# Efficient Learning for Time Series Models by Non-Negative Moment Matrix Factorization: *Supplemental Materials*

This document contains derivations for the method-of-moments algorithms used in the paper. The first section describes the general approach to deriving multiplicative update rules for the components of a non-negative matrix factorization (NMF). The next section discusses the convergence properties of NMF. Then we give the pseudocode of the sequence clustering algorithm for mixture of HMMs. And finally, the derivations for the mixture of HMMs, the switching HMM, and the factorial HMM are given. Some of the notation used here may differ slightly from that of the paper.

## 1 Matrix Factorization

A nonnegative matrix factorization is a relationship between two nonnegative matrices, the second of which is a function of other nonnegative matrices. For example, we may have:

$$\mathbf{P} \approx \mathbf{W}\,\mathbf{H}\ , \tag{1}$$

where $\mathbf{P} \in \mathbb{R}_+^{L\times L}$ is expressed in terms of $\mathbf{W} \in \mathbb{R}_+^{L\times J}$ and $\mathbf{H} \in \mathbb{R}_+^{J\times L}$. The + subscript indicates non-negativity. We can derive multiplicative update rules as part of a gradient descent scheme to optimize these factors. This approach was introduced in [Lee & Seung(2001)Lee and Seung]. In this paper, we minimize a Kullback-Liebler (KL) divergence error criterion between an empirical matrix **V** and its (theoretical) factorization **P**:

$$D(\mathbf{V},\mathbf{P}) = tr\left[\mathbf{V}^T \log\left(\frac{\mathbf{V}.}{\mathbf{P}}\right) - \mathbf{1}_{LL}\,\mathbf{V} + \mathbf{1}_{LL}\,\mathbf{P}\right] \propto\ tr\left[-\mathbf{V}^T \log(\mathbf{P}) + \mathbf{1}_{LL}\,\mathbf{P}\right] \tag{2}$$

We apply the rules of matrix calculus [Magnus & Neudecker(1999)Magnus and Neudecker] to find the partial derivatives of $D(\mathbf{V},\mathbf{P})$ with respect to the terms in the factorization:

$$\frac{\partial D}{\partial \mathbf{W}} = -\left(\frac{\mathbf{V}.}{\mathbf{P}}\right)\mathbf{H}^T + \mathbf{1}_{LL}\mathbf{H}^T \quad , \quad \frac{\partial D}{\partial \mathbf{H}} = -\mathbf{W}^T\left(\frac{\mathbf{V}.}{\mathbf{P}}\right) + \mathbf{W}^T\mathbf{1}_{LL}\ . \tag{3}$$

We choose the step sizes in a gradient descent scheme:

$$\mathbf{W} = \mathbf{W} - \boldsymbol{\eta}_{\mathbf{W}} \odot \frac{\partial D}{\partial \mathbf{W}} \quad , \quad \mathbf{H} = \mathbf{H} - \boldsymbol{\eta}_{\mathbf{H}} \odot \frac{\partial D}{\partial \mathbf{H}}\ , \tag{4}$$

to be:

$$\boldsymbol{\eta}_{\mathbf{W}} = \frac{\mathbf{W}\,.}{\mathbf{1}_{LL}\mathbf{H}^T} \qquad , \qquad \boldsymbol{\eta}_{\mathbf{H}} = \frac{\mathbf{H}\,.}{\mathbf{W}^T\mathbf{1}_{LL}} \ , \tag{5}$$

to derive the multiplicative updates:

$$\mathbf{W} = \mathbf{W} \odot \frac{\left(\frac{\mathbf{V}.}{\mathbf{P}}\right)\mathbf{H}^T\,.}{\mathbf{1}_{LL}\mathbf{H}^T} \qquad , \qquad \mathbf{H} = \mathbf{H} \odot \frac{\mathbf{W}^T\left(\frac{\mathbf{V}.}{\mathbf{P}}\right)\,.}{\mathbf{W}^T\mathbf{1}_{LL}} \ . \tag{6}$$

where $\odot$ indicates element-wise multiplication. If the matrices are supposed to be normalized in some way (e.g. column-wise), then this normalization is performed after the corresponding update. These updates are iterated until convergence. If the matrices are initializes with nonnegative values, they will stay this way throughout the optimization. We will follow this general procedure to derive all the update rules in this document.

### 1.1 Relationship with EM and Convergence of NMF

A set of update equations derived via the KL divergence criterion for discrete-valued models corresponds to an EM algorithm. This equivalence was discussed in [Ding et al.(2008)Ding, Li, and Peng] in the context of the non-negative matrix factorization:

$$\mathbf{P} \approx \mathbf{WH} \ , \tag{7}$$

and the corresponding probabilistic decomposition:

$$P\left(t,f\right) = \sum_z P\left(f|z\right)P\left(t,z\right) \ . \tag{8}$$

Multiplicative updates derived from the former are equivalent to EM updates (PLSI) [Hofmann(1999)] derived from the latter up to a normalization. In this paper, we present algorithms derived from matrix factorizations that have equivalent probabilistic decompositions. Thus, since they are equivalent to an EM algorithm (after appropriate normalization), they are guaranteed to converge to a local optimum of the cost function. This is observed in practice.

# 2 Pseudocode for mixture of HMMs

**Algorithm 1** Sequence clustering via learning mixture of HMMs with NMF-MoM.

```
Input: Sequences $\mathbf{x}_{1:N}$
Output: Estimated clustering assignments $\widehat{h}_{1:N}$.
1. Compute empirical moment estimate $\mathbf{V}$.
3. Estimate model parameters $\theta_{1:K}$ using NMF-MoM.
4. $\forall n \in \{1, \dots N\}, \widehat{h}_n = \text{argmax}_k p(\mathbf{x}_n | \theta_k)$.
```

# 3 Mixture of Hidden Markov Models

In this section, we derive the multiplicative update schemes to learn the parameters of a MHMM from any of the first four empirical moments.

## 3.1 Moment matrix decompositions

The probabilistic decompositions of the first four joint distributions are:

$$P(x_t) = \sum_k P(h_k) \sum_{r_t} P(x_t|r_t, h_k) P(r_t|h_k) \ , \tag{9}$$

$$P(x_{t:t+1}) = \sum_k P(h_k) \sum_{r_{t:t+1}} P(x_{t+1}|r_{t+1}, h_k) P(r_{t+1}|r_t, h_k) P(r_t|h_k) P(x_t|r_t, h_k) \ , \tag{10}$$

$$P(x_{t:t+2}) = \sum_k P(h_k) \sum_{r_{t:t+2}} P(x_{t+2}|r_{t+2}, h_k) P(r_{t+2}|r_{t+1}, h_k) \cdots \\ P(x_{t+1}|r_{t+1}, h_k) P(r_{t+1}|r_t, h_k) P(r_t|h_k) P(x_t|r_t, h_k) \ , \tag{11}$$

$$P(x_{t:t+3}) = \sum_k P(h_k) \sum_{r_{t:t+3}} P(x_{t+3}|r_{t+3}, h_k) P(r_{t+3}|r_{t+2}, h_k) P(x_{t+2}|r_{t+2}, h_k) \cdots \\ P(r_{t+2}|r_{t+1}, h_k) P(x_{t+1}|r_{t+1}, h_k) P(r_{t+1}|r_t, h_k) P(r_t|h_k) P(x_t|r_t, h_k) \ , \tag{12}$$

The corresponding moment matrices are:

$$P_1 = \mathrm{E}[\mathbf{x}_t] \ , \tag{13}$$

$$P_2 = \mathrm{E}[\mathbf{x}_{t+1} \otimes \mathbf{x}_t] \ , \tag{14}$$

$$P_3 = \mathrm{E}[\mathbf{x}_{t+2} \otimes \mathbf{x}_{t+1} \otimes \mathbf{x}_t] \ , \tag{15}$$

$$P_4 = \mathrm{E}[\mathbf{x}_{t+3} \otimes \mathbf{x}_{t+2} \otimes \mathbf{x}_{t+1} \otimes \mathbf{x}_t] \ , \tag{16}$$

where ⊗ denotes an outer product. The factorizations of these matrices are, respectively:

$$\mathbf{P}_1 = \tilde{\mathbf{O}} \tilde{\boldsymbol{\pi}} \ , \tag{17}$$

$$\mathbf{P}_2 = \tilde{\mathbf{O}} \bar{\mathbf{A}} \, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \ , \tag{18}$$

$$\mathbf{P}_3(:, j, :) = \tilde{\mathbf{O}} \bar{\mathbf{A}} \, diag\left[\tilde{\mathbf{O}}(j, :)\right] \bar{\mathbf{A}} \, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \ , \tag{19}$$

$$\mathbf{P}_4(:, j, k, :) = \tilde{\mathbf{O}} \bar{\mathbf{A}} \, diag\left[\tilde{\mathbf{O}}(j, :)\right] \bar{\mathbf{A}} \, diag\left[\tilde{\mathbf{O}}(k, :)\right] \bar{\mathbf{A}} \, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \ , \tag{20}$$

where:

