

# Fast Convolutive Nonnegative Matrix Factorization Through Coordinate and Block Coordinate Updates

Anthony Degleris, Benjamin Antin, Surya Ganguli, Alex H Williams

**Abstract**—Identifying recurring patterns in high-dimensional time series data is an important problem in many scientific domains. A popular model to achieve this is convolutive nonnegative matrix factorization (CNMF), which extends classic nonnegative matrix factorization (NMF) to extract short-lived temporal motifs from a long time series. Prior work has typically fit this model by multiplicative parameter updates—an approach widely considered to be suboptimal for NMF, especially in large-scale data applications. Here, we describe how to extend two popular and computationally scalable NMF algorithms—Hierarchical Alternating Least Squares (HALS) and Alternating Nonnegative Least Squares (ANLS)—for the CNMF model. Both methods demonstrate performance advantages over multiplicative updates on large-scale synthetic and real world data.

**Index Terms**—Convolutive nonnegative matrix factorization, hierarchical alternating least squares, alternating nonnegative least squares, coordinate descent

## I. INTRODUCTION

NMF models a matrix of nonnegative data,  $\mathbf{X}$ , as the product two low rank and nonnegative matrices, thus approximating each datapoint (a row or column of  $\mathbf{X}$ ) as a conical combination of basis features or latent factors [1], [2]. When the low rank assumption is appropriate, NMF often yields highly interpretable descriptions of data and thus is a highly effective tool for exploratory data analysis. NMF has been applied to high-dimensional time series data, with applications ranging from audio processing, image processing, neuroscience, and text mining [2]–[5].

However, many time series contain short-term temporal correlations or sequences of events that are not approximately low rank, and thus cannot be extracted by NMF. For example, audio recordings are typically represented and visualized as spectrograms, which display the frequency content of sound over time as the signal varies. Many sounds of interest have recognizable signatures in the frequency domain, which are not low rank—e.g. phonemes in human speech data may slightly change in frequency (pitch) over their production interval. Similarly, in time series data from neuroscience, it is common to find clusters of brain cells that fire in a rapid sequence [6], [7]. NMF could efficiently model these firing events if neurons fired simultaneously; however, sparse sequences of neural firing yield high rank data matrices. At best, NMF can

only describe such sequences through multiple latent factors. At worst, including additional factors to fit such structure may result in overfitting.

*Convolutive NMF* (CNMF) is a simple extension of the NMF model that overcomes these shortcomings. As its name suggests, CNMF introduces convolutional structure into the low rank model reconstruction, and thus captures short-term temporal dependencies in time series data [8], [9]. The CNMF model has been effective in a variety of applications, including neuroscience [10], medical data mining [11], and audio signal processing [12].

In recent years, algorithms for NMF have matured to a stage where it is computationally tractable to fit very large datasets [13], [14]. However, the CNMF model cannot be viewed as a special case of NMF, and thus these algorithmic improvements are not immediately transferable to CNMF. As a result, while many high-performance and computationally scalable code packages are available for NMF [15], [16], algorithms and implementations of CNMF are less mature.

The Multiplicative Update (MU) algorithm appears to be the most common optimization routine for CNMF in published literature [8]–[10], [12]. This method was originally developed for NMF [1], [17], and later adapted to CNMF [8], [9]. However, subsequent work found MU to be relatively ineffective for NMF [2], suggesting that MU may also be suboptimal for CNMF. Here we derive two new algorithms for the CNMF model—Hierarchical Alternating Least Squares (HALS) and Alternating Nonnegative Least Squares (ANLS)—both of which can be understood as extensions of successful NMF algorithms [2] and are special cases of coordinate and block coordinate descent [18], [19]. We show that HALS and ANLS outperform MU on CNMF models fit to large-scale data. Additionally, we derive several reformulations of the CNMF objective function which lead to new and useful interpretations of the model.

## II. BACKGROUND

### A. Notation

We denote a vector with  $P$  real-valued entries as  $\mathbf{x} \in \mathbb{R}^P$ , a  $P \times Q$  matrix as  $\mathbf{X} \in \mathbb{R}^{P \times Q}$ , and a  $P \times Q \times R$  tensor (in this paper, a tensor is an array with three indices) as  $\mathcal{X} \in \mathbb{R}^{P \times Q \times R}$ . If a matrix (or vector or tensor) has strictly nonnegative entries, we write  $\mathbf{X} \in \mathbb{R}_+^{P \times Q}$ , or alternatively  $\mathbf{X} \geq 0$ .

We denote the  $i$ th slice of the tensor  $\mathcal{X}$  along its first mode as  $\mathbf{X}_{i::}$ , which indicates the first index is fixed to  $i$  while the rest remain free. In our example above,  $\mathbf{X}_{i::}$  refers to a matrix

This work received support from the Department of Energy Computational Science Graduate Fellowship (CSGF) program, the Burroughs Wellcome Fund, the Alfred P. Sloan Foundation, the Simons Foundation, the McKnight Foundation, the James S. McDonell Foundation, and the Office of Naval Research.

The authors are with the Departments of Electrical Engineering (A.D., B.A.), Applied Physics (S.G.), and Statistics (A.H.W.), Stanford University, Stanford, CA 94305 USA (e-mail: ahwillia@stanford.edu).

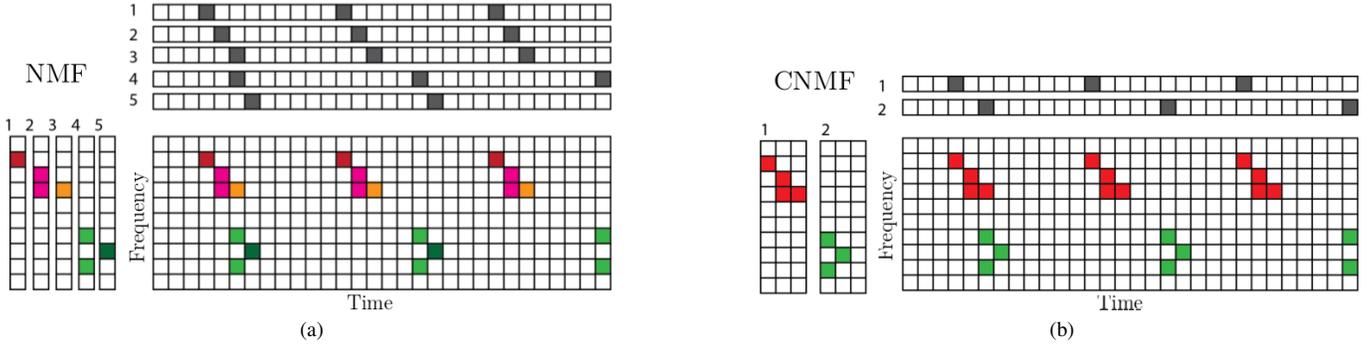


Fig. 1. Schematic illustration of NMF and CNMF models fitting the same dataset. (a) NMF models a data matrix  $\mathbf{X}$  (lower right) the product of  $\mathbf{W}$  (matrix with  $K = 5$  columns, left) and  $\mathbf{H}$  (matrix with  $K = 5$  rows, top). In the “sum-of-outer-products” interpretation of the model, each column of  $\mathbf{W}$  represents a group of simultaneously activated features, while the corresponding row of  $\mathbf{H}$  represents the times at which this group of features is active. (b) CNMF extends this “sum-of-outer-products” interpretation using a convolution operator instead of a vector outer product. Here, the same data are modeled using a tensor  $\mathcal{W}$  (tensor with  $K = 2$  slices, left) and  $\mathbf{H}$  (matrix with  $K = 2$  rows, top). Each of the  $K$  slices of  $\mathcal{W}$  can be thought as a spatiotemporal feature of temporal duration  $L$ , and the times at which each such feature is convolutionally activated are specified by the corresponding row of  $\mathbf{H}$ . This structure enables a much more compact and interpretable representation of this example time series.

of size  $Q \times R$ , whereas  $\mathbf{X}_{:,i}$  refers to a matrix of size  $P \times Q$ . In the following sections, we overload the notation  $\mathbf{W}_\ell = \mathbf{W}_{\ell,:}$  to refer to the matrix created by fixing the index of the tensor  $\mathcal{W}$  along the first mode. Concretely, if  $\mathcal{W}$  is a tensor of size  $L \times N \times K$ , then  $\mathbf{W}_\ell$  is a matrix of size  $N \times K$ .

