A unified framework for non-negative matrix and tensor factorisations with a smoothed Wasserstein loss

Stephen Y. Zhang
Department of Mathematics, University of British Columbia
Vancouver, BC Canada
syz@math.ubc.ca

Abstract

Non-negative matrix and tensor factorisations are a classical tool for finding low-dimensional representations of high-dimensional datasets. In applications such as imaging, datasets can be regarded as distributions supported on a space with metric structure. In such a setting, a loss function based on the Wasserstein distance of optimal transportation theory is a natural choice since it incorporates the underlying geometry of the data. We introduce a general mathematical framework for computing non-negative factorisations of both matrices and tensors with respect to an optimal transport loss. We derive an efficient computational method for its solution using a convex dual formulation, and demonstrate the applicability of this approach with several numerical illustrations with both matrix and tensor-valued data.

1. Introduction

Matrix and tensor factorisations are a classical tool for extracting low-dimensional structure from complex high-dimensional datasets. The seminal work of Lee and Seung [14] noted that real-world datasets are often naturally non-negative, and introduced non-negative matrix factorisation (NMF) for finding an approximate low-rank representation of a matrix-valued dataset that is easier to interpret. NMF can be interpreted in a sparse-coding sense, in that a small set of atoms is sought that can approximately generate the full dataset in its non-negative linear span. This general concept of finding linear representations in terms of a small number of components also extends to tensors [25], for which many notions of non-negative decompositions have been proposed, including the popular CANDECOMP/PARAFAC (CP) format [12].

Approximate factorisations are typically sought with respect to some divergence function [12, 11] such as the squared Frobenius norm or Kullback-Leibler divergence. Such divergences decompose elementwise in their matrix or tensor-valued arguments. In settings such as imaging the observed data lie naturally on a metric space, and elementwise divergences cannot take advantage of this additional structure. Recent works [20, 18, 22, 21] have focused on addressing this issue in the context of NMF by employing a Wasserstein loss that accounts for the geometry of the data by using optimal transport.

In this work, we generalise the smoothed dual approach of Rolet et al. [20] from matrices to the setting of tensors. The problem of finding non-negative tensor factorisations with a Wasserstein loss has remained untouched until very recently, when it was addressed by the work of Afshar et al. [1] in which a non-negative CP decomposition was sought via the primal formulation of optimal transport [8]. In contrast, the approach we consider proceeds via convex duality in order to take advantage of the availability of closed-form gradients in the dual problem [5, 6, 20].

Our work presents a unified framework for Wasserstein factorisation problems, since it handles the fully general case of finding a Tucker decomposition and includes non-negative CP decompositions and NMF as special cases.

2. Background

The reader is provided with an overview of our notation conventions in Supplement A.

2.1. Non-negative matrix factorisation

As a prelude to the more general problem of non-negative tensor factorisations, we discuss the case of NMF. Given a $m \times n$ non-negative matrix $X \in \mathbb{R}_{\geq 0}^{m \times n}$ and a target rank $1 \leq r \leq \min(m, n)$, the NMF problem [14, 24] seeks to find non-negative factor matrices $U \in \mathbb{R}_{\geq 0}^{m \times r}, V \in \mathbb{R}_{\geq 0}^{n \times r}$ such that we have a rank-r approximation $X \approx UV^\top = \sum_{k=1}^{r} U_k \otimes V_k$ in some appropriate sense. In terms of the

columns of X, we have equivalently

$$X_i \approx (UV^\top)_i = \sum_{k=1}^{r} U_k V_{ik}. \tag{1}$$

If columns of X are observations, then (1) represents each observation X_i as a linear combination of r atoms $\{U_k\}_{k=1}^r$ with non-negative coefficients $\{V_{ik}\}_{k=1}^r$. The factor matrix U thus contains an approximate r-element basis for the dataset.

Factors U and V are found by solving a minimisation problem of the form

$$\min_{U \in \mathbb{R}_{\geq 0}^{m \times r}, V \in \mathbb{R}_{\geq 0}^{n \times r}} \varphi(X, UV^\top). \tag{2}$$

In the above, $\varphi(\cdot, \cdot)$ is a suitably chosen loss function over matrices, commonly taken to be the previously mentioned squared Frobenius norm $\varphi(A, B) = \|A - B\|_F^2$ or the generalised Kullback-Leibler (KL) divergence $\varphi(A, B) = \mathrm{KL}(A|B)$ [24].

2.2. Non-negative tensor factorisation

Now consider a non-negative d-mode tensor $X \in \mathbb{R}_{\geq 0}^{n_1 \times \cdots \times n_d}$, for which we seek a low-rank, non-negative representation in a similar sense to NMF. The NMF problem can be directly generalised to d-mode tensors to yield the CP decomposition format [12]. Given a target rank r, we seek d factor matrices $A^{(i)} \in \mathbb{R}_{\geq 0}^{n_i \times r}, 1 \leq i \leq d$ such that

$$X \approx [A^{(1)}, \ldots, A^{(d)}] := \sum_{i=1}^{r} A_i^{(1)} \otimes \cdots \otimes A_i^{(d)}. \tag{3}$$

That is, X can be approximated as a sum of r rank-1 tensors, which we illustrate in Figure 1. For $d = 2$, one recovers the rank-r NMF problem.

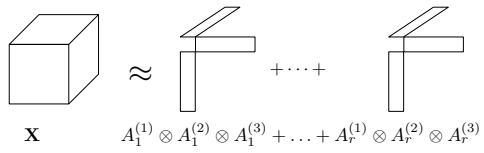


Figure 1: Illustration of the CP tensor decomposition format.