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{bmatrix} \quad , \quad \tilde{\boldsymbol{\pi}} = \begin{bmatrix} \pi_1 \boldsymbol{v}_1 \\ \pi_2 \boldsymbol{v}_2 \\ \vdots \\ \pi_K \boldsymbol{v}_K \end{bmatrix} \ . \tag{21}$$

## 3.2 $1^{\text{ST}}$ MOMENT FACTORIZATION

The KL error is:

$$D(\mathbf{V},\mathbf{P}_1) \propto tr\left[-\mathbf{V}^T \log(\mathbf{P}_1) + \mathbf{1}_{1L}\mathbf{P}_1\right] \ . \tag{22}$$

The partial derivatives are:

$$\frac{\partial D}{\partial \tilde{\mathbf{O}}} = -\mathbf{X}\tilde{\boldsymbol{\pi}}^T + \mathbf{1}_{L1}\tilde{\boldsymbol{\pi}}^T \qquad , \qquad \frac{\partial D}{\partial \tilde{\boldsymbol{\pi}}} = -\tilde{\mathbf{O}}^T\mathbf{X} + \tilde{\mathbf{O}}^T\mathbf{1}_{L1} \ , \tag{23}$$

where $\mathbf{X} = \frac{\mathbf{V}.}{\mathbf{P}_1}$. The multiplicative updates are:

$$\boxed{\tilde{\mathbf{O}} = \tilde{\mathbf{O}} \odot \frac{\mathbf{X}\tilde{\boldsymbol{\pi}}^T \ .}{\mathbf{1}_{L1}\tilde{\boldsymbol{\pi}}^T} \qquad , \qquad \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}} \odot \frac{\tilde{\mathbf{O}}^T\mathbf{X} \ .}{\tilde{\mathbf{O}}^T\mathbf{1}_{L1}} \ .} \tag{24}$$

These parameter matrices should be normalized column-wise after each update.

## 3.3 $2^{\text{ND}}$ MOMENT FACTORIZATION

We can use two methods. The first absorbs $diag(\tilde{\boldsymbol{\pi}})$ into $\tilde{\mathbf{A}}$ so that only two matrices are optimized (with no loss of information). The second optimizes all three parameter matrices.

### 3.3.1 TAKE 1

The factorization can be written as:

$$\mathbf{P} = \tilde{\mathbf{O}}\tilde{\mathbf{A}}\tilde{\mathbf{O}}^T = \begin{bmatrix}\mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_3\end{bmatrix} \begin{bmatrix} \pi_1\hat{\mathbf{A}}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \pi_2\hat{\mathbf{A}}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \pi_3\hat{\mathbf{A}}_3 \end{bmatrix} \begin{bmatrix}\mathbf{O}_1^T \\ \mathbf{O}_2^T \\ \mathbf{O}_3^T\end{bmatrix} \ . \tag{25}$$

The KL divergence error criterion is:

$$D(\mathbf{V},\mathbf{P}_2) \propto tr\left[-\mathbf{V}^T \log(\mathbf{P}_2) + \mathbf{1}_{LL}\mathbf{P}_2\right] \ . \tag{26}$$

The partial derivatives are:

$$\frac{\partial D}{\partial \tilde{\mathbf{O}}} = -\mathbf{X}^T\tilde{\mathbf{O}}\tilde{\mathbf{A}} - \mathbf{X}\tilde{\mathbf{O}}\tilde{\mathbf{A}}^T + \mathbf{1}_{LL}\tilde{\mathbf{O}}\left(\tilde{\mathbf{A}}^T + \tilde{\mathbf{A}}\right) \ , \tag{27}$$

$$\frac{\partial D}{\partial \tilde{\mathbf{A}}} = -\tilde{\mathbf{O}}^T\mathbf{X}\tilde{\mathbf{O}} + \tilde{\mathbf{O}}^T\mathbf{1}_{LL}\tilde{\mathbf{O}} \ , \tag{28}$$

where $\mathbf{X} = \frac{\mathbf{V}.}{\mathbf{P}_2}$. The resulting multiplicative updates are:

$$\boxed{\tilde{\mathbf{O}} = \tilde{\mathbf{O}} \odot \frac{\mathbf{X}^T\tilde{\mathbf{O}}\tilde{\mathbf{A}} + \mathbf{X}\tilde{\mathbf{O}}\tilde{\mathbf{A}}^T \ .}{\mathbf{1}_{LL}\tilde{\mathbf{O}}\left(\tilde{\mathbf{A}}^T + \tilde{\mathbf{A}}\right)} \quad , \quad \tilde{\mathbf{A}} = \tilde{\mathbf{A}} \odot \frac{\tilde{\mathbf{O}}^T\mathbf{X}\tilde{\mathbf{O}}}{\tilde{\mathbf{O}}^T\mathbf{1}_{LL}\tilde{\mathbf{O}}} \ .} \tag{29}$$

Considering normalization properties of the parameter matrices, the denominators simplify to:

$$\mathbf{1}_{L1}\left[\left(\tilde{\mathbf{A}}+\mathbf{I}\right)\mathbf{1}_{L1}\right]^{T} \qquad , \qquad \mathbf{1}_{JK,JK} \ . \tag{30}$$

After each iteration, $\tilde{\mathbf{O}}$ should be normalized column-wise and $\tilde{\mathbf{A}}$ should be normalized over the entire matrix.

### 3.3.2 TAKE 2

We can also derive updates that optimize $\tilde{\mathbf{A}}$ and $\tilde{\boldsymbol{\pi}}$ separately. The KL error is given by (26). The partial derivatives are:

$$\frac{\partial D}{\partial \tilde{\mathbf{O}}} = -\mathbf{X}\tilde{\mathbf{O}}\,diag\left(\tilde{\boldsymbol{\pi}}\right)\bar{\mathbf{A}}^{T} - \mathbf{X}^{T}\tilde{\mathbf{O}}\bar{\mathbf{A}}\,diag\left(\tilde{\boldsymbol{\pi}}\right) + \mathbf{1}_{LL}\tilde{\mathbf{O}}\left(diag\left(\tilde{\boldsymbol{\pi}}\right)\bar{\mathbf{A}}^{T} + \bar{\mathbf{A}}\,diag\left(\tilde{\boldsymbol{\pi}}\right)\right) \ , \tag{31}$$

$$\frac{\partial D}{\partial \bar{\mathbf{A}}} = -\tilde{\mathbf{O}}^{T}\mathbf{X}\tilde{\mathbf{O}}\,diag\left(\tilde{\boldsymbol{\pi}}\right) + \tilde{\mathbf{O}}^{T}\mathbf{1}_{LL}\tilde{\mathbf{O}}\,diag\left(\tilde{\boldsymbol{\pi}}\right) \ , \tag{32}$$

$$\frac{\partial D}{\partial \tilde{\boldsymbol{\pi}}} = -\left[\bar{\mathbf{A}}^{T}\tilde{\mathbf{O}}^{T}\mathbf{X}\odot\tilde{\mathbf{O}}^{T}\right]\mathbf{1}_{L1} + \bar{\mathbf{A}}^{T}\tilde{\mathbf{O}}^{T}\mathbf{1}_{L1}\odot\tilde{\mathbf{O}}^{T}\mathbf{1}_{L1} \ , \tag{33}$$

The multiplicative updates are:

$$\boxed{\begin{array}{c} \tilde{\mathbf{O}} = \tilde{\mathbf{O}}\odot\dfrac{\mathbf{X}\tilde{\mathbf{O}}\,diag\left(\tilde{\boldsymbol{\pi}}\right)\bar{\mathbf{A}}^{T} + \mathbf{X}^{T}\tilde{\mathbf{O}}\bar{\mathbf{A}}\,diag\left(\tilde{\boldsymbol{\pi}}\right) \ .}{\mathbf{1}_{LL}\tilde{\mathbf{O}}\left(diag\left(\tilde{\boldsymbol{\pi}}\right)\bar{\mathbf{A}}^{T} + \bar{\mathbf{A}}\,diag\left(\tilde{\boldsymbol{\pi}}\right)\right)} \ , \\ \bar{\mathbf{A}} = \bar{\mathbf{A}}\odot\dfrac{\tilde{\mathbf{O}}^{T}\mathbf{X}\tilde{\mathbf{O}} \ .}{\tilde{\mathbf{O}}^{T}\mathbf{1}_{LL}\tilde{\mathbf{O}}} \quad , \quad \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}\odot\dfrac{\left[\bar{\mathbf{A}}^{T}\tilde{\mathbf{O}}^{T}\mathbf{X}\odot\tilde{\mathbf{O}}^{T}\right]\mathbf{1}_{L1} \ .}{\bar{\mathbf{A}}^{T}\tilde{\mathbf{O}}^{T}\mathbf{1}_{L1}\odot\tilde{\mathbf{O}}^{T}\mathbf{1}_{L1}} \ . \end{array}} \qquad \begin{array}{r} (34) \\[2em] (35) \end{array}$$

Considering normalization properties of the parameter matrices, the denominators simplify to:

$$\mathbf{1}_{L1}\left[\left(\bar{\mathbf{A}}+\mathbf{I}\right)\tilde{\boldsymbol{\pi}}\right]^{T} \qquad , \qquad \mathbf{1}_{JK,JK} \qquad , \qquad \mathbf{1}_{JK,1} \ . \tag{36}$$

After each iteration, all matrices should be normalized column-wise.