The symbol  $\odot$  refers to element-wise multiplication of matrices, i.e.  $(\mathbf{A} \odot \mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$  (Hadamard product). Similarly, the notation  $\frac{\mathbf{A}}{\mathbf{B}}$  refers to element-wise division. In all cases, the *norm* of a vector, matrix, or tensor is defined as the root sum-of-squares. For a tensor  $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ , this is

$$\|\mathcal{X}\| = \left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \mathcal{X}_{ijk}^2 \right)^{1/2}.$$

The symbol  $\otimes$  refers to the Kronecker product between two matrices. If  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{B} \in \mathbb{R}^{p \times k}$ , then the Kronecker product  $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{mp \times nk}$  is

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B} & \dots & \mathbf{A}_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{m1}\mathbf{B} & \dots & \mathbf{A}_{mn}\mathbf{B} \end{bmatrix}.$$

If the matrix  $\mathbf{A}$  has columns  $\mathbf{a}_1, \dots, \mathbf{a}_n$ , then the vectorization of  $\mathbf{A}$  is defined as

$$\text{vec}(\mathbf{A}) = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}.$$

When discussing update rules that solve the NMF and CNMF problem, we use superscripts to denote the iteration number of the algorithm. For example,  $\mathbf{W}^{(i)}$  refers to the matrix  $\mathbf{W}$  at the  $i$ th iteration of an algorithm or update scheme,  $\mathbf{W}^{(i+1)}$  refers to the next iterate, and so on.

### B. The NMF and CNMF models

Given a data matrix  $\mathbf{X} \in \mathbb{R}_+^{N \times T}$ , NMF attempts to find two nonnegative factor matrices  $\mathbf{W} \in \mathbb{R}_+^{N \times K}$  and  $\mathbf{H} \in \mathbb{R}_+^{K \times T}$  that

roughly approximate  $\mathbf{X}$ . Formally, the NMF problem is:

$$\begin{aligned} & \underset{\mathbf{W}, \mathbf{H}}{\text{minimize}} && \|\mathbf{X} - \mathbf{W}\mathbf{H}\|^2 \\ & \text{subject to} && \mathbf{W} \geq 0, \mathbf{H} \geq 0. \end{aligned} \quad (1)$$

This model can be reformulated as a sum of outer products. Letting  $\mathbf{w}_1, \dots, \mathbf{w}_K$  be the columns of  $\mathbf{W}$  and  $\mathbf{h}_1^\top, \dots, \mathbf{h}_K^\top$  be the rows of  $\mathbf{H}$ , the objective is:

$$\begin{aligned} & \underset{\substack{\mathbf{w}_1, \dots, \mathbf{w}_K \\ \mathbf{h}_1, \dots, \mathbf{h}_K}}{\text{minimize}} && \|\mathbf{X} - \sum_{k=1}^K \mathbf{w}_k \mathbf{h}_k^\top\|^2 \\ & \text{subject to} && \mathbf{w}_k \geq 0, \mathbf{h}_k \geq 0 \quad \forall k \in \{1, 2, \dots, K\}. \end{aligned} \quad (2)$$

The NMF model can be effective when  $\mathbf{X}$  is nonnegative and approximately low rank. However, NMF may perform poorly as a feature extraction method when  $\mathbf{X}$  contains short-lived temporal motifs with high-rank structure. This is demonstrated schematically in Figure 1a. The data matrix  $\mathbf{X}$  represents a time series with  $T$  measurements, with each column representing a single measurement of  $N$  variables. For example,  $N$  could be the number of frequency bins in a spectrogram representation of an audio signal [9], or the number of recorded cells in a neural time series [10]. These time series can contain short-lived patterns that are not low-rank (in Fig. 1, two different recurring patterns are shown in shades of red and green). This results in NMF requiring many dimensions (i.e. a large choice for  $K$ ) to capture the structure in the data. This hampers interpretability as visible patterns in the data are split across multiple factors.

The convolutive NMF (CNMF) model was developed to address this shortcoming [8]. CNMF finds a matrix  $\mathbf{H} \in \mathbb{R}_+^{K \times T}$  and a tensor  $\mathcal{W} \in \mathbb{R}_+^{L \times N \times K}$  that minimizes the following objective:

$$\begin{aligned} & \underset{\mathcal{W}, \mathbf{H}}{\text{minimize}} && \|\mathbf{X} - \sum_{\ell=1}^L \mathcal{W}_\ell \mathbf{H} \mathbf{S}_{\ell-1}\|^2 \\ & \text{subject to} && \mathcal{W} \geq 0, \mathbf{H} \geq 0, \end{aligned} \quad (3)$$

where  $\mathbf{S}_\ell$  is a  $T \times T$  column right-shift matrix, defined as a matrix with ones along the  $\ell$ th upper diagonal and zeros

otherwise. If  $\mathbf{e}_i$  denotes the  $i$ th standard basis vector, then  $\mathbf{e}_i^\top \mathbf{S}_\ell = \mathbf{e}_{i+\ell}^\top$  when  $i + \ell \leq T$ . A visual demonstration makes the role of  $\mathbf{S}_\ell$  clear:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}, \quad \mathbf{AS}_1 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix}$$

$$\mathbf{AS}_2 = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{bmatrix}, \quad \dots$$

When  $L = 1$ , CNMF reduces exactly to NMF. Like ordinary NMF, CNMF also has a natural ‘‘sum of outer products’’ form. If we consider the slices  $\mathbf{W}_{::1}, \dots, \mathbf{W}_{::K} \in \mathbb{R}_+^{L \times N}$  and the row vectors  $\mathbf{h}_1^\top, \dots, \mathbf{h}_K^\top$  of the matrix  $\mathbf{H}$ , we can define the *convolution operator*  $*$  by

$$\mathbf{A} = \mathbf{W}_{::k}^\top * \mathbf{h}_k^\top, \quad \mathbf{A}_{nt} = \sum_{\ell=1}^L \mathbf{W}_{\ell nk} \mathbf{H}_{k,t-\ell}. \quad (4)$$

which can alternatively be written as

$$\mathbf{W}_{::k}^\top * \mathbf{h}_k^\top = \sum_{\tau=1}^T \mathbf{H}_{k\tau} \begin{bmatrix} \mathbf{0}_{\tau-1} & \mathbf{W}_{::k}^\top & \mathbf{0}_{T+1-L-\tau} \end{bmatrix}. \quad (5)$$

where  $\mathbf{0}_p$  signifies  $p$  columns of zeros. To make the notation more concise, we abbreviate the zero-padding as follows:

$$\left[ \mathbf{W}_{::k}^\top \right]_\tau = \begin{bmatrix} \mathbf{0}_{\tau-1} & \mathbf{W}_{::k}^\top & \mathbf{0}_{T+1-L-\tau} \end{bmatrix} \quad (6)$$

$$\mathbf{W}_{::k}^\top * \mathbf{h}_k^\top = \sum_{\tau=1}^T \mathbf{H}_{k\tau} \left[ \mathbf{W}_{::k}^\top \right]_\tau \quad (7)$$

Using equation 5, we rewrite the CNMF objective as

$$\begin{aligned} & \underset{\mathcal{W}, \mathbf{H}}{\text{minimize}} \quad \left\| \mathbf{X} - \sum_{k=1}^K \mathbf{W}_{::k}^\top * \mathbf{h}_k^\top \right\|^2 \\ & \text{subject to} \quad \mathcal{W} \geq 0, \mathbf{H} \geq 0. \end{aligned} \quad (8)$$

Here, each  $\mathbf{W}_{::k} \in \mathbb{R}^{N \times L}$  represents a short-lived temporal pattern, or *motif*, that may have full rank. The nonzero entries of each  $\mathbf{h}_k \in \mathbb{R}^T$  represent the times at which this motif occurs. For the idealized time series in Figure 1, CNMF pulls out a simpler and more interpretable description of data than NMF. In essence, CNMF extracts 2 recurring patterns, corresponding to  $K = 2$  factors in the model. In contrast, NMF requires  $K = 5$  model factors.

We note here briefly that different boundary conditions could be specified for the convolution operation in eq. 8. We adopted zero-padding for these boundary conditions as it appears to be the most standard choice in prior literature [9], [10]. Only minor modifications to our exposition would be needed to handle different choices. For example,  $\mathcal{H}$  could be re-specified as a  $K \times (T - L)$  matrix and each  $\mathbf{S}_\ell$  could be specified as a  $(T - L) \times T$  matrix to specify convolution without padding.

### C. Multiplicative Update (MU) Algorithms

The objective of the NMF problem (equation 1) is non-convex, and finding an exact solution is NP-hard in general [20]. This had led to extensive algorithmic research on NMF,

producing several effective heuristic algorithms [2] and conditions guaranteeing an exact solution in polynomial time [21], [22].