The Tucker decomposition format further generalises the CP format [12]. Given a tensor X and a d-tuple $(r_1, \ldots, r_d)$ specifying the *multilinear rank* of the decomposition, one seeks a core (also known as *mixing*) tensor $S \in \mathbb{R}_{\geq 0}^{r_1 \times \cdots \times r_d}$ and factor matrices $A^{(i)} \in \mathbb{R}_{\geq 0}^{n_i \times r_i}, 1 \leq i \leq d$ such that

$$\begin{aligned} X &\approx S[A^{(1)}, \cdots, A^{(d)}] \\ &:= \sum_{i_1=1}^{r_1} \cdots \sum_{i_d=1}^{r_d} S_{i_1, \ldots, i_d} A_{i_1}^{(1)} \otimes \cdots \otimes A_{i_d}^{(d)}. \end{aligned} \tag{4}$$

This can be interpreted similarly to the CP format, but with the core tensor S encoding interactions between the columns of the factor matrices $A^{(i)}$. Importantly, when $r_1 = \cdots = r_d = r$ and $S = \delta_{i_1, \ldots, i_d}$, we recover the CP format.

2.3. Optimal transport

Optimal transport (OT) deals with comparison of (probability) distributions supported on spaces with metric structure. Wasserstein distances have has found broad applications in statistics and machine learning in recent years [17], since they are sensitive to "horizontal" displacements, in contrast to other commonly used distances or divergences that are typically only sensitive to "vertical" discrepancies of distributions.

The central optimal transport problem is: given α, β probability distributions and a matrix C measuring the cost C_{ij} of transport from point i to point j, find the *coupling* γ (i.e. joint distribution having marginals α and β) solving

$$\mathrm{OT}(\alpha, \beta) := \inf_{\gamma \in \Gamma(\alpha, \beta)} \langle C, \gamma \rangle. \tag{5}$$

where $\Gamma(\alpha, \beta) = \{\gamma \geq 0 : \gamma \mathbf{1} = \alpha, \gamma^\top \mathbf{1} = \beta\}$ denotes the set of all couplings of (α, β). Importantly, in the case where α, β are supported on a space $\mathcal{X}$ with a distance function $d_{\mathcal{X}}$, one may pick $C(x, y) = d_{\mathcal{X}}(x, y)^2$. Then $(\alpha, \beta) \mapsto \mathrm{OT}(\alpha, \beta)$ establishes a natural metric on the space of probability distributions with finite second moment known as the 2-Wasserstein (W_2) metric (and more generally, one can define the p-Wasserstein metric) [17, Proposition 2.2].

In practice, the entropic regularisation [3], [17, Chapter 4] is often employed instead:

$$\begin{aligned} \mathrm{OT}_\varepsilon(\alpha, \beta) &:= \inf_{\gamma \in \Gamma(\alpha, \beta)} \langle C, \gamma \rangle + \varepsilon E(\gamma) \\ &= \inf_{\gamma \in \Gamma(\alpha, \beta)} \varepsilon \mathrm{H}(\gamma | e^{-C/\varepsilon}), \end{aligned} \tag{6}$$

where $\varepsilon > 0$ is the regularisation parameter. In the limit $\varepsilon \to 0^+$, the solution of (6) converges to that of (5) [17, Proposition 4.1]. The problem (6) can be solved efficiently using methods such as the Sinkhorn algorithm [3]. Since it is smooth and strictly convex, it is commonly used as a loss function that approximates the Wasserstein distance [5, 20, 22].

A shortcoming of the typical formulations of optimal transport (5, 6) is that the problem is posed over normalised distributions. In practice, input data may not be perfectly normalised, and rescaling could be undesirable, resulting in a potential loss of information. Various works [2, 8, 15] consider relaxations of optimal transport to deal with the case where α, β are allowed to be positive measures. We

introduce here *semi-unbalanced* transport, which takes the form

$$\mathrm{OT}_\varepsilon^\lambda(\alpha,\beta) = \inf_{\gamma:\gamma\mathbf{1}=\alpha} \varepsilon \mathrm{H}(\gamma|e^{-C/\varepsilon}) + \lambda \mathrm{KL}(\gamma^\top \mathbf{1}|\beta). \quad (7)$$

Here, the transport plan γ is required to agree only approximately with the second input measure via the soft marginal penalty $\mathrm{KL}(\cdot|\beta)$. The parameter λ controls the strength of the soft marginal constraint: sending $\lambda \to +\infty$, we recover the standard entropy-regularised problem (6). Like its balanced counterpart, unbalanced OT problems can be solved using a generalised Sinkhorn-like scheme [2].

3. Wasserstein tensor factorisation

3.1. Optimal transport as a distance on tensors

Optimal transport deals with distributions supported on spaces with metric structure, as introduced in Section 2.3. Tensors can be naturally cast in this framework by thinking in terms of a product of metric spaces. In concrete terms, suppose the ith mode of the tensor $X \in \mathbb{R}^{n_1\times\cdots\times n_d}$ corresponds a discrete metric space $(\mathcal{X}_i, d^{(i)})$. Then, the tensor X lives on a product of metric spaces $(\mathcal{X}, d_\mathcal{X}) = (\mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \cdots \oplus \mathcal{X}_d, d_\mathcal{X})$.

In general, the choice of the product metric $d_\mathcal{X}$ is not unique. Let us consider, however, the p-product metric [7] on $\mathcal{X}$ for $1 \le p < \infty$:

$$d_\mathcal{X}(\boldsymbol{x},\boldsymbol{y}) = (d^{(1)}(x_1,y_1)^p + \cdots + d^{(d)}(x_d,y_d)^p)^{\frac{1}{p}}. \quad (8)$$

This family of product metrics leads to a convenient formulation of optimal transport: to formulate the p-Wasserstein distance on a product of discrete spaces, we need to form the cost tensor C encoding cost for moving an unit mass from $(i_1,\ldots,i_d)$ to $(j_1,\ldots,j_d)$. For X a tensor with d modes, C has $2d$ modes and decomposes additively:

$$\begin{aligned} C_{i_1,\ldots,i_d,j_1,\ldots,j_d} &= d_\mathcal{X}((i_1,\ldots,i_d),(j_1,\ldots,j_d))^p \\ &= d^{(1)}(i_1,j_1)^p + \cdots + d^{(d)}(i_d,j_d)^p \\ &= C^{(1)}_{i_1,j_1} + \cdots + C^{(d)}_{i_d,j_d}, \end{aligned} \quad (9)$$

where $C^{(k)}$ is the cost matrix corresponding to the kth mode of the tensor X. We thus define the Wasserstein distance between tensors $X, Y \in \mathcal{P}(\mathcal{X})$ to be

$$\mathrm{OT}(X,Y) := \inf_{\gamma\in\Gamma(X,Y)} \langle C,\gamma\rangle, \quad (10)$$

where $\Gamma(X,Y)$ denotes the set of all possible couplings of the tensors X, Y, i.e.