## 3.4 3<sup>RD</sup> MOMENT FACTORIZATION

The KL error is:

$$D \propto \sum_{j=1}^{L} tr\left[-\mathbf{V}_{:j:}^{T}\log\left(\mathbf{P}_{:j:}\right) + \mathbf{1}_{LL}\mathbf{P}_{:j:}\right] \ . \tag{37}$$

### 3.4.1 Update of $\tilde{\mathbf{O}}$ matrix

Taking partial derivatives with respect to the first and third appearances of $\tilde{\mathbf{O}}$, we have:

$$\frac{\partial D}{\partial \tilde{\mathbf{O}}} = -\sum_j \mathbf{X}_j \mathbf{U}_{::j}^T + \mathbf{X}_j^T \mathbf{W}_{::j} + \mathbf{1}_{LL} \mathbf{U}_{::j}^T + \mathbf{1}_{LL} \mathbf{W}_{::j} \ , \tag{38}$$

and with respect to the second appearance of $\tilde{\mathbf{O}}$, we have:

$$\frac{\partial D}{\partial \tilde{\mathbf{O}}_{j:}} = -\mathbf{1}_{1L} \left[ \mathbf{X}_j^T \mathbf{H} \odot \mathbf{G}^T \right] + \left[ \mathbf{1}_{1L} \mathbf{H} \odot \mathbf{1}_{1L} \mathbf{G}^T \right] \ , \tag{39}$$

where

$$\mathbf{X}_j = \frac{\mathbf{V}_{:j:}}{\mathbf{P}_{:j:}} \ , \tag{40}$$

$$\mathbf{U}_{::j} = \bar{\mathbf{A}}\, diag\left(\tilde{\mathbf{O}}\left(j,:\right)\right) \mathbf{G} \ , \tag{41}$$

$$\mathbf{H} = \tilde{\mathbf{O}} \bar{\mathbf{A}} \ , \tag{42}$$

$$\mathbf{G} = \bar{\mathbf{A}}\, diag\left(\tilde{\boldsymbol{\pi}}\right) \tilde{\mathbf{O}}^T \ , \tag{43}$$

$$\mathbf{W}_{::j} = \mathbf{H}\, diag\left(\tilde{\mathbf{O}}\left(j,:\right)\right) \bar{\mathbf{A}}\, diag\left(\tilde{\boldsymbol{\pi}}\right) \ . \tag{44}$$

Thus, the multiplicative update is:

$$\tilde{\mathbf{O}} = \tilde{\mathbf{O}} \odot \frac{\left(\mathbf{1}_{1L} \left[ \mathbf{X}_j^T \mathbf{H} \odot \mathbf{G}^T \right]\right)_{rows:j} + \sum_j \mathbf{X}_j \mathbf{U}_{::j}^T + \mathbf{X}_j^T \mathbf{W}_{::j} \ .}{\mathbf{1}_{L1} \left[ \mathbf{1}_{1L} \mathbf{H} \odot \mathbf{1}_{1L} \mathbf{G}^T \right] + \mathbf{1}_{LL} \sum_j \mathbf{U}_{::j}^T + \mathbf{W}_{::j}} \ , \tag{45}$$

where the notation $\left(\mathbf{x}_j\right)_{rows:j}$ indicates that a matrix should be formed whose $j^{\text{th}}$ row is equal to $\mathbf{x}_j$. Simplifying the denominator via normalization properties, we get:

$$\mathbf{1}_{L1} \left[ \left( \bar{\mathbf{A}}^2 + \bar{\mathbf{A}} + \mathbf{I} \right) \tilde{\boldsymbol{\pi}} \right]^T \ . \tag{46}$$

### 3.4.2 Update of $\bar{\mathbf{A}}$ matrix

Taking the partial derivative with respect to $\bar{\mathbf{A}}$, we get:

$$\begin{aligned} \frac{\partial D}{\partial \bar{\mathbf{A}}} = -\sum_j & \tilde{\mathbf{O}}^T \mathbf{X}_j \mathbf{G}^T\, diag\left(\tilde{\mathbf{O}}\left(j,:\right)\right) + diag\left(\tilde{\mathbf{O}}\left(j,:\right)\right) \mathbf{H}^T \mathbf{X}_j \mathbf{M} \\ & + \tilde{\mathbf{O}}^T \mathbf{1}_{LL} \mathbf{G}^T\, diag\left(\mathbf{1}^T \tilde{\mathbf{O}}\right) + diag\left(\mathbf{1}^T \tilde{\mathbf{O}}\right) \mathbf{H}^T \mathbf{1}_{LL} \mathbf{M} \ . \end{aligned} \tag{47}$$

where $\mathbf{G}$ and $\mathbf{H}$ are defined as in the updates for $\tilde{\mathbf{O}}$ and:

$$\mathbf{M} = \tilde{\mathbf{O}}\, diag\left(\tilde{\boldsymbol{\pi}}\right) \ . \tag{48}$$

So, we have the following multiplicative update for $\bar{\mathbf{A}}$:

$$\bar{\mathbf{A}} = \bar{\mathbf{A}} \odot \frac{\sum_j \tilde{\mathbf{O}}^T \mathbf{X}_j \mathbf{G}^T\, diag\left(\tilde{\mathbf{O}}\left(j,:\right)\right) + diag\left(\tilde{\mathbf{O}}\left(j,:\right)\right) \mathbf{H}^T \mathbf{X}_j \mathbf{M} \ .}{\tilde{\mathbf{O}}^T \mathbf{1}_{LL} \mathbf{G}^T\, diag\left(\mathbf{1}^T \tilde{\mathbf{O}}\right) + diag\left(\mathbf{1}^T \tilde{\mathbf{O}}\right) \mathbf{H}^T \mathbf{1}_{LL} \mathbf{M}} \ . \tag{49}$$

Taking normalization properties into account, the denominator simplifies to:

$$\mathbf{1}_{JK,1} \left[ \left( \bar{\mathbf{A}} + \mathbf{I} \right) \tilde{\boldsymbol{\pi}} \right]^T \ . \tag{50}$$

### 3.4.3 UPDATE OF $\tilde{\boldsymbol{\pi}}$ VECTOR

Taking the partial derivative with respect to $\tilde{\boldsymbol{\pi}}$, we get:

$$\frac{\partial D}{\partial \tilde{\boldsymbol{\pi}}} = \sum_j -\left[\mathbf{Y}_{::j}^T \mathbf{X}_j \odot \tilde{\mathbf{O}}^T\right] \mathbf{1}_{L1} + \left[\mathbf{Y}_{::j}^T \mathbf{1}_{L1} \odot \tilde{\mathbf{O}}^T \mathbf{1}_{L1}\right] \ , \tag{51}$$

where

$$\mathbf{Y}_{::j} = \tilde{\mathbf{O}} \bar{\mathbf{A}}\, diag\left(\tilde{\mathbf{O}}\left(j, :\right)\right) \bar{\mathbf{A}} \ . \tag{52}$$

Thus, we have the multiplicative update for $\tilde{\boldsymbol{\pi}}$:

$$\tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}} \odot \frac{\sum_j \left[\mathbf{Y}_{::j}^T \mathbf{X}_j \odot \tilde{\mathbf{O}}^T\right] \mathbf{1}_{L1} \ .}{\sum_j \mathbf{Y}_{::j}^T \mathbf{1}_{L1} \odot \tilde{\mathbf{O}}^T \mathbf{1}_{L1}} \ . \tag{53}$$

Application of normalization properties results in a denominator equal to $\mathbf{1}_{JK,1}$.

