corresponds to solving:

$$\begin{aligned} &\nabla_{g_k} \left[G(g|\tilde{g}) + \mu_{\mathcal{G}} \|g\|_1 \right] = 0 \\ &\Leftrightarrow a h_k^{2-\beta} + b h_k - c = 0 \text{ since } \hat{g_k} \in (0, \infty). \end{aligned} \tag{23}$$

where:

- $a = \mu_{\mathcal{G}} > 0$ by hypothesis,
- $b = \sum_f u_{fk} \left(\frac{\tilde{v_f}}{\tilde{g_k}} \right)^{\beta - 1} \geq 0$ given u_{fk}, $\tilde{v_f} = [Ug]_f$ and $\tilde{g_k}$ nonnegative for all f, k.
- $c = \sum_f u_{fk} v_f \left(\frac{\tilde{v_f}}{\tilde{g_k}} \right)^{\beta - 2} \geq 0$ given u_{fk}, v_f, $\tilde{v_f}$ and $\tilde{g_k}$ nonnegative for all f, k.

For $\beta = 1$, Equation (23) becomes $a g_k + b g_k - c = 0$. The positive (real) root is computed as follows:

$$\begin{aligned} \hat{g_k} &= \frac{c|_{\beta=1}}{a + b|_{\beta=1}} \\ &= \tilde{g_k} \frac{\sum_f u_{fk} \frac{v_f}{\tilde{v_f}}}{\mu_{\mathcal{G}} + \sum_f u_{fk}} \end{aligned} \tag{24}$$

which is nonnegative since $\mu_{\mathcal{G}} > 0$ and given u_{fk}, v_f ,$\tilde{v_f}$ and $\tilde{g_k}$ nonnegative for all f, k. Equation (24) can be expressed in matrix form as follows:

$$\hat{g} = \tilde{g} \odot \frac{[U^T \frac{[v]}{[Ug]}]}{[\mu_{\mathcal{G}} e_K + U^T e_F]} \tag{25}$$

where e_K and e_F are all-ones column vectors of appropriate size.

Some important insights are discussed here-under:

- Case $\beta = 2$: as discussed earlier, the existence of a unique minimizer for Problem (22) on $(0, \infty)$ is guaranteed for $\beta < 2$. Let us illustrate this theoretical bound by considering the case $\beta = 2$. For such value Equation (23) becomes: $a + bg_k - c = 0$, hence leading to the following expression for the minimizer $\hat{g_k}$: $\hat{g_k} = \frac{c|_{\beta=2} - a}{b|_{\beta=2}}$. Since there is a negative term in the numerator, there is no guarantee anymore that $\hat{g_k} \in (0, \infty)$ for any nonnegative c and positive a.

- Other values of β: similar rationale can be followed to derive updates in closed-form for $\beta \in (-\infty, 1) \cup \{5/4, 4/3, 3/2\}$. The derivation of the updates for the core tensor in the case $\beta = 0$ is detailed in Appendix A.

- Complexity and numerical costs: as opposed to factor updates, in most practical cases we cannot compute and store the matrix U explicitly since it is much larger than the dataset itself. To tackle this issue, we follow the approach proposed in [4] that is all products Ut for any $t := \text{vec}(\mathcal{T})$ are computed using the following identity:

  $$(W \boxtimes H \boxtimes Q)t = \text{vec}(\mathcal{T} \times_1 W \times_2 H \times_3 Q) \tag{26}$$

  Note that the products $U^T t$ are computed similarly since the transposition is distributive with respect to the Kronecker product.

Algorithm 1 summarizes our method to tackle (5) for the β-divergences in the particular case detailed above for $\beta = 1$ which we refer to as Sparse KL-NTD algorithm. The updates for factors H and Q can be derived in the same way, by considering matricization (6) along modes 2 and 3 of the problem.

Algorithm 1 Sparse KL-NTD

Require: A tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, an initialization $\mathcal{G} \in \mathbb{R}_+^{K_1 \times K_2 \times K_3}$, $W \in \mathbb{R}_+^{I_1 \times K_1}$, $H \in \mathbb{R}_+^{I_2 \times K_2}$ and $Q \in \mathbb{R}^{I_3 \times K_3}$, the factorization ranks (K_1, K_2, K_3), a maximum number of iterations, maxiter, and weight vectors $\mu_{\mathcal{G}} > 0$, $\mu_W > 0$, $\mu_H > 0$, $\mu_Q > 0$.

Ensure: A multirank-(K_1, K_2, K_3) NTD $(\mathcal{G}, W, H, Q)$ of $\mathcal{X}$.

```
for $it = 1$ : maxiter do
    % Update of factor $W$
    $U \leftarrow (\mathcal{G} \times_2 H \times_3 Q)_{(1)}$
    $V \leftarrow \mathcal{X}_{(1)}$
    $C \leftarrow eU^T$
    $S \leftarrow 4\mu_W W \odot \left(\frac{[V]}{[WU]} U^T\right)$
    $W \leftarrow \frac{[C^{\cdot 2} + S]^{\cdot \frac{1}{2}} - C}{2\mu_W}$
    $H$ and $Q$ are updated in a similar way as $W$.
    % Update of core tensor $\mathcal{G}$
    $\mathcal{N} \leftarrow (\mathcal{G} \times_1 W \times_2 H \times_3 Q)^{\cdot(-1)} \odot \mathcal{X}$
    $\mathcal{G} \leftarrow \frac{[\mathcal{N} \times_1 W^T \times_2 H^T \times_3 Q^T]}{[\mu_{\mathcal{G}} \mathcal{E} + \mathcal{E} \times_1 W^T \times_2 H^T \times_3 Q^T]}$ with $\mathcal{E}$ an all-ones tensor of dimension $K_1 \times K_2 \times K_3$
end for
return $\mathcal{G}$, $W$, $H$ and $Q$
```

Computational cost. The computational cost of Algorithm 1 is asymptotically....

References

[1] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

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

[3] Valentin Leplat, Nicolas Gillis, and Jérôme Idier. Multiplicative updates for nmf with beta-divergences under disjoint equality constraints. *SIAM Journal on Matrix Analysis and Applications*, 42(2):730–752, 2021.

[4] Axel Marmoret, Florian Voorwinden, Valentin Leplat, Jérémy E. Cohen, and Frédéric Bimbot. Nonnegative tucker decomposition with beta-divergence for music structure analysis of audio signals, 2021.

[5] E. Rechtschaffen. Real roots of cubics: explicit formula for quasi-solutions. *The Mathematical Gazette*, (524):268–276, 2008.

[6] S. A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

Appendix A: Factor and core tensor updates for $\beta = 0$