One such heuristic algorithm for NMF is the Multiplicative Update (MU) algorithm. The MU algorithm repeatedly updates  $\mathbf{W}$  and  $\mathbf{H}$  according to the following update rule [17]

$$\mathbf{W}^{(i+1)} = \mathbf{W}^{(i)} \odot \frac{\mathbf{X}\mathbf{H}^{(i)\top}}{\left[ \mathbf{W}^{(i)}\mathbf{H}^{(i)} \right] \mathbf{H}^{(i)\top}}, \quad (9)$$

where the index  $i$  refers to the current iteration of the algorithm. By the symmetry of the NMF problem (the objective can be expressed as  $\|\mathbf{X}^T - \mathbf{H}^T \mathbf{W}^T\|$ ), the same update rule can be applied to  $\mathbf{H}$ . In reality, the MU algorithm is actually just gradient descent with per-parameter scaling factors [18]. Its popularity stems from several desirable properties—the update rule is monotonic, simple to implement, and preserves nonnegativity.

Since NMF is a special case of CNMF (with  $L = 1$ ), solving the latter problem exactly is also NP-hard. Accordingly, heuristic algorithms are also used to fit CNMF, most notably a generalization of MU [9]. In this case, the update rules are

$$\mathbf{W}_\ell^{(i+1)} = \mathbf{W}_\ell^{(i)} \odot \frac{\mathbf{X}(\mathbf{H}^{(i)}\mathbf{S}_\ell)^\top}{\widehat{\mathbf{X}}_\ell^{(i)}(\mathbf{H}^{(i)}\mathbf{S}_\ell)^\top}, \quad (10)$$

$$\mathbf{H}^{(i+1)} = \mathbf{H}^{(i)} \odot \frac{\sum_{\ell=1}^L \mathbf{W}_\ell^{(i)\top} \mathbf{X} \mathbf{S}_{-\ell}}{\sum_{\ell=1}^L \mathbf{W}_\ell^{(i)\top} \widehat{\mathbf{X}}^{(i)} \mathbf{S}_{-\ell}}, \quad (11)$$

where  $\widehat{\mathbf{X}}^{(i)} = \sum_{\ell=1}^L \mathbf{W}_\ell^{(i)} \mathbf{H}^{(i)} \mathbf{S}_{\ell-1}$  is our reconstruction of  $\mathbf{X}$ . As in the NMF case, MU is easy to implement and has been applied frequently to fit CNMF. Nevertheless, past work has shown MU to be suboptimal for fitting NMF compared to other coordinate descent algorithms [2], [18]. We reasoned that exploiting similar coordinate and block-coordinate updates would lead to performance benefits in the case of CNMF.

### D. Hierarchical Alternating Least Squares (HALS) for NMF

Hierarchical alternating least squares (HALS) is a coordinate descent method used to fit NMF [2], [23]. Each update step solves a constrained optimization problem exactly for a single column of  $\mathbf{W}$  or row of  $\mathbf{H}$ . To update a single column  $\mathbf{w}_p$ , we reformulate the NMF objective as

$$\begin{aligned} J(\mathbf{W}, \mathbf{H}) &= \left\| \mathbf{X} - \sum_{k=1}^K \mathbf{w}_k \mathbf{h}_k^\top \right\|^2 \\ &= \left\| \left( \mathbf{X} - \sum_{k \neq p} \mathbf{w}_k \mathbf{h}_k^\top \right) - \mathbf{w}_p \mathbf{h}_p^\top \right\|^2 \end{aligned} \quad (12)$$

and fix all variables except for the  $p^{\text{th}}$  column of  $\mathbf{W}$ .<sup>1</sup> Minimizing over  $\mathbf{w}_p$  is a convex problem, and the Karush-

<sup>1</sup>As in the case of MU, the symmetry of the NMF problem allows us to use the same rule for  $\mathbf{H}$ . For a more detailed derivation of the HALS updates for NMF, see [23].

Kuhn-Tucker (KKT) conditions for optimality generate the closed-form update rules [23]:

$$\mathbf{w}_p^{(i+1)} = \max \left( 0, \frac{\left[ \mathbf{X} - \sum_{k \neq p} \mathbf{w}_k^{(i)} \mathbf{h}_k^{(i)\top} \right] \mathbf{h}_p^{(i)}}{\|\mathbf{h}_p^{(i)}\|^2} \right). \quad (13)$$

Numerical experiments suggest this update rule notably outperforms MU [2]. One possible explanation for this is that although both algorithms have a similar flop count to update all of  $\mathbf{W}$ , HALS solves many exact problems whereas MU computes a single, inexact gradient step.

### E. Alternating Nonnegative Least Squares (ANLS) for NMF

Another popular approach to the NMF problem is to fix  $\mathbf{W}$  or  $\mathbf{H}$ , and to solve the resulting convex sub-problem exactly. This leads to an algorithm known as *Alternating Nonnegative Least Squares (ANLS)*, whose updates are:

$$\mathbf{W}^{(i+1)} = \arg \min_{\mathbf{W} \geq 0} \|\mathbf{X} - \mathbf{W}\mathbf{H}^{(i)}\|^2, \quad (14)$$

$$\mathbf{H}^{(i+1)} = \arg \min_{\mathbf{H} \geq 0} \|\mathbf{X} - \mathbf{W}^{(i+1)}\mathbf{H}\|^2. \quad (15)$$

Each of these updates amounts to solving a *nonnegative least squares* problem, which has been extensively studied in the optimization literature [24], [25]. Thus, one can leverage existing nonnegative least squares solvers to compute the solution to each sub-problem. A variety of such solvers are available, including active-set methods, quasi-newton methods, and projected gradient methods [2]. This motivates us to also extend the ANLS approach to fit CNMF.

## III. NEW ALGORITHMS FOR CNMF

### A. HALS

In this section, we demonstrate how to extend HALS to fit the CNMF model, highlighting the key reformulations used in deriving the update rule.

*Updating  $\mathbf{W}$ :* Recall the CNMF objective from (3). The sum  $\sum_{\ell} \mathbf{W}_{\ell} \mathbf{H} \mathbf{S}_{\ell-1}$  can be written as a block matrix product by defining

$$\tilde{\mathbf{W}} = [\mathbf{W}_1 \quad \mathbf{W}_2 \quad \dots \quad \mathbf{W}_L], \quad \tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \mathbf{S}_0 \\ \mathbf{H} \mathbf{S}_1 \\ \vdots \\ \mathbf{H} \mathbf{S}_{L-1} \end{bmatrix}.$$

Using the fact that  $\sum_{\ell} \mathbf{W}_{\ell} \mathbf{H} \mathbf{S}_{\ell-1} = \tilde{\mathbf{W}} \tilde{\mathbf{H}}$ , we can reformulate the CNMF objective as

$$\begin{aligned} & \underset{\mathbf{W}, \mathbf{H}}{\text{minimize}} && \|\mathbf{X} - \tilde{\mathbf{W}} \tilde{\mathbf{H}}\|^2 \\ & \text{subject to} && \tilde{\mathbf{W}} \geq 0, \tilde{\mathbf{H}} \geq 0, \\ & && \tilde{\mathbf{H}}_{\ell K:(\ell+1)K,:} = \tilde{\mathbf{H}}_{0:K,:} \mathbf{S}_{\ell} \\ & && \text{for all } \ell = \{0, \dots, L-1\}, \end{aligned} \quad (16)$$

where the last constraint ensures that  $\tilde{\mathbf{H}}$  has the block matrix structure described above. This reveals an important fact: the CNMF approximation is an NMF factorization with linear constraints on  $\tilde{\mathbf{H}}$ .

Due to this reformulation, it is clear that the HALS update rule for NMF extends easily to  $\mathbf{W}$ . When updating  $\mathbf{W}$ , we treat  $\mathbf{H}$  as fixed, and thus we can ignore the linear constraints in 16. Letting  $\tilde{\mathbf{w}}_p \in \mathbb{R}^N$  be the  $p$ th column of  $\tilde{\mathbf{W}}$  and  $\tilde{\mathbf{h}}_p \in \mathbb{R}^T$  be the  $p$ th row of  $\tilde{\mathbf{H}}$ , we have the update rule

$$\tilde{\mathbf{w}}_p^{(i+1)} := \max \left( 0, \frac{\left[ \mathbf{X} - \sum_{j \neq p} \tilde{\mathbf{w}}_j^{(i)} \tilde{\mathbf{h}}_j^{(i)\top} \right] \tilde{\mathbf{h}}_p^{(i)}}{\|\tilde{\mathbf{h}}_p^{(i)}\|^2} \right). \quad (17)$$

This rule allows us to update  $\mathbf{w}_{\ell,:,k}$  using  $p = (l-1)L + k$ . Note that in practice, the matrices  $\tilde{\mathbf{H}}, \tilde{\mathbf{W}}$  do not need to be explicitly instantiated. Each  $\tilde{\mathbf{w}}_p$  is simply an array view into  $\mathbf{W}$ , and each block matrix  $\mathbf{H} \mathbf{S}_{\ell}$  comprising  $\tilde{\mathbf{H}}$  can be computed on demand.