### 3.4.4 OVERALL UPDATES

Overall, we have the following multiplicative updates:

$$\boxed{\begin{aligned}
\tilde{\mathbf{O}} &= \tilde{\mathbf{O}} \odot \frac{\left(\mathbf{1}_{1L}\left[\mathbf{X}_j^T \mathbf{H} \odot \mathbf{G}^T\right]\right)_{rows:j} + \sum_j \mathbf{X}_j \mathbf{U}_{::j}^T + \mathbf{X}_j^T \mathbf{W}_{::j} \ .}{\mathbf{1}_{L1}\left[\mathbf{1}_{1L}\mathbf{H} \odot \mathbf{1}_{1L}\mathbf{G}^T\right] + \mathbf{1}_{LL}\sum_j \mathbf{U}_{::j}^T + \mathbf{W}_{::j}} \ , && (54)\\
\bar{\mathbf{A}} &= \bar{\mathbf{A}} \odot \frac{\sum_j \tilde{\mathbf{O}}^T \mathbf{X}_j \mathbf{G}^T\, diag\left(\tilde{\mathbf{O}}\left(j, :\right)\right) + diag\left(\tilde{\mathbf{O}}\left(j, :\right)\right) \mathbf{H}^T \mathbf{X}_j \mathbf{M} \ .}{\tilde{\mathbf{O}}^T \mathbf{1}_{LL} \mathbf{G}^T\, diag\left(\mathbf{1}^T \tilde{\mathbf{O}}\right) + diag\left(\mathbf{1}^T \tilde{\mathbf{O}}\right) \mathbf{H}^T \mathbf{1}_{LL} \mathbf{M}} \ , && (55)\\
\tilde{\boldsymbol{\pi}} &= \tilde{\boldsymbol{\pi}} \odot \frac{\sum_j \left[\mathbf{Y}_{::j}^T \mathbf{X}_j \odot \tilde{\mathbf{O}}^T\right] \mathbf{1}_{L1} \ .}{\sum_j \mathbf{Y}_{::j}^T \mathbf{1}_{L1} \odot \tilde{\mathbf{O}}^T \mathbf{1}_{L1}} \ . && (56)
\end{aligned}}$$

After each iteration, all matrices should be normalized column-wise.

## 3.5 $4^{\text{TH}}$ MOMENT FACTORIZATION

The KL error is:

$$D \propto \sum_j \sum_k tr\left[-\mathbf{V}_{:jk:}^T \log\left(\mathbf{P}_{:jk:}\right) + \mathbf{1}_{LL} \mathbf{P}_{:jk:}\right] \ . \tag{57}$$

The partial derivative with respect to $\tilde{\mathbf{O}}$ is:

$$
\begin{aligned}
\frac{\partial D}{\partial \tilde{\mathbf{O}}} = & -\Big[ \sum_j \sum_k \mathbf{Y}_{:jk:} \big(\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big] \bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big] \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\, \tilde{\mathbf{O}}^T\big)^T \cdots \\
& + \sum_k \Big(\mathbf{1}_{1L}\Big[\mathbf{Y}^T_{:jk:}\tilde{\mathbf{O}}\bar{\mathbf{A}} \odot \big(\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T\big)^T\Big]\Big)_{rows:\,j} \cdots \\
& + \sum_j \Big(\mathbf{1}_{1L}\Big[\mathbf{Y}^T_{:jk:}\tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}} \odot \big(\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T\big)^T\Big]\Big)_{rows:k} \cdots \\
& + \sum_j \sum_k \mathbf{Y}^T_{:jk:}\tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\Big] \cdots \\
& + \Big[\sum_j \sum_k \mathbf{1}_{LL}\big(\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T\big)^T \cdots \\
& + \sum_k \mathbf{1}_{L1}\Big[\mathbf{1}_{1L}\tilde{\mathbf{O}}\bar{\mathbf{A}} \odot \mathbf{1}_{1L}\,\big(\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T\big)^T\Big] \cdots \\
& + \sum_j \mathbf{1}_{L1}\Big[\mathbf{1}_{1L}\tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}} \odot \mathbf{1}_{1L}\,\big(\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T\big)^T\Big] \cdots \\
& + \sum_j \sum_k \mathbf{1}_{LL}\tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\Big] \qquad (58) \\
= & -\sum_j \sum_k \Big[\mathbf{Y}_{:jk:}\mathbf{G}^T_{::jk} + \mathbf{Y}^T_{:jk:}\mathbf{L}_{::jk}\Big] - \sum_k \Big(\mathbf{1}_{1L}\Big[\mathbf{Y}^T_{:jk:}\mathbf{Q} \odot \mathbf{H}^T_{::k}\Big]\Big)_{rows:\,j} - \sum_j \Big(\mathbf{1}_{1L}\Big[\mathbf{Y}^T_{:jk:}\mathbf{M}_{::j} \odot \mathbf{R}^T\Big]\Big)_{rows:k} \cdots \\
& + \mathbf{1}_{LL}\bar{\mathbf{G}}^T + \big[\mathbf{1}_{LL}\mathbf{Q} \odot \mathbf{1}_{LL}\bar{\mathbf{H}}^T\big] + \big[\mathbf{1}_{LL}\bar{\mathbf{M}} \odot \mathbf{1}_{LL}\mathbf{R}^T\big] + \mathbf{1}_{LL}\bar{\mathbf{L}}\ . \qquad (59)
\end{aligned}
$$

where:

$$
\begin{aligned}
\mathbf{Y}_{:jk:} &= \frac{\mathbf{V}_{:jk:}\ \cdot}{\mathbf{P}_{:jk:}}\ , && (60) \\
\mathbf{G}_{::jk} &= \bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T = \bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\mathbf{H}_{::k}\ , && (61) \\
\mathbf{H}_{::k} &= \bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T = \bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\mathbf{R}\ , && (62) \\
\mathbf{M}_{::j} &= \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}} = \mathbf{Q}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}}\ , && (63) \\
\mathbf{L}_{::jk} &= \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}\big(j,:\big)\big]\bar{\mathbf{A}}\, diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}}) = \mathbf{M}_{::j}\ diag\big[\tilde{\mathbf{O}}(k,:)\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\ , && (64) \\
\mathbf{Q} &= \tilde{\mathbf{O}}\bar{\mathbf{A}}\ , && (65) \\
\mathbf{R} &= \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T\ , && (66)
\end{aligned}
$$

and a bar above a matrix indicates, for example:

$$
\bar{\mathbf{G}} = \sum_j \sum_k \mathbf{G}_{::jk} = \bar{\mathbf{A}}\, diag\big[\mathbf{1}_{1L}\tilde{\mathbf{O}}\big]\bar{\mathbf{A}}\, diag\big[\mathbf{1}_{1L}\tilde{\mathbf{O}}\big]\bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}})\,\tilde{\mathbf{O}}^T\ . \qquad (67)
$$

The partial derivative with respect to $\bar{\mathbf{A}}$ is:

$$
\begin{aligned}
\frac{\partial D}{\partial \bar{\mathbf{A}}} = \sum_j \sum_k - \Big[ & \tilde{\mathbf{O}}^T \mathbf{Y}_{:jk:} \left( diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \right)^T \\
& + \left( \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \right)^T \mathbf{Y}_{:jk:} \left( diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \right)^T \\
& + \left( \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \right)^T \mathbf{Y}_{:jk:} \left( diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \right)^T \Big] \\
& + \sum_j \sum_k \Big[ \left( \mathbf{1}_{LL} \tilde{\mathbf{O}} \right)^T \left( diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \right)^T \\
& + \left( \mathbf{1}_{LL} \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \right)^T \left( diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \right)^T \\
& + \left( \mathbf{1}_{LL} \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \right)^T \left( diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T \right)^T \Big]
\end{aligned} \tag{68}
$$

$$
\begin{aligned}
&= - \sum_j \sum_k \tilde{\mathbf{O}}^T \mathbf{Y}_{:jk:} \mathbf{S}^T_{::jk} + \mathbf{U}^T_{::j} \mathbf{Y}_{:jk:} \mathbf{W}^T_{::k} + \mathbf{X}^T_{::jk} \mathbf{Y}_{:jk:} \mathbf{Z} \\
&+ \tilde{\mathbf{O}}^T \mathbf{1}_{LL} \bar{\mathbf{S}}^T + \bar{\mathbf{U}}^T \mathbf{1}_{LL} \bar{\mathbf{W}}^T + \bar{\mathbf{X}}^T \mathbf{1}_{LL} \mathbf{Z}
\end{aligned} \tag{69}
$$

$$
\begin{aligned}
&= - \left[ \tilde{\mathbf{O}}^T \left( \sum_j \sum_k \mathbf{Y}_{:jk:} \mathbf{S}^T_{::jk} \right) + \left( \sum_j \sum_k \mathbf{U}^T_{::j} \mathbf{Y}_{:jk:} \mathbf{W}^T_{::k} \right) + \left( \sum_j \sum_k \mathbf{X}^T_{::jk} \mathbf{Y}_{:jk:} \right) \mathbf{Z} \right] \\
&+ \tilde{\mathbf{O}}^T \mathbf{1}_{LL} \bar{\mathbf{S}}^T + \bar{\mathbf{U}}^T \mathbf{1}_{LL} \bar{\mathbf{W}}^T + \bar{\mathbf{X}}^T \mathbf{1}_{LL} \mathbf{Z} \ ,
\end{aligned} \tag{70}
$$

where:

$$
\mathbf{S}_{::jk} = diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T = diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, \mathbf{W}_{::k} \ , \tag{71}
$$

$$
\mathbf{U}_{::j} = \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \ , \tag{72}
$$

$$
\mathbf{W}_{::k} = diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}}\, diag(\tilde{\boldsymbol{\pi}}) \tilde{\mathbf{O}}^T = diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}}\, \mathbf{Z}^T \ , \tag{73}
$$

$$
\mathbf{X}_{::jk} = \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] = \mathbf{U}_{::j} \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \ , \tag{74}
$$

$$
\mathbf{Z} = \tilde{\mathbf{O}}\, diag(\tilde{\boldsymbol{\pi}}) \ . \tag{75}
$$