$$\begin{aligned} \Gamma(X,Y) = \{\gamma \in \mathcal{P}(\mathcal{X}\otimes\mathcal{X}) : \\ \sum_{j_1,\ldots,j_d} \gamma_{i_1,\ldots,i_d,j_1,\ldots,j_d} = X_{i_1,\ldots,i_d}, \\ \sum_{i_1,\ldots,i_d} \gamma_{i_1,\ldots,i_d,j_1,\ldots,j_d} = Y_{j_1,\ldots,j_d}\}. \end{aligned} \quad (11)$$

Choice of normalisation	Σ
Row-normalised factor matrices	$\{A^{(i)}\mathbf{1} = \mathbf{1}\}$
Column-normalised factor matrices	$\{(A^{(i)})^\top\mathbf{1} = \mathbf{1}\}$
Fully normalised factor matrices	$\{\langle A^{(i)}, \mathbf{1}\rangle = 1\}$
Fully normalised core tensor	$\{\langle S, 1\rangle = 1\}$

Table 1: Normalisation constraints on factors

3.2. Problem setup

We will first discuss the problem of finding non-negative tensor factorisations in the fully general case of a Tucker decomposition, since the settings of CP tensor decompositions and NMF fit naturally in this framework as special cases (see Supplement B). Our approach to the Wasserstein tensor factorisation (WTF) problem is based on the approach to NMF introduced by Rolet et al. [20], where the authors utilise duality properties of entropically smoothed optimal transport to efficiently solve a series of smooth, convex subproblems for the factor matrices.

As previously, we consider a non-negative d-mode tensor $X \in \mathbb{R}_{\ge 0}^{n_1\times\cdots\times n_d}$, for which we seek a Tucker decomposition $S[A^{(1)},\ldots,A^{(d)}]$, where $A^{(i)} \in \mathbb{R}_{\ge 0}^{n_i\times r_i}$ is the ith factor matrix and $S \in \mathbb{R}_{\ge 0}^{r_1\times\cdots\times r_d}$ is the core tensor. Let $\Phi : \mathbb{R}_{\ge 0}^{n_1\times\cdots\times n_d} \times \mathbb{R}_{\ge 0}^{n_1\times\cdots\times n_d} \to [0,\infty)$ be a loss function on tensors based on optimal transport, which we define in further detail in Section 3.6. For now, we will require that Φ is smooth and convex in its second argument.

The fundamental WTF problem can be written then as

$$\min_{S,A^{(1)},\ldots,A^{(d)}\ge 0} \Phi(X, S[A^{(1)},\ldots,A^{(d)}]). \quad (12)$$

In addition, we may optionally impose normalisation constraints on the decomposition components. This is a natural constraint for such settings where X contains histogram or count data, since it resolves the issue of multiplicative non-uniqueness in the scaling of factor matrices (that is, without normalisation constraints the decomposition does not change when any two factor matrices are multiplied by κ and κ^{-1} respectively). We list some useful normalisation constraints in Table 1. For ease of notation, we will denote by Σ_i the constraint set for the ith factor matrix, and Σ_0 that of the core tensor.

Following [20], we relax the non-negativity constraint by using a entropy barrier function. For future convenience (see Sections 3.3 and 3.4), we choose to incorporate normalisation constraints into the barrier function: we define $E_{\Sigma_i}(x) = E(x) + \iota(x \in \Sigma_i)$, where ι denotes the indicator function of a convex set

$$\iota(x \in A) = \begin{cases} 0, & x \in A; \\ +\infty, & \text{otherwise.} \end{cases} \quad (13)$$

This yields the following smooth and unconstrained WTF problem

$$\min_{S,A^{(1)},\ldots,A^{(d)}} \Phi(X, S[A^{(1)},\ldots,A^{(d)}]) + \rho_0 E_{\Sigma_0}(S) + \sum_{i=1}^{d} \rho_i E_{\Sigma_i}(A^{(i)}), \tag{14}$$

and this is the form of the problem which we seek to solve. The problem (14) is convex individually in each of the factors, but not jointly. We may therefore seek a local minimum by performing a block coordinate descent in each of the core tensor and factor matrices. For each convex subproblem, we proceed via convex duality [19]. We state now a result on the Legendre transform (see Supplement C for a definition) of the entropy functional [20, 5].

Proposition 1 (Legendre transform of entropy)**.** *The Legendre transform of the unconstrained entropy $\alpha \mapsto E(\alpha)$ is $E^*(u) = \langle \exp(u), \mathbf{1}\rangle$.*

Now let Σ be a constraint set. Then up to a constant, the Legendre transform of the constrained entropy $\alpha \mapsto E_\Sigma(\alpha)$ is:

$$E^*_\Sigma(u) = \begin{cases} \log\langle \exp(u), \mathbf{1}\rangle, & \Sigma = \{\alpha : \langle \alpha, \mathbf{1}\rangle = 1\} \\ \sum_i \log\langle \exp(u_i), \mathbf{1}\rangle, & \Sigma = \{\alpha : \alpha^\top \mathbf{1} = \mathbf{1}\} \\ \sum_i \log\langle \exp((u^\top)_i), \mathbf{1}\rangle, & \Sigma = \{\alpha : \alpha \mathbf{1} = \mathbf{1}\} \end{cases} \tag{15}$$

We note that the first definition applies for vector, matrix or tensor-valued α, whilst the latter two apply only to matrix-valued α.