*Updating  $\mathbf{H}$ :* Deriving an update rule for  $\mathbf{H}$  is more complicated due to the convolutive structure imposed on  $\mathbf{H}$ . We first consider the outer product form of the CNMF objective from equation 8 and expand the convolution operator

$$\begin{aligned} & \left\| \mathbf{X} - \sum_{k=1}^K \mathbf{W}_{::k}^{\top} * \mathbf{h}_k^{\top} \right\|^2 \\ &= \left\| \mathbf{X} - \sum_{k=1}^K \sum_{\tau=1}^T \mathbf{H}_{k\tau} [\mathbf{W}_{::k}^{\top}]_{\tau} \right\|^2 \\ &= \left\| \mathbf{E}^{(i)} - \mathbf{H}_{kt} [\mathbf{W}_{::k}^{\top}]_t \right\|^2, \end{aligned} \quad (18)$$

where  $[\mathbf{W}_{::k}^{\top}]_t$  is the matrix  $\mathbf{W}_{::k}^{\top}$  padded with  $t-1$  columns of zeros on the left and  $T+1-L-t$  columns of zeros on the right (first defined in (6), and where we define  $\mathbf{E}^{(i)}$  as

$$\mathbf{E}^{(i)} = \left( \mathbf{X}^{(i)} - \sum_{(p,\tau) \neq k,t} \mathbf{H}_{p\tau}^{(i)} [\mathbf{W}_{::p}^{(i)\top}]_{\tau} \right)_{:,t:t+L-1}. \quad (19)$$

This equation is reminiscent of equation 12, and indeed leads to a related update rule. Fixing all variables but a single entry  $\mathbf{H}_{kt}$ , we can derive the Lagrangian and corresponding Karush-Kuhn-Tucker (KKT) conditions for optimality (see Appendix B-B). This leads us to a closed form update rule for a single entry of  $\mathbf{H}$ :

$$\mathbf{H}_{kt}^{(i+1)} = \max \left( 0, \frac{\text{Tr}(\mathbf{W}_{::k}^{(i)} \mathbf{E}^{(i)})}{\|\mathbf{W}_{::k}^{(i)}\|^2} \right), \quad (20)$$

which completes the generalization of HALS to CNMF.

Indeed, when  $L = 1$ , the HALS update rule for NMF (eq. 13) can be recovered exactly. Specifically, when  $L = 1$ , both  $\mathbf{E}^{(i)}$  and  $\mathbf{W}_{::k}^{(i)}$  reduce to length- $N$  vectors:  $\mathbf{E}^{(i)}$  is column  $t$  of the residual matrix, and  $\mathbf{W}_{::k}^{(i)}$  is the  $k$ th low rank factor. Thus, the numerator term  $\text{Tr}(\mathbf{W}_{::k}^{(i)} \mathbf{E}^{(i)})$  reduces to a vector inner product. To update an entire row of  $\mathbf{H}$  at once, as is standard for HALS in NMF, the numerator term may be extended to be a matrix-vector product, recovering eq. 13.

However, when  $L > 1$ , updating the full row of  $\mathbf{H}$  in closed form is not feasible. Specifically, the update rule for  $\mathbf{H}_{kt}$  is dependent on the current value of  $\mathbf{H}_{k\tau}$  for all  $t < \tau < t+L$ , meaning that one can only simultaneously update every

$L^{\text{th}}$  entry in the  $k^{\text{th}}$  row in  $\mathbf{H}$ . Thus, there are two potential extensions of HALS for CNMF, when updating  $\mathbf{H}$ :

- *Update  $\mathbf{H}$  in blocks of size  $T/L$ .* Iterate over  $\ell = 1, \dots, L-1$  and, starting at position  $\ell$ , update every  $L^{\text{th}}$  entry of row  $k$  in  $\mathbf{H}$ . In principle, this could be achieved by appropriately truncating and reshaping the residual matrix  $\mathbf{E}^{(i)}$ .
- *Update single entries of  $\mathbf{H}$ .* Iterate over  $t = 1, \dots, T$  and update  $\mathbf{H}_{kt}$  by equation 20. Note that one need not compute the full residual matrix; only columns ranging from  $t$  to  $t+L$  of  $\mathbf{E}^{(i)}$  should be computed. This results in  $O(NL)$  total floating point operations to update a single entry of  $\mathbf{H}$ .

In both cases, the relevant entries in  $\mathbf{E}^{(i)}$  should be updated after each parameter update. The second option listed above (pure coordinate descent) is simpler to implement, and thus we focused on this variant in our numerical experiments [18].

As in HALS for NMF, adding  $\ell_1$  regularization with weight  $\alpha$  amounts to subtracting  $\alpha$  from the numerator, and adding  $\ell_2$  regularization with weight  $\beta$  amounts to adding  $\beta$  to the denominator [23]. These extensions to the above algorithm could be used to identify regularized and sparse CNMF models. As we are primarily interested in computational performance, we did not explore the statistical benefits of such regularization methods in detail.

## B. ANLS

In this section, we derive an Alternating Nonnegative Least Squares update rule for CNMF. We'll make use of two different formulations of the CNMF model, one for the update of  $\mathcal{W}$  and one for the update of  $\mathbf{H}$ .

*Updating  $\mathcal{W}$ :* If we fix all entries of the matrix  $\mathbf{H}$ , updating the tensor  $\mathcal{W}$  using ANLS is straightforward. We recall the formulation from (16), in which we expressed the CNMF model as a product of block matrices:  $\hat{\mathbf{X}} = \tilde{\mathbf{W}}\tilde{\mathbf{H}}$ . In this form, it is clear that we can update the matrix  $\tilde{\mathbf{W}}$  using an off-the-shelf Nonnegative Least Squares solver. Since the matrix  $\tilde{\mathbf{W}}$  is simply a reshaped version of the tensor  $\mathcal{W}$ , this suffices for updating  $\mathcal{W}$ . Concretely, we have the following update rule:

$$\tilde{\mathbf{W}}^{\tilde{(i+1)}} = \arg \min_{\tilde{\mathbf{W}} \geq 0} \|\mathbf{X} - \tilde{\mathbf{W}}\tilde{\mathbf{H}}^{\tilde{(i)}}\|^2. \quad (21)$$

*Updating  $\mathbf{H}$ :* The update of  $\mathbf{H}$  requires us to use a different formulation of the CNMF model. First, we recall the following fact (see, e.g., [26]):

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A}) \text{vec}(\mathbf{B}), \quad (22)$$

for any matrices  $\mathbf{A}, \mathbf{B}, \mathbf{C}$  (assuming appropriate dimensions). This leads us to the following vectorized version of the CNMF model:

$$\text{vec}(\hat{\mathbf{X}}) = \underbrace{\sum_{\ell=1}^L (\mathbf{S}_{\ell-1}^T \otimes \mathbf{W}_{\ell})}_{\mathbf{V}} \text{vec}(\mathbf{H}).$$

When  $\mathcal{W}$  is fixed, the above equation allows us to update  $\mathbf{H}$  by solving a single Nonnegative Least Squares problem, as we

did in the case of  $\mathcal{W}$ . With this definition, the ANLS update rule for  $\mathbf{H}$  is given by

$$\mathbf{H}^{(i+1)} = \arg \min_{\mathbf{H} \geq 0} \|\text{vec}(\mathbf{X}) - \mathbf{V} \text{vec}(\mathbf{H})\|^2. \quad (23)$$

Thus, we have cast the CNMF optimization problem as an Alternating Nonnegative Least Squares problem. In practice, the matrix  $\mathbf{V} \in \mathbb{R}_+^{NT \times KT}$  may be too large to fit in memory. One way around this is to use a matrix-free method, which requires access to the matrix  $\mathbf{V}$  only through it's matrix vector product. For example, Projected Gradient Descent and Fast Iterative Shrinkage Thresholding (FISTA) are good candidate methods if efficient implementations of  $\mathbf{V}\mathbf{z}$  and  $\mathbf{V}^T\mathbf{z}$  are available [24], [25]. In practice, we find that directly solving (23) at each iteration is inefficient. However, the formulation above leads to two insights.