The partial derivative with respect to $\tilde{\boldsymbol{\pi}}$ is:

$$
\begin{aligned}
\frac{\partial D}{\partial \tilde{\boldsymbol{\pi}}} &= \sum_j \sum_k - \left[ \left( \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}} \right)^T \mathbf{Y}_{:jk:} \odot \tilde{\mathbf{O}}^T \right] \mathbf{1}_{L1} \\
&+ \sum_j \sum_k \left[ \left( \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}} \right)^T \mathbf{1}_{L1} \odot \tilde{\mathbf{O}}^T \mathbf{1}_{L1} \right]
\end{aligned} \tag{76}
$$

$$
= - \left[ \left( \sum_j \sum_k \mathbf{B}^T_{::jk} \mathbf{Y}_{:jk:} \right) \odot \tilde{\mathbf{O}}^T \right] \mathbf{1} + \left[ \bar{\mathbf{B}}^T \mathbf{1}_{L1} \odot \tilde{\mathbf{O}}^T \mathbf{1}_{L1} \right] \ , \tag{77}
$$

where

$$
\mathbf{B}_{::jk} = \tilde{\mathbf{O}}\bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(j,:)\right] \bar{\mathbf{A}}\, diag\left[\tilde{\mathbf{O}}(k,:)\right] \bar{\mathbf{A}} \ . \tag{78}
$$

We can simplify these expressions by packing the $\mathbf{Y}_{:jk:}$ terms into a large matrix:

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathbf{Y}}_{11} & \tilde{\mathbf{Y}}_{12} & \cdots & \tilde{\mathbf{Y}}_{1L} \\ \tilde{\mathbf{Y}}_{21} & \tilde{\mathbf{Y}}_{22} & \cdots & \tilde{\mathbf{Y}}_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{Y}}_{L1} & \tilde{\mathbf{Y}}_{L2} & \cdots & \tilde{\mathbf{Y}}_{LL}\ . \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{:11:} & \mathbf{Y}_{:12:} & \cdots & \mathbf{Y}_{:1L:} \\ \mathbf{Y}_{:21:} & \mathbf{Y}_{:22:} & \cdots & \mathbf{Y}_{:2L:} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Y}_{:L1:} & \mathbf{Y}_{:L2:} & \cdots & \mathbf{Y}_{:LL:}\ , \end{bmatrix} \tag{79}$$

and defining block-vec and block-transpose operations such that:

$$bvec\left(\tilde{\mathbf{Y}}\right) = \begin{bmatrix} \tilde{\mathbf{Y}}_{11} \\ \tilde{\mathbf{Y}}_{21} \\ \vdots \\ \tilde{\mathbf{Y}}_{L1} \\ \tilde{\mathbf{Y}}_{12} \\ \tilde{\mathbf{Y}}_{22} \\ \vdots \end{bmatrix} \quad , \quad \tilde{\mathbf{Y}}^{BT} = \begin{bmatrix} \tilde{\mathbf{Y}}_{11} & \tilde{\mathbf{Y}}_{21} & \cdots & \tilde{\mathbf{Y}}_{L1} \\ \tilde{\mathbf{Y}}_{12} & \tilde{\mathbf{Y}}_{22} & \cdots & \tilde{\mathbf{Y}}_{L2} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{Y}}_{1L} & \tilde{\mathbf{Y}}_{2L} & \cdots & \tilde{\mathbf{Y}}_{LL}\ . \end{bmatrix} \tag{80}$$

With this notation, we can write the update rules as:

$$\tilde{\mathbf{O}} = \tilde{\mathbf{O}} \odot \frac{bvec\left(\tilde{\mathbf{Y}}\right)^{BT} bvec\left(\tilde{\mathbf{G}}_T\right) + bvec\left(\tilde{\mathbf{Y}}\right)^T bvec\left(\tilde{\mathbf{L}}\right) + \left(\mathbf{1}_{1L}\left[\tilde{\mathbf{Y}}_{i:}^T \mathbf{Q} \odot \tilde{\mathbf{H}}_T\right] + \mathbf{1}_{1L}\left[\tilde{\mathbf{Y}}_{:i}^T \tilde{\mathbf{M}} \odot \mathbf{R}^T\right]\right)_{rows:i}}{\mathbf{1}_{LL}\left(\bar{\mathbf{G}}^T + \bar{\mathbf{L}}\right) + \left[\mathbf{1}_{LL}\mathbf{Q} \odot \mathbf{1}_{LL}\bar{\mathbf{H}}^T\right] + \left[\mathbf{1}_{LL}\bar{\mathbf{M}} \odot \mathbf{1}_{LL}\mathbf{R}^T\right]} \tag{81}$$

$$\bar{\mathbf{A}} = \tilde{\mathbf{A}} \odot \frac{\tilde{\mathbf{O}}^T\, bvec\left(\tilde{\mathbf{Y}}\right)^{BT} bvec\left(\tilde{\mathbf{S}}_T\right) + \tilde{\mathbf{U}}^T \tilde{\mathbf{Y}} \tilde{\mathbf{W}}_T + bvec\left(\tilde{\mathbf{X}}\right)^T bvec\left(\tilde{\mathbf{Y}}\right)\mathbf{Z}}{\tilde{\mathbf{O}}^T \mathbf{1}_{LL}\bar{\mathbf{S}}^T + \bar{\mathbf{U}}^T \mathbf{1}_{LL}\bar{\mathbf{W}}^T + \bar{\mathbf{X}}^T \mathbf{1}_{LL}\mathbf{Z}} \tag{82}$$

$$\tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}} \odot \frac{\left[bvec\left(\tilde{\mathbf{B}}\right)^T bvec\left(\tilde{\mathbf{Y}}\right) \odot \tilde{\mathbf{O}}^T\right]\mathbf{1}_{L1}}{\bar{\mathbf{B}}^T \mathbf{1}_{L1} \odot \tilde{\mathbf{O}}^T \mathbf{1}_{L1}} \tag{83}$$

where block matrices (indicated by ~'s) are defined in the same way that $\tilde{\mathbf{Y}}$ is defined and the notation $\tilde{\mathbf{S}}_T$ indicates that the block matrix contains $\mathbf{S}^T$'s. After each iteration, all matrices should be normalized column-wise.

# 4 Switching Hidden Markov Model

In this section, we derive a multiplicative update scheme to learn the parameters of a switching HMM.

## 4.1 Expressing moment matrix in terms of parameters

The probabilistic decomposition of the 2nd-order joint emission density is:

$$P(\mathbf{x}_t, \mathbf{x}_{t+1}) = \sum_{h_t} \sum_{h_{t+1}} P(h_t, h_{t+1}) \cdots \\ \sum_{r_t} \sum_{r_{t+1}} P(\mathbf{x}_{t+1} | r_{t+1}, h_{t+1}) P(r_{t+1} | r_t, h_t, h_{t+1}) P(r_t | h_t) P(\mathbf{x}_t | r_t, h_t) \ . \tag{84}$$

The matrix factorization can be expressed as:

$$\mathbf{P} = \sum_i \sum_j \mathbf{O}_i \tilde{\mathbf{A}}_{(ij)} \mathbf{O}_j^T = \tilde{\mathbf{O}} \tilde{\mathbf{A}} \tilde{\mathbf{O}}^T \ , \tag{85}$$

where $\tilde{\mathbf{A}}_{(ij)}$ denotes the $(i,j)^{\text{th}}$ block of:

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_1 B_{11} & \mathbf{A}_{12} B_{12} & \cdots & \mathbf{A}_{1K} B_{1K} \\ \mathbf{A}_{21} B_{21} & \mathbf{A}_2 B_{22} & \cdots & \mathbf{A}_{2K} B_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{K1} B_{K1} & \mathbf{A}_{K2} B_{K2} & \cdots & \mathbf{A}_K B_{KK} \end{bmatrix} diag(\tilde{\boldsymbol{\pi}}) \ , \tag{86}$$

and **B** is the switch state transition matrix.