We now formulate the dual problems for the factor matrices and core tensor. Proofs are deferred to Supplement D.

3.3. Tensor decompositions – factor matrices

For the moment let S be held fixed, and we seek to optimise over only the factor matrices $A^{(k)}$. We recall from Section 2.2 that in the case where $r_1 = \cdots = r_d = r$ and $S_{i_1,\ldots,i_d} = \delta_{i_1,\ldots,i_d}$, this corresponds exactly to the CP decomposition format [12].

We consider the convex and formally unconstrained subproblem for a single factor matrix $A^{(k)}$,

$$\min_{A^{(k)}} \Phi(X, S[A^{(1)},\ldots,A^{(d)}]) + \rho_k E_{\Sigma_k}(A^{(k)}). \tag{16}$$

The corresponding dual problem is given by the following Proposition.

Proposition 2 (Dual problem for factor matrices)**.** *The dual problem for the kth factor matrix $A^{(k)}$, corresponding to (16), is*

$$\min_{U \in \mathbb{R}^{n_1 \times \cdots \times n_d}} \Phi^*(X, U) + \rho_k E^*_{\Sigma_k}\left(\frac{-1}{\rho_k}\Xi^{(k)}(U)\right), \tag{17}$$

where we have written $\Xi^{(k)}$ to be a linear function of U:

$$\Xi^{(k)}(U) = \left[U \times_{j\geq k+1} (A^{(j)})^\top\right]_{(k)} \left[S \times_{j \leq k-1} A^{(j)}\right]^\top_{(k)}. \tag{18}$$

This problem is smooth and convex in the variable U.

Furthermore, at optimality, the primal variable $A^{(k)\star}$ can be recovered from the optimal dual variable $U^\star$ as the solution of

$$\sup_{A^{(k)}} \frac{-1}{\rho_k}\left\langle A^{(k)}, \Xi^{(k)}(U^\star)\right\rangle - E_{\Sigma_k}(A^{(k)}). \tag{19}$$

In particular, letting $Z = \exp\left(\frac{-1}{\rho_k}\Xi^{(k)}(U^\star)\right)$, we get

$$A^{(k)\star} = \begin{cases} Z, & \Sigma_k = \{\}, \\ Z/\langle Z, \mathbf{1}\rangle, & \Sigma_k = \{A^{(k)} : \langle A^{(k)}, \mathbf{1}\rangle = 1\}, \\ \mathrm{diag}(Z\mathbf{1})^{-1} Z, & \Sigma_k = \{A^{(k)} : A^{(k)}\mathbf{1} = 1\} \\ Z\,\mathrm{diag}(Z^\top \mathbf{1})^{-1}, & \Sigma_k = \{A^{(k)} : (A^{(k)})^\top \mathbf{1} = 1\} \end{cases} \tag{20}$$

For the choices of Φ and Σ_k that we consider, the dual problem (17) is a smooth, unconstrained and convex problem in the dual variable U, and can thus be solved using general gradient-based methods.

3.4. Tensor decompositions – core tensor

Now we will consider optimising over the core tensor S, for fixed factor matrices $A^{(i)}$. The subproblem for S reads

$$\min_S \Phi(X, S[A^{(1)},\ldots,A^{(d)}]) + \rho_0 E_{\Sigma_0}(S) \tag{21}$$

We state now the corresponding dual problem.

Proposition 3. *The dual problem corresponding to (21) is*

$$\min_{U \in \mathbb{R}^{n_1 \times \cdots \times n_d}} \Phi^*(X, U) + \rho_0 E^*_{\Sigma_0}\left(\frac{-1}{\rho_0}\Omega(U)\right), \tag{22}$$

where we have written $\Omega(U) = U \times_{j=1}^d (A^{(j)})^\top$. This problem is smooth and convex in the variable U.

Furthermore, at optimality, the primal variable S can be recovered from the optimal dual variable $U^\star$ as the solution of

$$\sup_S \left\langle \frac{-1}{\rho_0}\Omega(U^\star), S\right\rangle - E_{\Sigma_0}(S). \tag{23}$$

In particular, letting $Z = \exp\left(\frac{-1}{\rho_0}\Omega(U^\star)\right)$, we get

$$S^\star = \begin{cases} Z, & \Sigma_0 = \{\} \\ Z/\langle Z, \mathbf{1}\rangle, & \Sigma_0 = \{S : \langle S, \mathbf{1}\rangle = 1\} \end{cases}. \tag{24}$$

As with Section 3.3, this is a smooth, unconstrained and convex problem in the dual variable U.

3.5. Legendre transform for optimal transport loss

The strategy we proposed in Sections 3.3, 3.4 relies on having access to the Legendre transform of $\Phi(X, \cdot)$. We state now two crucial results about the Legendre transform of $\beta \mapsto \mathrm{OT}_\varepsilon(\alpha, \beta)$ and $\beta \mapsto \mathrm{OT}_\varepsilon^\lambda(\alpha, \beta)$, which will allow us to compute $\Phi^*(X, \cdot)$ for certain choices of Φ.

Proposition 4 (Semi-unbalanced transport)**.** *The Legendre transform of $\beta \mapsto \mathrm{OT}_\varepsilon^\lambda(\alpha, \beta)$ is*

$$\mathrm{OT}_\varepsilon^{\lambda *}(\alpha, u) = -\varepsilon \langle \alpha, \log(\alpha \oslash (Kf(u))) - \mathbf{1} \rangle, \qquad (25)$$

where $f(u) = \left(\frac{\lambda}{\lambda - u}\right)^{\lambda/\varepsilon}$, and at optimality, the relationship between the primal variable $\beta^\star$ and dual variable $u^\star$ is

$$\beta^\star = \left(\frac{\lambda}{\lambda - u^\star}\right) f(u^\star) \odot K^\top \frac{\alpha}{Kf(u^\star)}. \qquad (26)$$

where in the above we have written $K = e^{-C/\varepsilon}$ to be the Gibbs kernel [17].