Noting that  $\mathbf{V}$  is a block-toeplitz matrix, it becomes clear that the update for  $\mathbf{H}$  is actually a higher-dimensional analogue to the standard nonnegative deconvolution problem studied in the literature [27]. The difference is that here, the coefficient matrix is block-toeplitz rather than toeplitz. This suggests the possibility of leveraging the convolutional structure of the problem using approaches which have been applied in the deconvolution case. We leave this to future work.

The second insight is that updating a single column of  $\mathbf{H}$  with the other columns held fixed is simply a Nonnegative Least Squares problem in  $K$  variables which does not require explicitly storing the matrix  $\mathbf{V}$ . Therefore, one approach to solving (23) is block coordinate descent, updating a single column at a time. Since block coordinate descent converges to the optimal solution for Nonnegative Least Squares problems [28], this approach would eventually reach the optimal solution for (23). In practice, it is not necessary to exactly solve (23) at each iteration. For the purposes of our numerical experiments, we make a single pass of coordinate descent at each iteration (updating each column exactly once), using the block-principal pivoting method described in [29].

## IV. NUMERICAL EXPERIMENTS

In this section, we compare all three algorithms on synthetic and experimental data. We find that HALS and ANLS both converge significantly faster than MU, and that their relative performance to MU increases with dataset size. For example, on a large audio dataset, we find that HALS converges roughly five times faster than MU. This effect occurs consistently regardless of random initialization.

In each figure, we measure reconstruction error (loss) by the scaled norm of the residual,  $\|\mathbf{X} - \hat{\mathbf{X}}\|/\|\mathbf{X}\|$ . All results are obtained via the Julia [30] code using version 1.0, published in the GitHub repository at [github.com/degleis1/CMF.jl](https://github.com/degleis1/CMF.jl), which contains implementations of all algorithms and Jupyter notebooks to reproduce figures. We use the Sherlock compute cluster at Stanford to run all simulations, using two cores (Broadwell) with 16GB of memory per core. In all experiments, all algorithms were given the same random initialization.

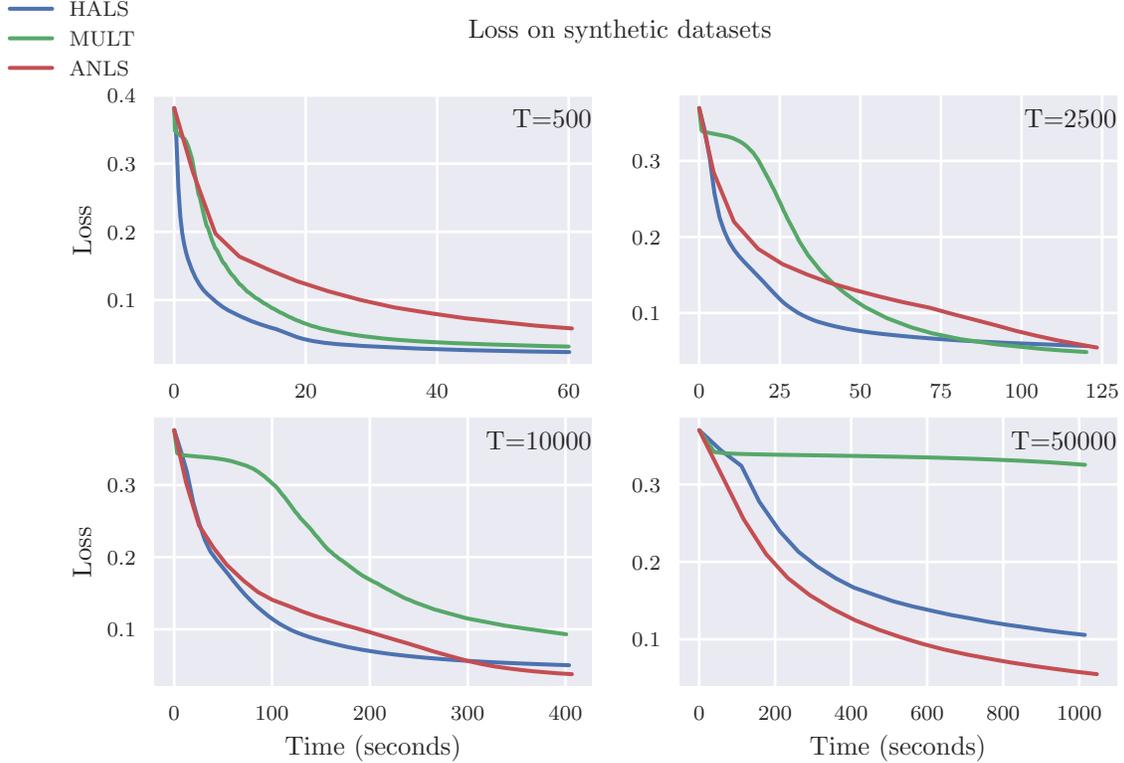


Fig. 2. Algorithm performance on synthetic data. The vertical axis denotes normalized loss of the CNMF model,  $\|\mathbf{X} - \hat{\mathbf{X}}\|/\|\mathbf{X}\|$ ; the horizontal axis denotes cumulative computation time. As dataset size increases (denoted by number of timebins  $T$ ) the performance of HALS and ANLS improves relative to multiplicative updates. For  $T = 500$  and  $T = 2500$  all three algorithms perform similarly. For  $T = 10000$ , multiplicative updates takes significantly longer to converge. Finally, for  $T = 50000$ , multiplicative updates makes little to no progress in the allotted time (1000 seconds), whereas both ANLS and HALS rapidly converge.

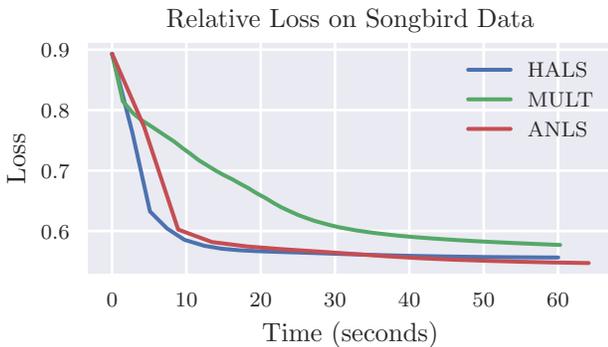


Fig. 3. Algorithm performance on a songbird spectrogram from [10]. ANLS and HALS perform similarly and nearly converge after 20 seconds; multiplicative updates takes approximately three times as long to achieve the same objective value.

#### A. Synthetic Data

In this experiment, we test each algorithm on synthetic data of various sizes. The synthetic datasets were generated from a CNMF model with added noise, as follows:

- The dimensional parameters were chosen to be  $N = 250$ ,  $L = 20$ ,  $K = 5$ . We generated and examined four

otherwise identical datasets with  $T = 500$ ,  $T = 2500$ ,  $T = 10000$ , and  $T = 50000$ .

- Each  $\mathbf{w}_{:nk}$  (the length- $L$  fibers of  $\mathbf{W}$ ) followed a randomly shifted Gaussian curve. Specifically, let  $f(\tau; \mu_{nk}, \sigma)$  denote a univariate Gaussian probability distribution function with mean  $\mu$  and standard deviation  $\sigma$ . We set  $\sigma = 0.2$  and sampled  $\mu_{nk}$  uniformly at random between  $-1$  and  $1$ . We then randomly sampled amplitude parameters from a symmetric Dirichlet distribution  $\boldsymbol{\alpha}_n \sim \text{Dir}(0.1)$ , achieving approximately sparse vectors  $\boldsymbol{\alpha}_n \in \mathbb{R}_+^K$  representing loadings across each component. Finally, we set  $\mathbf{W}_{\ell nk} = \alpha_{nk} f(2\ell/L - 1; \mu_{nk}, \sigma)$ , for  $\ell = 1, \dots, L$ . This procedure was repeated for each feature  $n = 1, \dots, N$  and component  $k = 1, \dots, K$ .
- Each element in  $\mathbf{H}$  was set to zero with probability 0.1, and otherwise randomly sampled from an exponential distribution with a rate parameter  $\lambda = 1$ . Similar to our construction of  $\mathbf{W}$ , this produced a synthetic dataset with sparse factors, in agreement with previously reported results on real data (e.g. [10]).
- The ground truth matrix is given by  $\mathbf{X}^{\text{true}} = \sum_{\ell=1}^L \mathbf{W}_{\ell} \mathbf{H} \mathbf{S}_{\ell-1}$ . We then added truncated Gaussian noise,  $\mathbf{X}_{nt} = \max(0, \mathbf{X}_{nt}^{\text{true}} + e_{nt})$  where each  $e_{nt}$  was drawn uniformly from a standard normal distribution (zero-mean and unit standard deviation). The matrix  $\mathbf{X}$