## 4.2 Deriving Updates from Factorization

The KL error function is:

$$D(\mathbf{V}, \mathbf{P}) \propto tr\left[-\mathbf{V}^T \log(\mathbf{P}) + \mathbf{1}_{LL} \mathbf{P}\right] \ . \tag{87}$$

The partial derivatives are:

$$\frac{\partial D}{\partial \tilde{\mathbf{O}}} = -\mathbf{X} \tilde{\mathbf{O}} \tilde{\mathbf{A}}^T - \mathbf{X}^T \tilde{\mathbf{O}} \tilde{\mathbf{A}} + \mathbf{1}_{LL} \tilde{\mathbf{O}} \left(\tilde{\mathbf{A}}^T + \tilde{\mathbf{A}}\right) \ , \tag{88}$$

$$\frac{\partial D}{\partial \mathbf{A}_{ij}} = -\mathbf{O}_i^T \mathbf{X} \left(B_{ij} \pi_j \, diag\left(\boldsymbol{v}_j\right) \mathbf{O}_j^T\right)^T + \mathbf{O}_i^T \mathbf{1}_{LL} \left(B_{ij} \pi_j \, diag\left(\boldsymbol{v}_j\right) \mathbf{O}_j^T\right)^T \ , \tag{89}$$

$$\frac{\partial D}{\partial B_{ij}} = -\mathbf{1}_{1L} \left[\mathbf{X} \odot \frac{\pi_j}{B_{ij}} \mathbf{O}_i \tilde{\mathbf{A}}_{(ij)} \mathbf{O}_j^T\right] \mathbf{1}_{L1} + \mathbf{1}_{1L} \left[\frac{\pi_j}{B_{ij}} \mathbf{O}_i \tilde{\mathbf{A}}_{(ij)} \mathbf{O}_j^T\right] \mathbf{1}_{L1} \ , \tag{90}$$

$$\frac{\partial D}{\partial \tilde{\boldsymbol{\pi}}} = -\left[\tilde{\mathbf{A}}^T \tilde{\mathbf{O}}^T \mathbf{X} \odot \tilde{\mathbf{O}}^T\right] \mathbf{1}_{L1} + \tilde{\mathbf{A}}^T \tilde{\mathbf{O}}^T \mathbf{1}_{L1} \odot \tilde{\mathbf{O}}^T \mathbf{1}_{L1} \ , \tag{91}$$

where $\mathbf{X} = \frac{\mathbf{V}.}{\mathbf{P}}$. The multiplicative updates are:

$$\boxed{\begin{array}{ll} \tilde{\mathbf{O}} = \tilde{\mathbf{O}} \odot \frac{\mathbf{X}\tilde{\mathbf{O}}\tilde{\mathbf{A}}^T + \mathbf{X}^T\tilde{\mathbf{O}}\tilde{\mathbf{A}}\ .}{\mathbf{1}_{LL}\tilde{\mathbf{O}}\left(\tilde{\mathbf{A}}^T + \tilde{\mathbf{A}}\right)} \quad , & \mathbf{A}_{ij} = \mathbf{A}_{ij} \odot \frac{\mathbf{O}_i^T\mathbf{X}\mathbf{O}_j\ .}{\mathbf{O}_i^T\mathbf{1}_{LL}\mathbf{O}_j} \ , \\ B_{ij} = B_{ij} \odot \frac{\mathbf{1}_{1L}\left[\mathbf{X}\odot\mathbf{O}_i\tilde{\mathbf{A}}_{(ij)}\mathbf{O}_j^T\right]\mathbf{1}_{L1}\ .}{\mathbf{1}_{1L}\left[\mathbf{O}_i\tilde{\mathbf{A}}_{(ij)}\mathbf{O}_j^T\right]\mathbf{1}_{L1}} \quad , & \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}} \odot \frac{\left[\tilde{\mathbf{A}}^T\tilde{\mathbf{O}}^T\mathbf{X}\odot\tilde{\mathbf{O}}^T\right]\mathbf{1}_{L1}\ .}{\tilde{\mathbf{A}}^T\tilde{\mathbf{O}}^T\mathbf{1}_{L1}\odot\tilde{\mathbf{O}}^T\mathbf{1}_{L1}} \ . \end{array}} \quad \begin{array}{r}(92)\\(93)\end{array}$$

After each iteration, all matrices should be normalized column-wise.

## 4.3 3RD MOMENT MATRIX FACTORIZATION

The probabilistic decomposition of the 3rd-order joint emission density is:

$$P\left(\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{x}_{t+2}\right) = \sum_{h_t}\sum_{h_{t+1}}\sum_{h_{t+2}} P\left(h_{t+2}|h_{t+1}\right) P\left(h_{t+1}|h_t\right) P\left(h_t\right) \cdots$$
$$\sum_{r_t}\sum_{r_{t+1}}\sum_{r_{t+2}} P\left(\mathbf{x}_{t+2}|r_{t+2}, h_{t+2}\right) P\left(r_{t+2}|r_{t+1}, h_{t+1}, h_{t+2}\right) P\left(\mathbf{x}_{t+1}|r_{t+1}, h_{t+1}\right) P\left(r_{t+1}|r_t, h_t, h_{t+1}\right) P\left(r_t|h_t\right) P\left(\mathbf{x}_t|r_t, h_t\right) \ . \tag{94}$$

The corresponding moment tensor factorization is:

$$\mathbf{P}_{:j:} = \sum_i\sum_k\sum_l \mathbf{O}_i B_{ik}\mathbf{A}_{ik}\, diag\left(\mathbf{b}_j\mathbf{O}_k\right) B_{kl}\mathbf{A}_{kl}\, diag\left(\pi_l \boldsymbol{\nu}_l\right)\mathbf{O}_l^T \tag{95}$$
$$= \tilde{\mathbf{O}}\tilde{\mathbf{A}}\, diag\left(\mathbf{b}_j\tilde{\mathbf{O}}\right)\tilde{\mathbf{A}}\, diag\left(\tilde{\boldsymbol{\pi}}\right)\tilde{\mathbf{O}}^T \ , \tag{96}$$

where:

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_1 B_{11} & \mathbf{A}_{12} B_{12} & \cdots & \mathbf{A}_{1K} B_{1K} \\ \mathbf{A}_{21} B_{21} & \mathbf{A}_2 B_{22} & \cdots & \mathbf{A}_{2K} B_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{K1} B_{K1} & \mathbf{A}_{K2} B_{K2} & \cdots & \mathbf{A}_K B_{KK} \end{bmatrix} \ . \tag{97}$$

As far as $\tilde{\mathbf{O}}$ and $\tilde{\boldsymbol{\pi}}$ are concerned, the corresponding update rules are identical to those of the MHMM (3rd order). The remaining partial derivatives are:

$$\frac{\partial D}{\partial \mathbf{A}_{ik}} = -\sum_j B_{ik}\mathbf{O}_i^T\mathbf{X}_j \left(\sum_l diag\left(\mathbf{b}_j\mathbf{O}_k\right) B_{kl}\mathbf{A}_{kl}\, diag\left(\pi_l \boldsymbol{\nu}_l\right)\mathbf{O}_l^T\right)^T + \left(\sum_{i'}\mathbf{O}_{i'} B_{i'i}\mathbf{A}_{i'i}\, diag\left(\mathbf{b}_j\mathbf{O}_i\right) B_{ik}\right)^T \mathbf{X}_j \left(diag\left(\pi_k\boldsymbol{\nu}_k\right)\mathbf{O}_k^T\right)^T \tag{98}$$

$$+\left[B_{ik}\mathbf{O}_i^T\mathbf{1}_{LL}\left(\sum_l diag\left(\mathbf{1}_{1L}\mathbf{O}_k\right) B_{kl}\mathbf{A}_{kl}\, diag\left(\pi_l\boldsymbol{\nu}_l\right)\mathbf{O}_l^T\right)^T + \left(\sum_{i'}\mathbf{O}_{i'} B_{i'i}\mathbf{A}_{i'i}\, diag\left(\mathbf{1}_{1L}\mathbf{O}_i\right) B_{ik}\right)^T \mathbf{1}_{LL}\left(diag\left(\pi_k\boldsymbol{\nu}_k\right)\mathbf{O}_k^T\right)^T\right] \tag{99}$$

$$= -B_{ik}\mathbf{O}_i^T\left(\sum_j \mathbf{X}_j\mathbf{G}_{kj}^T\right) - B_{ik}\left(\sum_j \mathbf{H}_{ij}\mathbf{X}_j\right)\mathbf{O}_k\, diag\left(\pi_k\boldsymbol{\nu}_k\right) + B_{ik}\mathbf{O}_i^T\mathbf{1}_{LL}\left(\sum_j \mathbf{G}_{kj}^T\right) + B_{ik}\left(\sum_j \mathbf{H}_{ij}\right)\mathbf{1}_{LL}\mathbf{O}_k\, diag\left(\pi_k\boldsymbol{\nu}_k\right)\ , \tag{100}$$

$$\frac{\partial D}{\partial B_{ik}} = \sum_j -\mathbf{1}_{1L}\left[\mathbf{X}_j \odot \sum_l \mathbf{P}^{ikl}_{:j:}\right]\mathbf{1}_{L1} - \mathbf{1}_{1L}\left[\mathbf{X}_j \odot \sum_{i'}\mathbf{P}^{i'ik}_{:j:}\right]\mathbf{1}_{L1} + \mathbf{1}_{1L}\left[\sum_l \mathbf{P}^{ikl}_{:j:}\right]\mathbf{1}_{L1} + \mathbf{1}_{1L}\left[\sum_{i'}\mathbf{P}^{i'ik}_{:j:}\right]\mathbf{1}_{L1}\ , \tag{101}$$