Proof. See Supplement D. □

Since we recover OT_ε from $\mathrm{OT}_\varepsilon^\lambda$ in the limit $\lambda \to +\infty$, we expect their Legendre transforms to also coincide in the limit. We note that for Proposition 4, in the limit $\lambda \to +\infty$ we have:

$$\lim_{\lambda \to +\infty} \left(\frac{\lambda}{\lambda - u}\right)^{\lambda/\varepsilon} = e^{u/\varepsilon}. \qquad (27)$$

Thus we recover the known result for balanced transport (see e.g. [5, Theorem 2.4]):

Proposition 5 (Balanced transport)**.** *The Legendre transform of $\beta \mapsto \mathrm{OT}_\varepsilon(\alpha, \beta)$ is*

$$\mathrm{OT}_\varepsilon^*(\alpha, u) = -\varepsilon \left\langle \alpha, \log(\alpha \oslash (Ke^{u/\varepsilon})) - \mathbf{1} \right\rangle \qquad (28)$$

Furthermore, at optimality, the relationship between the primal variable $\beta^\star$ and dual variable $u^\star$ is

$$\beta^\star = e^{u^\star/\varepsilon} \odot K^\top \frac{\alpha}{Ke^{u^\star/\varepsilon}}. \qquad (29)$$

3.6. Optimal transport as a loss functional

We discuss now choices of the loss functional Φ which may be useful in certain contexts. Although Rolet et al. [20] employ the well-known entropy-regularised optimal transport loss (6), we will use instead the more flexible semi-unbalanced loss (7).

In particular, in applications sometimes input data may not be perfectly normalised [18], or may have a multimodal distribution [22]. In such settings, a strict requirement for mass transport may result in excessive sensitivity to noise in the input data. Further, for factorisation problems we found empirically that using the semi-unbalanced loss can improve numerical stability. This can be explained by noting that a candidate low-rank approximation may not always perfectly match the input in terms of total mass.

In the following, we work with the semi-unbalanced loss $\mathrm{OT}_\varepsilon^\lambda(\cdot, \cdot)$, since we formally recover balanced transport in the limit $\lambda \to +\infty$.

Proposition 6 (Smoothed Wasserstein loss on tensors)**.** *Let $\mathcal{X}$ be a d-dimensional discrete product metric space as discussed in Section 3.1. Suppose $X, \hat{X} \in \mathcal{M}_+(\mathcal{X})$. Then, the smoothed Wasserstein loss directly applied on tensors is $\Phi(X, \hat{X}) = \mathrm{OT}_\varepsilon^\lambda(X, \hat{X})$, where $C_{i_1,\ldots,i_d,j_1,\ldots,j_d} = \sum_{k=1}^d C^{(k)}_{i_k,j_k}$ is a cost tensor that decomposes along the modes. We write $C^{(k)}$ to be the cost matrix for the kth mode.*

Let $U \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a tensor of dual variables corresponding to $\hat{X}$. Then we have $\Phi^*(X, U) = \mathrm{OT}_\varepsilon^{\lambda *}(X, U)$, where we interpret the formula (25) of Proposition 4 in terms of inner products, elementwise operations, and contractions on tensors.

A key step in evaluating $\mathrm{OT}_\varepsilon^{\lambda *}$ is a convolution with the Gibbs kernel $K = e^{-C/\varepsilon}$. For a cost (9) that decomposes additively along different modes, this amounts to a series of n-mode products (see Supplement E) for which there exist efficient parallel computation schemes [10]. Therefore, there is no need to directly deal with the (prohibitively large) full cost tensor C.

For imaging applications [9], 3-mode tensors arise naturally by stacking two-dimensional images. In such a setting, the obvious choice is a sum of Wasserstein losses over images, which are slices of the tensor X.

Proposition 7 (Smoothed Wasserstein loss along slices)**.** *Let $\mathcal{X}$ be a two-dimensional discrete metric space and suppose that X is a 3-mode tensor such that the slice $X_{i,:,:} \in \mathcal{M}_+(\mathcal{X})$ is a matrix containing 2-dimensional information such as an image. Then in terms of matricisations, the columns of $X_{(1)}^\top$ are the vectorised images and so a Wasserstein loss that decomposes along slices is*

$$\Phi(X, \hat{X}) = \sum_{i=1}^{n_1} \mathrm{OT}_\varepsilon^\lambda \left(\left(X_{(1)}^\top\right)_i, \left(\hat{X}_{(1)}^\top\right)_i \right), \qquad (30)$$

where $\mathrm{OT}_\varepsilon^\lambda$ is a smoothed Wasserstein distance between 1-dimensional histograms, with a cost matrix encoding distances between vectorised images. This approach addresses the setting of sparse image coding using tensors introduced by [9].

Let U be a tensor of dual variables corresponding to $\hat{X}$. Note that each scalar entry of $\hat{X}$ appears only once in the

sum (30). *Thus, the Legendre transform decomposes along the sum and we have:*

$$\Phi^*(X, U) = \sum_{i=1}^{n_1} \mathrm{OT}_\varepsilon^{\lambda *} \left(\left(X_{(1)}^\top \right)_i, \left(U_{(1)}^\top \right)_i \right). \qquad (31)$$

3.7. Implementation

To find a local minimum of (14), we perform in general a block coordinate descent in $A^{(1)}, \ldots, A^{(d)}, S$. Each convex subproblem is solved via its dual problem, which admits closed form gradients [5, 20]. We use L-BFGS for moderately sized problems, and gradient descent for large problems. We employ the PyTorch automatic differentiation engine [16] and the Tensorly tensor routine library [13]. Since the problem (14) is non-convex, a good initialisation may improve the result: while a random initialisation of the core tensor and factor matrices as done by [1] can be used, in Section 4 we opt to employ the commonly used non-negative SVD initialisation [12].