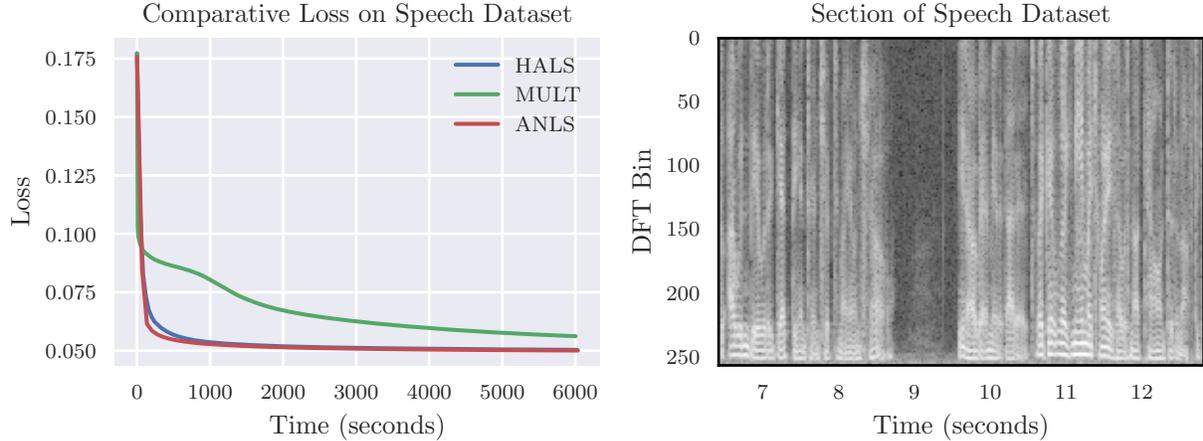


Fig. 4. Comparison of algorithms on a large speech dataset. **Left:** Normalized loss,  $\|\mathbf{X} - \widehat{\mathbf{X}}\|/\|\mathbf{X}\|$ , achieved by each algorithm as a function of computation time. Both HALS and ANLS converge significantly faster than multiplicative updates. **Right:** a small slice of the speech dataset, representative of the full recording.

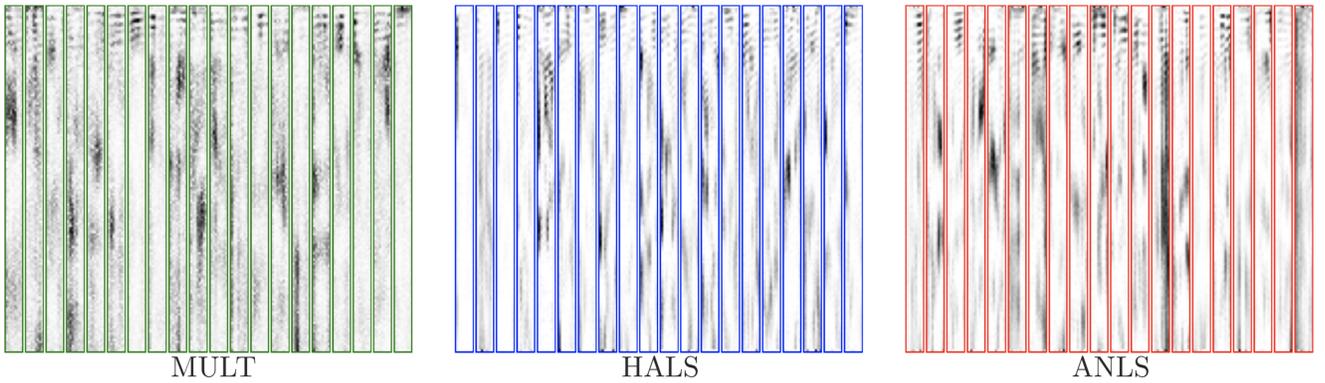


Fig. 5. The twenty components  $W_{:,k}$  recovered by multiplicative updates, HALS, and ANLS. The vertical axis spans frequency (DFT bin) and the horizontal axis spans time. One rectangle surrounded by a colored border is a single  $W_{:,k}$ . The borders of each component are colored to match the previous loss plots. For all three algorithms, the components are perceptually similar but appear in different orders. Specifically, each algorithm recovers harmonic stacks that correspond to different sounds frequently spoken during the recording.

was given as input to all algorithms.

Convergence on synthetic data is shown in Figure 2. For small dataset sizes, all three algorithms give similar performance. As dataset size grows, however, HALS and ANLS converge much more quickly than MU. This is best illustrated when  $T = 50000$  columns. On this large dataset, MU fails to converge within the 1000 second limit.

### B. Results on a Songbird Spectrogram

In this experiment, we fit the CNMF model on a songbird spectrogram from [10] (available at [github.com/FeeLab/seqNMF](https://github.com/FeeLab/seqNMF)). The dimensions of the data matrix are 141 DFT bins (rows) by 4440 timebins (columns), and we use a motif length of  $L = 50$  and  $K = 3$  factors. The timebins are sampled at 200 Hz. We run each algorithm for 60 seconds and plot the relative loss over time. We find that both HALS and ANLS converge after around 20 seconds, whereas multiplicative updates fails to converge within the 60 second time limit (Fig.

3). All algorithms find perceptually similar components (data not shown).

### C. Qualitative Results on a Large Speech Dataset

In this experiment, we fit the CNMF model on a large dataset consisting of two males speaking as part of an interview. Following the procedure in [31], we down-sample the audio recording to 8KHz and compute a magnitude spectrum using an FFT window of 512 samples, and an overlap of 384. This yields a data matrix of size  $257 \times 20149$  which we fit using  $K = 20$  components and motif length of  $L = 12$  time-steps. As a final preprocessing step, we log-transform the spectrogram and add a constant (so that all entries are nonnegative).

We observe that HALS and ANLS converge to their final loss roughly 5x faster than MU. A small section of the magnitude spectrogram, along with a convergence comparison, is shown in Figure 4.

A natural question is whether the components found by HALS and ANLS are similar to those found by MU. We find that this is indeed the case. Figure 5 shows that components recovered by all three algorithms are perceptually similar, each containing distinctive horizontal bands which correspond to the harmonics found in human speech. The components extracted in this experiment look similar to those found by [31].

## V. CONCLUSION

In this paper, we have shown how to extend two popular algorithms for NMF, HALS and ANLS, to the Convolutional NMF problem. Both algorithms offer faster convergence rates than MU, with speedups of around 5x noted on a large dataset, and were observed to recover qualitatively similar motifs. In situations where the practitioner must perform a parameter search over regularization strengths or the number of motifs, this speedup is of practical value. Future research could investigate improvements to the ANLS algorithm by incorporating specialized nonnegative least squares solvers and potentially exploiting the block Toeplitz structure of eq. 23. To handle even larger datasets, randomized variants of the CNMF algorithms described here could also be developed, in analogy to recently proposed randomized variants of HALS in NMF [14]. Overall, we expect these improvements to enable convolutional factor modeling on a variety of high-dimensional time series data with much longer durations than what has been previously explored.