where:

$$\mathbf{P}^{ikl}_{:j:} = \mathbf{O}_i B_{ik}\mathbf{A}_{ik}\, diag\left(\mathbf{b}_j\mathbf{O}_k\right) B_{kl}\mathbf{A}_{kl}\, diag\left(\pi_l\boldsymbol{\nu}_l\right)\mathbf{O}_l^T\ , \tag{102}$$

$$\mathbf{G}_{kj} = \sum_l diag\left(\mathbf{b}_j\mathbf{O}_k\right) B_{kl}\mathbf{A}_{kl}\, diag\left(\pi_l\boldsymbol{\nu}_l\right)\mathbf{O}_l^T = diag\left(\mathbf{b}_j\mathbf{O}_k\right)\tilde{\mathbf{A}}_{(k:)}\, diag\left(\tilde{\boldsymbol{\pi}}\right)\tilde{\mathbf{O}}^T\ , \tag{103}$$

$$\mathbf{H}_{ij} = \sum_{i'}\mathbf{O}_{i'} B_{i'i}\mathbf{A}_{i'i}\, diag\left(\mathbf{b}_j\mathbf{O}_i\right) = \tilde{\mathbf{O}}\tilde{\mathbf{A}}_{(:i)}\, diag\left(\mathbf{b}_j\mathbf{O}_i\right)\ . \tag{104}$$

The remaining multiplicative updates are:

$$\boxed{\begin{aligned} \mathbf{A} &= \mathbf{A}\odot\frac{\tilde{\mathbf{O}}^T\left(\sum_j \mathbf{X}_j\mathbf{G}_{:j}^T\right) + \left(\sum_j \mathbf{H}_{:j}\mathbf{X}_j\right)\tilde{\mathbf{O}}\, diag\left(\tilde{\boldsymbol{\pi}}\right)\ .}{\tilde{\mathbf{O}}^T\mathbf{1}_{LL}\left(\sum_j \mathbf{G}_{:j}^T\right) + \left(\sum_j \mathbf{H}_{:j}\right)\mathbf{1}_{LL}\tilde{\mathbf{O}}\, diag\left(\tilde{\boldsymbol{\pi}}\right)}\ , \\ B_{ik} &= B_{ik}\odot\frac{\mathbf{1}_{1L}\left[\sum_j\left(\mathbf{X}_j\odot\sum_l \mathbf{P}^{ikl}_{:j:}\right)\right]\mathbf{1}_{L1} + \mathbf{1}_{1L}\left[\sum_j\left(\mathbf{X}_j\odot\sum_{i'}\mathbf{P}^{i'ik}_{:j:}\right)\right]\mathbf{1}_{L1}\ .}{\mathbf{1}_{1L}\left[\sum_j\sum_l \mathbf{P}^{ikl}_{:j:}\right]\mathbf{1}_{L1} + \mathbf{1}_{1L}\left[\sum_j\sum_{i'}\mathbf{P}^{i'ik}_{:j:}\right]\mathbf{1}_{L1}}\ . \end{aligned}} \qquad \begin{aligned} &(105) \\ &(106) \end{aligned}$$

# 5 Factorial Hidden Markov Model

In this section, we will use the notation $\mathbf{J}_{ij}$ to indicate a matrix of ones of size $J_i \times J_j$, where $J_i$ is the number of states in the $i^{\text{th}}$ chain.

## 5.1 2$^{\text{ND}}$ moment factorization

We consider an additive emission model with uniform mixing:

$$P\left(\mathbf{x}_t|r_t^1,r_t^2\right) = \frac{1}{2}\left[P\left(\mathbf{x}_t|r_t^1\right) + P\left(\mathbf{x}_t|r_t^2\right)\right] \ . \tag{107}$$

### 5.1.1 Expressing moment matrix in terms of parameters

The joint emission distribution can be decomposed as:

$$\begin{aligned}
P\left(\mathbf{x}_t,\mathbf{x}_{t+1}\right) &= \sum_{r_t^1}\sum_{r_t^2}\sum_{r_{t+1}^1}\sum_{r_{t+1}^2} P\left(\mathbf{x}_t|r_t^1,r_t^2\right) P\left(r_t^1,r_t^2,r_{t+1}^1,r_{t+1}^2\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^1,r_{t+1}^2\right) && (108)\\
&= \frac{1}{4}\sum_{r_t^1}\sum_{r_t^2}\sum_{r_{t+1}^1}\sum_{r_{t+1}^2}\left[\sum_i P\left(\mathbf{x}_t|r_t^i\right)\right] P\left(r_t^1,r_{t+1}^1\right) P\left(r_t^2,r_{t+1}^2\right)\left[\sum_j P\left(\mathbf{x}_{t+1}|r_{t+1}^j\right)\right] && (109)\\
&\propto \sum_{r_t^1}\sum_{r_t^2}\sum_{r_{t+1}^1}\sum_{r_{t+1}^2}\left[P\left(\mathbf{x}_t|r_t^1\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^1\right) P\left(r_t^1,r_{t+1}^1\right)\right] P\left(r_t^2,r_{t+1}^2\right) \ \cdots\\
&\qquad + \left[P\left(\mathbf{x}_t|r_t^1\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^2\right) P\left(r_t^1\right) P\left(r_t^2,r_{t+1}^2\right)\right] P\left(r_{t+1}^1|r_t^1\right) \ \cdots\\
&\qquad + \left[P\left(\mathbf{x}_t|r_t^2\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^1\right) P\left(r_t^2\right) P\left(r_t^1,r_{t+1}^1\right)\right] P\left(r_{t+1}^2|r_t^2\right) \ \cdots\\
&\qquad + \left[P\left(\mathbf{x}_t|z_t^2\right) P\left(\mathbf{x}_{t+1}|z_{t+1}^2\right) P\left(z_t^2,z_{t+1}^2\right)\right] P\left(z_t^1,z_{t+1}^1\right) \ . && (110)
\end{aligned}$$

Splitting this into four terms and pushing the summations in gives:

$$\begin{aligned}
P\left(\mathbf{x}_t,\mathbf{x}_{t+1}\right) = &\sum_{r_t^1}\sum_{r_{t+1}^1} P\left(\mathbf{x}_t|r_t^1\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^1\right) P\left(r_t^1,r_{t+1}^1\right) \ \cdots\\
&+\sum_{z_t^1}\sum_{z_t^2}\sum_{r_{t+1}^2} P\left(\mathbf{x}_t|z_t^1\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^2\right) P\left(z_t^1\right) P\left(r_t^2,r_{t+1}^2\right) \ \cdots\\
&+\sum_{r_t^1}\sum_{r_t^2}\sum_{r_{t+1}^1} P\left(\mathbf{x}_t|z_t^2\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^1\right) P\left(z_t^2\right) P\left(r_t^1,r_{t+1}^1\right) \ \cdots\\
&+\sum_{r_t^2}\sum_{r_{t+1}^2} P\left(\mathbf{x}_t|r_t^2\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^2\right) P\left(r_t^2,r_{t+1}^2\right) \ , && (111)
\end{aligned}$$

which we can re-arrange as:

$$
\begin{aligned}
P\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right) = &\sum_{r_t^1}\sum_{r_{t+1}^1} P\left(\mathbf{x}_t|r_t^1\right) P\left(r_t^1, r_{t+1}^1\right) P\left(\mathbf{x}_{t+1}|z_{t+1}^1\right) \cdots \\
&+ \left(\sum_{r_t^2}\sum_{r_{t+1}^2} P\left(r_t^2, r_{t+1}^2\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^2\right)\right)\left(\sum_{r_t^1} P\left(\mathbf{x}_t|r_t^1\right) P\left(r_t^1\right)\right) \cdots \\
&+ \left(\sum_{z_t^1}\sum_{r_{t+1}^1} P\left(r_t^1, r_{t+1}^1\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^1\right)\right)\left(\sum_{r_t^2} P\left(\mathbf{x}_t|r_t^2\right) P\left(r_t^2\right)\right) \cdots \\
&+ \sum_{r_t^2}\sum_{r_{t+1}^2} P\left(\mathbf{x}_t|r_t^2\right) P\left(r_t^2, r_{t+1}^2\right) P\left(\mathbf{x}_{t+1}|r_{t+1}^2\right) \qquad (112)
\end{aligned}
$$