4. Results

4.1. Simulated data – 3-mode tensor

We first deal with the setting where the input tensor is a histogram lying on a product of metric spaces. We take the space $\mathcal{X} = \widehat{\mathcal{X}}^3$ where $\widehat{\mathcal{X}}$ is a regularly spaced grid with 128 points, equipped with the squared product metric, i.e. (8) with $p = 2$. The cost tensor decomposes following (9). Cost matrices $C^{(i)}$, $i = 1, 2, 3$ are normalised to unit mean.

We construct a ground truth tensor that is the mixture of three separable distributions: $X_{\text{true}} = \sum_{i=1}^{3} \alpha_i \otimes \beta_i \otimes \gamma_i$, where $\{\alpha_i, \beta_i, \gamma_i\}_{i=1}^3$ are discrete univariate Gaussians supported on $\widehat{\mathcal{X}}$ that we illustrate in Figure 2(a). We now consider a scenario where we have access only to limited sample observations from the ground truth. Given these samples, we seek to recover the separable components of the underlying distribution by finding a low-rank approximation to the (high-rank) observed tensor. Here, we take the tensor X to be an empirical distribution drawn from X_{true}. We next apply WTF with parameters $\varepsilon = 0.01, \rho_i = 10^{-3}, \lambda = 25$ to find a rank-3 approximation to X, with the additional constraint that the learned univariate components be normalised. For comparison, we also computed rank-3 approximations using a standard non-negative CP factorisation with a Frobenius loss, and the SWIFT algorithm [1] with the identical parameters $\varepsilon = 0.01, \lambda = 25$.

We show the recovered univariate factors in Figure 2(a), and also visualise projections of the recovered tensors in Figure 2(b). From this, we see that WTF recovers smooth atoms that are faithful to the ground truth. In contrast, Frobenius-CP finds irregular atoms that contain many 'spikes' and fail to capture the underlying structure in the empirical distribution. This behaviour can be partially

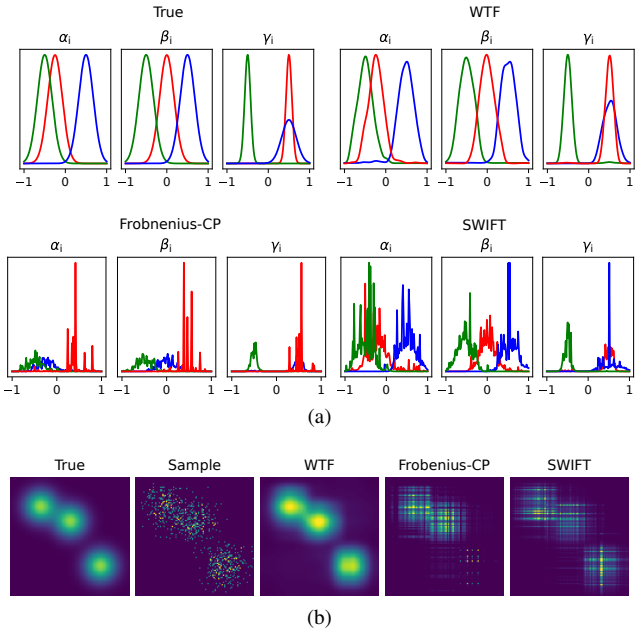


Figure 2: (a) True and recovered atoms visualised as univariate distributions, using WTF, Frobenius, and SWIFT decomposition methods. (b) True, sampled, and recovered tensors visualised as projections onto the first two dimensions.

explained by the pointwise nature of the Frobenius norm, which is insensitive to the spatial structure in the noisy input tensor. Finally, SWIFT produces atoms which correspond roughly to the true factors (i.e. localised in the correct region). However, these are significantly more noisy than those recovered by WTF. We hypothesise that the difference in behaviour observed between WTF and SWIFT partly due to the different formulation of optimal transport on tensors – WTF employs optimal transport with the natural product metric (9) which allows mass to be transported 'globally' on the product space. In contrast, SWIFT uses a sum of optimal transport terms on the i-mode fibers of the tensor [1] and thus for each term, there is the limitation that mass is constrained to be transported along one-dimensional fibers.

4.2. Simulated data – stacked images

We construct a tensor X of dimensions $100 \times 32 \times 32$, for which the ith slice $X_{i,:,:}$ is a 32×32 discrete distribution constructed as a mixture of three separable bivariate distributions $X_{i,:,:} = {Z_i}^{-1} \sum_{k=1}^{3} \alpha_k^{(i)} \otimes \beta_k^{(i)}$, where $\{\alpha_k^{(i)}, \beta_k^{(i)}\}_{k=1}^3$ are discretised Gaussians on `linspace(-1, 1, 32)`, and Z_i is a normalising constant. The observed univariate distributions $\{\alpha_k^{(i)}, \beta_k^{(i)}\}_{k=1}^3$ are constructed by applying random, normally-distributed translations to some fixed 'ground truth' distributions $\{\alpha_k, \beta_k\}_{k=1}^3$. We illustrate in Figure 3 the simulated

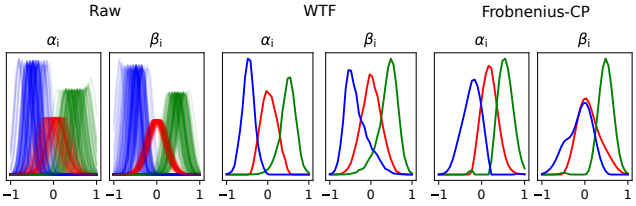


Figure 3: (Left) Marginal distributions $\{\alpha_k^{(i)}, \beta_k^{(i)} : k = 1, 2, 3\}_{i=1}^{100}$, where k is indexed by color, after applying random translational noise. (Center and right) Atoms learned by WTF and Frobenius-CP decompositions, respectively, as univariate distributions that generate the rank-1 bivariate atoms.