## APPENDIX A

### REFORMULATIONS OF THE CNMF OBJECTIVE

The *classical form* of the CNMF approximation is

$$\mathbf{f}(\mathcal{W}, \mathbf{H}) = \sum_{\ell=1}^L \mathbf{W}_\ell \mathbf{H} \mathbf{S}_{\ell-1} \quad (24)$$

We define the *convolution operator* as

$$\mathbf{W}_{::k}^\top * \mathbf{h}_k^\top = \sum_{\tau=1}^T \mathbf{H}_{k\tau} [\mathbf{W}_{::k}^\top]_\tau \quad (25)$$

where  $[\mathbf{W}_{::k}^\top]_\tau = [\mathbf{0}_{\tau-1} \quad \mathbf{W}_{::k}^\top \quad \mathbf{0}_{T+1-L-\tau}]$  and  $\mathbf{0}_p$  is a  $N \times p$  matrix of zeros. This allows us to write the *outer product form* of the CNMF approximation

$$\mathbf{f}(\mathcal{W}, \mathbf{H}) = \sum_{k=1}^K \mathbf{W}_{::k}^\top * \mathbf{h}_k^\top \quad (26)$$

Another useful formulation comes from considering Kronecker identities. Given three matrices  $\mathbf{A}$ ,  $\mathbf{X}$ ,  $\mathbf{B}$ , we know

$$\text{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \otimes \mathbf{A}) \text{vec}(\mathbf{X}) \quad (27)$$

from [26]. This leads us to the *Kronecker form* of the CNMF approximation, which is

$$\text{vec}(\mathbf{f}(\mathcal{W}, \mathbf{H})) = \sum_{\ell=1}^L (\mathbf{S}_{\ell-1}^\top \otimes \mathbf{W}_\ell) \text{vec}(\mathbf{H}) \quad (28)$$

$$= \sum_{\ell=1}^L (\mathbf{S}_{1-\ell} \otimes \mathbf{W}_\ell) \text{vec}(\mathbf{H}) \quad (29)$$

We define the matrix  $\mathbf{V} = \sum_{\ell=1}^L \mathbf{S}_{1-\ell} \otimes \mathbf{W}_\ell$ , which is also written as

$$\mathbf{V} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{W}_2 & \mathbf{W}_1 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_L & \mathbf{W}_{L-1} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_L & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{W}_1 \end{bmatrix} \quad (30)$$

where  $\mathbf{0}$  is a  $T \times T$  matrix of zeros. Thus the Kronecker form is concisely written as  $\text{vec}(\mathbf{f}(\mathcal{W}, \mathbf{H})) = \mathbf{V} \text{vec}(\mathbf{H})$ .

Alternatively, we can take the transpose of equation (24) and apply (27) to write the *Toeplitz form* of the CNMF approximation

$$\text{vec}(\mathbf{f}(\mathcal{W}, \mathbf{H})^\top) = \sum_{l=1}^L (\mathbf{W}_l \otimes \mathbf{S}_{1-l}) \text{vec}(\mathbf{H}^\top) \quad (31)$$

which is also written as

$$\text{vec}(\mathbf{f}(\mathcal{W}, \mathbf{H})^\top) = \begin{bmatrix} \mathcal{T}(\mathbf{w}_{:11}) & \dots & \mathcal{T}(\mathbf{w}_{:1K}) \\ \vdots & \ddots & \vdots \\ \mathcal{T}(\mathbf{w}_{:N1}) & \dots & \mathcal{T}(\mathbf{w}_{:NK}) \end{bmatrix} \quad (32)$$

where  $\mathcal{T}(\mathbf{v}) \in \mathbb{R}^{T \times T}$  is a Toeplitz matrix defined for any vector  $\mathbf{v} \in \mathbb{R}^L$  as

$$\mathcal{T}(\mathbf{v}) = \begin{bmatrix} \mathbf{v}_1 & 0 & \dots & 0 \\ \mathbf{v}_2 & \mathbf{v}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_L & \mathbf{v}_{L-1} & \dots & 0 \\ 0 & \mathbf{v}_L & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{v}_1 \end{bmatrix} \quad (33)$$

i.e. the  $\ell$ th diagonal below the main diagonal is equal to  $\mathbf{v}_{\ell+1}$ .

## APPENDIX B

### DERIVATIONS OF THE HALS UPDATE RULES

#### A. HALS for NMF

Consider the NMF objective, written as

$$\begin{aligned} & \underset{\substack{\mathbf{w}_1, \dots, \mathbf{w}_k \\ \mathbf{h}_1, \dots, \mathbf{h}_k}}{\text{minimize}} \quad \left\| \mathbf{X} - \sum_{k=1}^K \mathbf{w}_k \mathbf{h}_k^\top \right\|^2 \\ & \text{subject to} \quad \mathbf{w}_k \geq 0, \mathbf{h}_k \geq 0 \quad \forall k \in \{1, 2, \dots, K\} \end{aligned}$$

We will derive a closed-form update rule that updates a single column  $\mathbf{w}_k$  or a single row  $\mathbf{h}_k^\top$ . By the symmetry of the problem, it suffices to derive this update rule for  $\mathbf{w}_k$  only. First choose  $k$  and let  $\mathbf{E} = \mathbf{X} - \sum_{p \neq k} \mathbf{w}_p \mathbf{h}_p^\top$ . Our minimization problem is now

$$\begin{aligned} & \underset{\mathbf{w}_k}{\text{minimize}} \quad \left\| \mathbf{E} - \mathbf{w}_k \mathbf{h}_k^\top \right\|^2 \\ & \text{subject to} \quad \mathbf{w}_k \geq 0 \end{aligned}$$

Applying the identity  $\|\mathbf{X}\|^2 = \text{Tr}(\mathbf{X}^T\mathbf{X})$ , we can rewrite  $J(\mathbf{w}_k) = \|\mathbf{E} - \mathbf{w}_k\mathbf{h}_k^T\|^2$  as

$$\begin{aligned} J(\mathbf{w}_k) &= \text{Tr}(\mathbf{E}^T\mathbf{E}) + \text{Tr}(\mathbf{h}_k\mathbf{w}_k^T\mathbf{w}_k\mathbf{h}_k^T) - 2\text{Tr}(\mathbf{E}^T\mathbf{w}_k\mathbf{h}_k^T) \\ &= \text{Tr}(\mathbf{E}^T\mathbf{E}) + \text{Tr}(\mathbf{h}_k^T\mathbf{h}_k\mathbf{w}_k^T\mathbf{w}_k) - 2\text{Tr}(\mathbf{h}_k^T\mathbf{E}^T\mathbf{w}_k) \\ &= \|\mathbf{E}\|^2 + \|\mathbf{h}_k\|^2\|\mathbf{w}_k\|^2 - 2\mathbf{h}_k^T\mathbf{E}^T\mathbf{w}_k \end{aligned}$$

Next, we write the Lagrangian as

$$\mathcal{L}(\mathbf{w}_k, \lambda) = \text{Tr}(\mathbf{E}^T\mathbf{E}) + \|\mathbf{h}_k\|^2\|\mathbf{w}_k\|^2 - 2\mathbf{h}_k^T\mathbf{E}^T\mathbf{w}_k - \lambda^T\mathbf{w}_k$$

which has gradient

$$\nabla_{\mathbf{w}_k}\mathcal{L}(\mathbf{w}_k, \lambda) = 2\|\mathbf{h}_k\|^2\mathbf{w}_k - 2\mathbf{E}\mathbf{h}_k - \lambda$$

Setting this equal to zero gives us the KKT conditions

$$\mathbf{w}_k = \frac{\mathbf{E}\mathbf{h}_k + \frac{1}{2}\lambda}{\|\mathbf{h}_k\|^2} \quad (34)$$

$$\lambda \geq 0 \quad (35)$$

$$\mathbf{w}_k \geq 0 \quad (36)$$

$$(\mathbf{w}_k)_i\lambda_i = 0 \quad \forall i = 1, 2, \dots, N \quad (37)$$

If  $(\mathbf{E}\mathbf{h}_k)_i \geq 0$ , then we must have  $\lambda_i = 0$  to satisfy equation (37). If  $(\mathbf{E}\mathbf{h}_k)_i \geq 0$ , then we must have  $(\mathbf{w}_k)_i = 0$ . This leads to the closed form solution

$$\mathbf{w}_k = \max\left(0, \frac{\mathbf{E}\mathbf{h}_k}{\|\mathbf{h}_k\|^2}\right) \quad (38)$$

### B. HALS for CNMF

For the CNMF model, the HALS update rule loses its symmetry across  $\mathbf{W}$  and  $\mathbf{H}$ . However, as demonstrated in Section III-A, the update rule for  $\mathbf{W}$  can be derived using the HALS update rule for NMF. To derive the update rule for  $\mathbf{H}$ , we begin by choosing  $k, t$  and defining  $\mathbf{E} = \mathbf{X} - \sum_{(p,\tau) \neq (k,t)} \mathbf{H}_{p,\tau} [\mathbf{W}_{::p}^T]_{\tau}$ . From (18), we can update  $\mathbf{H}_{kt}$  with the optimization problem

$$\begin{aligned} \underset{\mathbf{H}_{kt}}{\text{minimize}} \quad & J(\mathbf{H}_{kt}) = \|\mathbf{E} - \mathbf{H}_{kt}[\mathbf{W}_{::k}^T]_t\|^2 \\ \text{subject to} \quad & \mathbf{H}_{kt} \geq 0 \end{aligned} \quad (39)$$