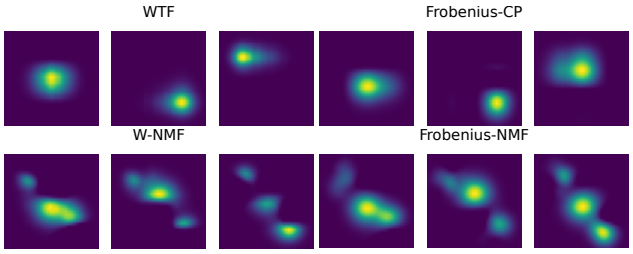


Figure 4: Atoms learned by WTF, Frobenius-CP, W-NMF and Frobenius-NMF.

dataset, showing the randomly shifted marginal distributions that generate the observed bivariate distributions. Examining the averaged distribution reveals that the noisy observations fluctuate around three modes located along the diagonal, corresponding to the 'ground truth' distributions $\alpha_1 \otimes \beta_1, \alpha_2 \otimes \beta_2$, and $\alpha_3 \otimes \beta_3$.

To find a rank-3 CP decomposition, we applied a standard non-negative CP factorisation with a Frobenius loss as well as WTF with $\varepsilon = 0.01, \rho_i = 0.01, \lambda = 25$. Both methods learn a decomposition where each slice $X_{i,\cdot,\cdot}$ is represented as a mixture of $r = 3$ rank-1 matrices. For WTF, we imposed the additional constraint that the learned univariate atoms lie in the simplex. We show the separable atoms learned by the respective methods also in Figure 3. For ease of comparison, we show normalised atoms in the case of Frobenius-CP.

It is clear that the atoms learned by WTF provide a reasonable summary of the input data. The presence of three distinct unimodal atoms in each dimension is evident, and the spatial arrangement agrees with the distribution of the noisy inputs. On the other hand, Frobenius-CP appears to struggle with the presence of translational noise due to the pointwise nature of the loss function, yielding two atoms in the second dimension that share a mode.

As an alternative to seeking tensor decompositions, we could vectorise the 32×32 matrices as columns to form a 1024×100 matrix and then apply NMF with Frobenius and Wasserstein losses (Frobenius-NMF and WNMF) respectively. For WNMF we used the same parameters as for WTF previously, and require that the components be normalised. We show in Figure 4 the atoms learned by WTF, Frobenius-CP, WNMF and Frobenius-NMF respectively. From this it is clear that the atoms found by matrix factorisation are high-rank, each capturing partial information across all three components of the mixture. In contrast, the atoms found by tensor factorisation are separable, and each atom clearly corresponds to only a single mode.

4.3. Learning basis for faces

The AT&T Olivetti faces dataset [1] consists of 400 images (40 subjects, 10 images per subject). Images were resized to 32×32 and normalised to have unit mass. The dataset was randomly split into a set of training and test images, each of which contained 200 images (5 images per individual in each set). We constructed from this $200 \times 32 \times 32$ tensors X_{train} and X_{test} by stacking images from the respective sets as slices along the first mode. WTF was applied to the training data X_{train} to find CP decompositions of varying rank $r \in \{10, 20, \ldots, 100\}$ with parameters $\varepsilon = 10^{-3}, \rho_i = 5 \times 10^{-3} \times r^{-1}, \lambda = 10$, with the constraint that learned atoms lie in the simplex. We also applied a standard non-negative CP decomposition with a squared Frobenius norm loss (which we denote F-CP).

To examine the learned factors, motivated by the observations of [9, 23] we expect the separable basis elements to roughly correspond to spatially localised features of the input. To investigate this, we applied spectral clustering to the basis images learned by WTF and Frobenius-CP respectively, and show their superpositions by cluster in Figure 5(a-b). We observe that atoms learned by WTF can be grouped into clusters that highlight spatial regions corresponding to prominent features of the face, such as forehead, cheekbone, nose, etc. On the other hand, the atoms learned by Frobenius-CP effectively fail to cluster, suggesting that each the atoms do not spatially segregate into distinct features.

To assess the usefulness of the learned basis for supervised classification, we projected the test dataset X_{test} onto the basis learned from the training set. This was done by solving (17) for a coefficient matrix $A_{\text{test}}^{(1)}$ whilst holding the factor matrices encoding atoms $\{A^{(2)}, A^{(3)}\}$ fixed. This problem is convex, so we are guaranteed a unique solution. The rows of $A_{\text{test}}^{(1)}$ are the coordinates of the images in the basis $\{A^{(2)}, A^{(3)}\}$ learned from the training dataset. Following [21], we then use 1-nearest neighbour classification with a cosine distance $(x, y) \mapsto 1 - \cos(\angle(x, y))$ to assign each test image to one of the 40 individual labels. In Figure 6(a) we summarise the accuracy of this approach over 10

[1] this dataset is accessible at http://www.cs.nyu.edu/~roweis/

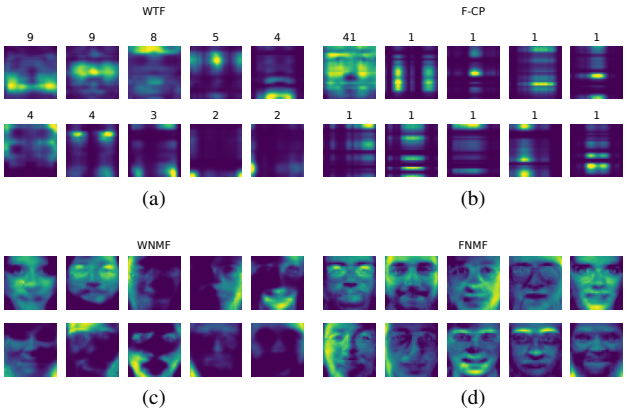
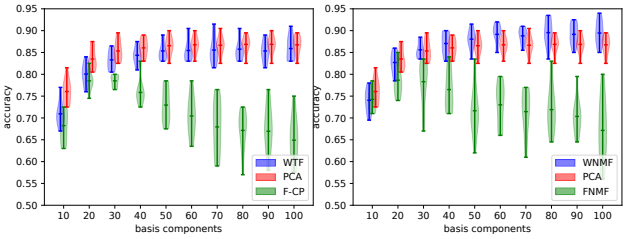


Figure 5: (a, b) Superpositions of rank-1 atoms after clustering for finding a tensor approximation of rank 50. The numbers display the number of rank-1 atoms in each cluster. (c, d) Full-rank atoms found by WNMF and Frobenius-NMF respectively, seeking a basis of size 10.



(a) Tensor factorisations (b) Matrix factorisations

Figure 6: Classification accuracy as a function of the number of basis components for (a) tensor factorisations (separable atoms) and (b) matrix factorisations (full rank atoms).

random train-test splits as a function of the basis size (factorisation rank) r. As a reference for performance, we also display the accuracy for a basis of r principal components (PCs). Note that the PCs are full-rank 32×32 matrices that are not non-negative, compared to the rank-1 non-negative basis atoms sought by tensor factorisation methods. Thus, each full-rank atom contains 16-fold more entries than a rank-1 atom. We find that classification using the WTF basis learned from X_{train} achieves performance comparable to the PCA basis, despite the additional restrictions for WTF that the basis elements be rank-1 and non-negative. In contrast, the performance of Frobenius-CP degrades as the number of components increases, suggesting that the basis elements found using the pointwise Frobenius loss have a poor ability to generalise beyond the training examples.

We next compare the tensor representation found by WTF against the equivalent representation found by matrix decompositions using Wasserstein-NMF (W-NMF). Training and test datasets were constructed as in the tensor case, except we took X_{train} and X_{test} to be 1024×200 matrices with columns corresponding to the vectorised images. We sought a decomposition using both Frobenius (F-NMF) and Wasserstein (W-NMF) losses. For W-NMF, we used parameters identical to WTF. Each atom is a vector of length 1024 which represents a 32×32 matrix with no constraints on rank, in contrast to the rank-1 constraint in the case of the tensor representation.

In Figure 5(c-d) we show the individual atoms found by WNMF and Frobenius-NMF respectively. Curiously, as in the case of tensor factorisations, the atoms found by WNMF visually correspond to localised facial features. In contrast, all of the atoms found by FNMF redundantly capture the full structure of the face. We assessed the performance of the bases found by WNMF and FNMF for classification of the test dataset X_{test}. As shown in Figure 6(b), we find that WNMF achieves a classification accuracy that is on-par or higher than PCA. On the other hand, as in the case of tensor decompositions, the accuracy of FNMF decreases as the number of basis components is increased. Finally, we note that for a fixed number of basis components r, the matrix representation requires effectively 4.6 fold more stored entries than the CP tensor representation, and each full-rank atom is equivalent to 16 rank-1 entries in terms of stored entries. However, for the same number of basis elements we find that WTF achieves a classification accuracy that is comparable to WNMF. This suggests that the tensor format is more efficient for representing image data [9, 23].

5. Conclusion

Motivated by practical settings where observed data lie on a space with metric structure, we formulated the problem of finding non-negative factorisations of matrices and tensors using a Wasserstein loss and propose to solve it numerically via the dual formulation. Alone the way, we derived a closed-form Legendre transform for the semi-unbalanced Wasserstein loss (7) , which to our knowledge has not been previously reported in the literature. Avenues for future work include generalising our approach to deal with sparse data as in [1], as well as exploring alternative choices of barrier functions for the non-negativity constraint. One direction of interest is to develop a methodology where the operation of taking linear combinations of atoms is replaced with taking the Wasserstein barycenter [4], as was done in the setting of matrix factorisations by Schmitz et al. [22].

Acknowledgements

The author would like to thank Elina Robeva for an introduction to the theory of tensors, Hugo Lavenant for an introduction to the duality theory of smoothed optimal transport, and Igor Pinheiro for insightful discussions.

References

[1] Ardavan Afshar, Kejing Yin, Sherry Yan, Cheng Qian, Joyce C Ho, Haesun Park, and Jimeng Sun. Swift: Scalable wasserstein factorization for sparse nonnegative tensors. *arXiv preprint arXiv:2010.04081*, 2020. 1, 6, 8

[2] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. 2, 3

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013. 2

[4] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014. 8

[5] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016. 1, 2, 4, 5, 6

[6] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018. 1

[7] Michel Marie Deza and Elena Deza. Encyclopedia of distances. In *Encyclopedia of distances*, pages 1–583. Springer, 2009. 3

[8] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. *arXiv preprint arXiv:1506.05439*, 2015. 1, 2

[9] Tamir Hazan, Simon Polak, and Amnon Shashua. Sparse image coding using a 3d non-negative tensor factorization. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 50–57. IEEE, 2005. 5, 7, 8

[10] Oguz Kaya and Bora Uçar. High performance parallel algorithms for the tucker decomposition of sparse tensors. In *2016 45th International Conference on Parallel Processing (ICPP)*, pages 103–112. IEEE, 2016. 5

[11] Yong-Deok Kim and Seungjin Choi. Nonnegative tucker decomposition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1

[12] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 1, 2, 4, 6

[13] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. Tensorly: Tensor learning in python. *Journal of Machine Learning Research*, 20(26):1–6, 2019. 6

[14] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 1

[15] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018. 2

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6

[17] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2, 5

[18] Wei Qian, Bin Hong, Deng Cai, Xiaofei He, Xuelong Li, et al. Non-negative matrix factorization with sinkhorn distance. In *IJCAI*, pages 1960–1966, 2016. 1, 5

[19] Ralph Rockafellar. Duality and stability in extremum problems involving convex functions. *Pacific Journal of Mathematics*, 21(1):167–187, 1967. 4

[20] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638. PMLR, 2016. 1, 2, 3, 4, 5, 6

[21] Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011. 1, 7

[22] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018. 1, 2, 5, 8

[23] Amnon Shashua and Anat Levin. Linear image coding for regression and classification using the tensor-rank principle. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 7, 8

[24] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2012. 1, 2

[25] Max Welling and Markus Weber. Positive tensor factorization. *Pattern Recognition Letters*, 22(12):1255–1261, 2001. 1