Since  $[\mathbf{W}_{::k}^T]_t$  only interacts with  $L$  columns of  $\mathbf{E}$ , we can define  $\mathbf{R} = \mathbf{E}_{:,t+L-1}$  and write (39) as

$$\begin{aligned} \underset{\mathbf{H}_{kt}}{\text{minimize}} \quad & J(\mathbf{H}_{kt}) = \|\mathbf{R} - \mathbf{H}_{kt}\mathbf{W}_{::k}^T\|^2 \\ \text{subject to} \quad & \mathbf{H}_{kt} \geq 0 \end{aligned} \quad (40)$$

Since  $\mathbf{H}_{kt}$  is just a scalar, it is quite simple to derive a closed form update rule. The corresponding Lagrangian is

$$\begin{aligned} \mathcal{L}(\mathbf{H}_{kt}, \lambda) &= \|\mathbf{R}\|^2 + \mathbf{H}_{kt}^2\|\mathbf{W}_{::k}^T\|^2 \\ &\quad - 2\mathbf{H}_{kt}\text{Tr}(\mathbf{W}_{::k}\mathbf{R}) - \lambda\mathbf{H}_{kt} \end{aligned} \quad (41)$$

which has gradient

$$\begin{aligned} \nabla_{\mathbf{H}_{kt}}\mathcal{L}(\mathbf{H}_{kt}, \lambda) &= 2\mathbf{H}_{kt}\|\mathbf{W}_{::k}^T\|^2 \\ &\quad - 2\text{Tr}(\mathbf{W}_{::k}\mathbf{R}) - \lambda \end{aligned} \quad (42)$$

This gives us the KKT conditions

$$\mathbf{H}_{kt} = \frac{\text{Tr}(\mathbf{W}_{::k}\mathbf{R}) + \frac{1}{2}\lambda}{\|\mathbf{W}_{::k}^T\|^2} \quad (43)$$

$$\mathbf{H}_{kt} \geq 0 \quad (44)$$

$$\lambda \geq 0 \quad (45)$$

$$\lambda\mathbf{H}_{kt} = 0 \quad (46)$$

Equation (46), which is referred to as the complementary slackness condition, implies that either  $\lambda$  or  $\mathbf{H}_{kt}$ , must be zero. This allows us to update  $\mathbf{H}_{kt}$  using the closed form update rule from Section III-A.

### REFERENCES

- [1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, p. 788, Oct. 1999. [Online]. Available: <https://doi.org/10.1038/44565%20http://10.0.4.14/44565>.
- [2] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, no. 257, 2014.
- [3] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, Oct. 2003, pp. 177–180.
- [4] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 161–173, Jan. 2009, ISSN: 0196-2892. DOI: 10.1109/TGRS.2008.2002882.
- [5] J. K. Liu, H. M. Schreyer, A. Onken, F. Rozenblit, M. H. Khani, V. Krishnamoorthy, S. Panzeri, and T. Gollisch, "Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization," *Nature Communications*, vol. 8, no. 1, p. 149, 2017, ISSN: 2041-1723. DOI: 10.1038/s41467-017-00156-9. [Online]. Available: <https://doi.org/10.1038/s41467-017-00156-9>.
- [6] R. H. R. Hahnloser, A. A. Kozhevnikov, and M. S. Fee, "An ultra-sparse code underlies the generation of neural sequences in a songbird," *Nature*, vol. 419, no. 6902, pp. 65–70, 2002, ISSN: 1476-4687. DOI: 10.1038/nature00974. [Online]. Available: <https://doi.org/10.1038/nature00974>.
- [7] S. Fujisawa, A. Amarasingham, M. T. Harrison, and G. Buzsáki, "Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex," *Nature neuroscience*, vol. 11, no. 7, p. 823, 2008.
- [8] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, C. G. Puntonet and A. Prieto, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 494–499, ISBN: 978-3-540-30110-3.

- [9] P. Smaragdis *et al.*, “Convolutional speech bases and their application to supervised speech separation,” *IEEE Transactions on audio speech and language processing*, vol. 15, no. 1, p. 1, 2007.
- [10] E. L. Mackevicius, A. H. Bahle, A. H. Williams, S. Gu, N. I. Denissenko, M. S. Goldman, and M. S. Fee, “Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience,” *BioRxiv*, 2018. DOI: 10.1101/273128. eprint: <https://www.biorxiv.org/content/early/2018/03/02/273128.full.pdf>. [Online]. Available: <https://www.biorxiv.org/content/early/2018/03/02/273128>.
- [11] V. Ramanarayanan, A. Katsamanis, and S. Narayanan, “Automatic data-driven learning of articulatory primitives from real-time mri data using convolutional nmf with sparseness constraints,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [12] J. Zhou, R. Liang, L. Zhao, L. Tao, and C. Zou, “Unsupervised learning of phonemes of whispered speech in a noisy environment based on convolutional non-negative matrix factorization,” *Information Sciences*, vol. 257, pp. 115–126, 2014.
- [13] R. Kannan, G. Ballard, and H. Park, “A high-performance parallel algorithm for nonnegative matrix factorization,” *SIGPLAN Not.*, vol. 51, no. 8, 9:1–9:11, Feb. 2016, ISSN: 0362-1340. DOI: 10.1145/3016078.2851152. [Online]. Available: <http://doi.acm.org/10.1145/3016078.2851152>.
- [14] N. B. Erichson, A. Mendible, S. Wihlborn, and J. N. Kutz, “Randomized nonnegative matrix factorization,” *Pattern Recognition Letters*, vol. 104, pp. 1–7, 2018, ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2018.01.007>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167865518300138>.
- [15] B. Zupan *et al.*, “Nimfa: A python library for nonnegative matrix factorization,” *Journal of Machine Learning Research*, vol. 13, no. 3, pp. 849–853, 2012.
- [16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [17] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in neural information processing systems*, 2001, pp. 556–562.
- [18] J. Kim, Y. He, and H. Park, “Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework,” *Journal of Global Optimization*, vol. 58, no. 2, pp. 285–319, Feb. 2014, ISSN: 1573-2916. DOI: 10.1007/s10898-013-0035-4. [Online]. Available: <https://doi.org/10.1007/s10898-013-0035-4>.
- [19] S. J. Wright, “Coordinate descent algorithms,” *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, Jun. 2015, ISSN: 1436-4646. DOI: 10.1007/s10107-015-0892-3. [Online]. Available: <https://doi.org/10.1007/s10107-015-0892-3>.
- [20] S. Vavasis, “On the complexity of nonnegative matrix factorization,” *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2010. DOI: 10.1137/070709967.
- [21] D. Donoho and V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts?” In *Advances in neural information processing systems*, 2004, pp. 1141–1148.
- [22] S. Arora, R. Ge, R. Kannan, and A. Moitra, “Computing a nonnegative matrix factorization—provably,” *SIAM Journal on Computing*, vol. 45, no. 4, pp. 1582–1611, 2016. DOI: 10.1137/130913869.
- [23] A. Cichocki, R. Zdunek, and S.-i. Amari, “Hierarchical algorithms for nonnegative matrix and 3d tensor factorization,” in *International Conference on Independent Component Analysis and Signal Separation*, Springer, 2007, pp. 169–176.
- [24] C.-J. Lin, “Projected gradient methods for nonnegative matrix factorization,” *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.
- [25] R. A. Polyak, “Projected gradient method for non-negative least square,” *Contemp Math*, vol. 636, pp. 167–179, 2015.
- [26] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*. Cambridge University Press, 1991. DOI: 10.1017/CBO9780511840371.
- [27] J. T. Vogelstein, A. M. Packer, T. A. Machado, T. Sippy, B. Babadi, R. Yuste, and L. Paninski, “Fast nonnegative deconvolution for spike train inference from population calcium imaging,” *Journal of neurophysiology*, vol. 104, no. 6, pp. 3691–3704, 2010.
- [28] A. Beck and L. Tetruashvili, “On the convergence of block coordinate descent type methods,” *SIAM journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.
- [29] J. Kim and H. Park, “Fast nonnegative matrix factorization: An active-set-like method and comparisons,” *SIAM Journal on Scientific Computing*, vol. 33, no. 6, pp. 3261–3281, 2011.
- [30] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, “Julia: A fresh approach to numerical computing,” *SIAM review*, vol. 59, no. 1, pp. 65–98, 2017.
- [31] P. D. O’grady and B. A. Pearlmutter, “Discovering convolutional speech phones using sparseness and non-negativity,” in *International Conference on Independent Component Analysis and Signal Separation*, Springer, 2007, pp. 520–527.