In matrix notation, this is:

$$
\begin{aligned}
\mathbf{P}_2 &= \mathbf{O}_1\hat{\mathbf{A}}_1\mathbf{O}_1^T + \left(\mathbf{O}_2\hat{\mathbf{A}}_2\mathbf{1}_{J_2 1}\right)\left(\mathbf{1}_{1J_1}\hat{\mathbf{A}}_1\mathbf{O}_1^T\right) + \left(\mathbf{O}_1\hat{\mathbf{A}}_1\mathbf{1}_{J_1 1}\right)\left(\mathbf{1}_{1J_2}\hat{\mathbf{A}}_2\mathbf{O}_2^T\right) + \mathbf{O}_2\hat{\mathbf{A}}_2\mathbf{O}_2^T \qquad (113) \\
&= \mathbf{O}_1\hat{\mathbf{A}}_1\mathbf{O}_1^T + \mathbf{O}_2\hat{\mathbf{A}}_2\mathbf{J}_{12}^T\hat{\mathbf{A}}_1\mathbf{O}_1^T + \mathbf{O}_1\hat{\mathbf{A}}_1\mathbf{J}_{12}\hat{\mathbf{A}}_2\mathbf{O}_2^T + \mathbf{O}_2\hat{\mathbf{A}}_2\mathbf{O}_2^T \ . \qquad (114)
\end{aligned}
$$

Grouping terms together, we can also write this as:

$$
\mathbf{P}_2 = \tilde{\mathbf{O}}\tilde{\mathbf{A}}\tilde{\mathbf{O}}^T = \begin{bmatrix} \mathbf{O}_1 & \mathbf{O}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{A}}_1 & \hat{\mathbf{A}}_1\mathbf{J}_{12}\hat{\mathbf{A}}_2 \\ \hat{\mathbf{A}}_2\mathbf{J}_{12}^T\hat{\mathbf{A}}_1 & \hat{\mathbf{A}}_2 \end{bmatrix} \begin{bmatrix} \mathbf{O}_1^T \\ \mathbf{O}_2^T \end{bmatrix} \ . \qquad (115)
$$

### 5.1.2 Deriving updates from factorization

The KL divergence error criterion is:

$$
D\left(\mathbf{V}, \mathbf{P}_2\right) \propto \ tr\left[-\mathbf{V}^T \log\left(\mathbf{P}_2\right) + \mathbf{1}_{LL}\mathbf{P}_2\right] \ . \qquad (116)
$$

The partial derivatives are:

$$
\begin{aligned}
\frac{\partial D}{\partial \mathbf{O}_1} &= -\mathbf{X}^T\mathbf{O}_1\hat{\mathbf{A}}_1 - \mathbf{X}\mathbf{O}_1\hat{\mathbf{A}}_1^T - \mathbf{X}^T\mathbf{O}_2\hat{\mathbf{A}}_2\mathbf{J}_{12}^T\hat{\mathbf{A}}_1 - \mathbf{X}\mathbf{O}_2\hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T \ \cdots \\
&\quad + \mathbf{1}_{LL}\mathbf{O}_1\left(\hat{\mathbf{A}}_1 + \hat{\mathbf{A}}_1^T\right) + \mathbf{1}_{LL}\mathbf{O}_2\left(\hat{\mathbf{A}}_2\mathbf{J}_{12}^T\hat{\mathbf{A}}_1 + \hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T\right) \ , \qquad (117) \\
\frac{\partial D}{\partial \mathbf{O}_2} &= -\mathbf{X}^T\mathbf{O}_2\hat{\mathbf{A}}_2 - \mathbf{X}\mathbf{O}_2\hat{\mathbf{A}}_2^T - \mathbf{X}^T\mathbf{O}_1\hat{\mathbf{A}}_1\mathbf{J}_{12}\hat{\mathbf{A}}_2 - \mathbf{X}\mathbf{O}_1\hat{\mathbf{A}}_1^T\mathbf{J}_{12}\hat{\mathbf{A}}_2^T \ \cdots \\
&\quad + \mathbf{1}_{LL}\mathbf{O}_2\left(\hat{\mathbf{A}}_2 + \hat{\mathbf{A}}_2^T\right) + \mathbf{1}_{LL}\mathbf{O}_1\left(\hat{\mathbf{A}}_1\mathbf{J}_{12}\hat{\mathbf{A}}_2 + \hat{\mathbf{A}}_1^T\mathbf{J}_{12}\hat{\mathbf{A}}_2^T\right) \ , \qquad (118) \\
\frac{\partial D}{\partial \hat{\mathbf{A}}_1} &= -\mathbf{O}_1^T\mathbf{X}\mathbf{O}_1 - \mathbf{O}_1^T\mathbf{X}\mathbf{O}_2\hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T - \mathbf{J}_{12}\hat{\mathbf{A}}_2^T\mathbf{O}_2^T\mathbf{X}\mathbf{O}_1 \ \cdots \\
&\quad + \mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_1 + \mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2\hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T + \mathbf{J}_{12}\hat{\mathbf{A}}_2^T\mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_1 \ , \qquad (119) \\
\frac{\partial D}{\partial \hat{\mathbf{A}}_2} &= -\mathbf{O}_2^T\mathbf{X}\mathbf{O}_2 - \mathbf{O}_2^T\mathbf{X}\mathbf{O}_1\hat{\mathbf{A}}_1^T\mathbf{J}_{12} - \mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T\mathbf{O}_1^T\mathbf{X}\mathbf{O}_2 \ \cdots \\
&\quad + \mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_2 + \mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_1\hat{\mathbf{A}}_1^T\mathbf{J}_{12} + \mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2 \ , \qquad (120)
\end{aligned}
$$

where $\mathbf{X} = \frac{\mathbf{V}_{\cdot}}{\mathbf{P}_2}$. The resulting multiplicative updates are:

$$\mathbf{O}_1 = \mathbf{O}_1 \odot \frac{\mathbf{X}^T\mathbf{O}_1\hat{\mathbf{A}}_1 + \mathbf{X}\mathbf{O}_1\hat{\mathbf{A}}_1^T + \mathbf{X}^T\mathbf{O}_2\hat{\mathbf{A}}_2\mathbf{J}_{12}^T\hat{\mathbf{A}}_1 + \mathbf{X}\mathbf{O}_2\hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T}{\mathbf{1}_{LL}\mathbf{O}_1\left(\hat{\mathbf{A}}_1 + \hat{\mathbf{A}}_1^T\right) + \mathbf{1}_{LL}\mathbf{O}_2\left(\hat{\mathbf{A}}_2\mathbf{J}_{12}^T\hat{\mathbf{A}}_1 + \hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T\right)} , \tag{121}$$

$$\mathbf{O}_2 = \mathbf{O}_2 \odot \frac{\mathbf{X}^T\mathbf{O}_2\hat{\mathbf{A}}_2 + \mathbf{X}\mathbf{O}_2\hat{\mathbf{A}}_2^T + \mathbf{X}^T\mathbf{O}_1\hat{\mathbf{A}}_1\mathbf{J}_{12}\hat{\mathbf{A}}_2 + \mathbf{X}\mathbf{O}_1\hat{\mathbf{A}}_1^T\mathbf{J}_{12}\hat{\mathbf{A}}_2^T}{\mathbf{1}_{LL}\mathbf{O}_2\left(\hat{\mathbf{A}}_2 + \hat{\mathbf{A}}_2^T\right) + \mathbf{1}_{LL}\mathbf{O}_1\left(\hat{\mathbf{A}}_1\mathbf{J}_{12}\hat{\mathbf{A}}_2 + \hat{\mathbf{A}}_1^T\mathbf{J}_{12}\hat{\mathbf{A}}_2^T\right)} , \tag{122}$$

$$\hat{\mathbf{A}}_1 = \hat{\mathbf{A}}_1 \odot \frac{\mathbf{O}_1^T\mathbf{X}\mathbf{O}_1 + \mathbf{O}_1^T\mathbf{X}\mathbf{O}_2\hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T + \mathbf{J}_{12}\hat{\mathbf{A}}_2^T\mathbf{O}_2^T\mathbf{X}\mathbf{O}_1}{\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_1 + \mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2\hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T + \mathbf{J}_{12}\hat{\mathbf{A}}_2^T\mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_1} , \tag{123}$$

$$\hat{\mathbf{A}}_2 = \hat{\mathbf{A}}_2 \odot \frac{\mathbf{O}_2^T\mathbf{X}\mathbf{O}_2 + \mathbf{O}_2^T\mathbf{X}\mathbf{O}_1\hat{\mathbf{A}}_1^T\mathbf{J}_{12} + \mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T\mathbf{O}_1^T\mathbf{X}\mathbf{O}_2}{\mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_2 + \mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_1\hat{\mathbf{A}}_1^T\mathbf{J}_{12} + \mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2} . \tag{124}$$

Taking advantage of normalization properties of the parameter matrices, we can simplify the denominators to:

$$2\mathbf{1}_{L1}\left(\hat{\mathbf{A}}_1 + \hat{\mathbf{A}}_1^T\right) \quad , \quad 2\mathbf{1}_{L2}\left(\hat{\mathbf{A}}_2 + \hat{\mathbf{A}}_2^T\right) \quad , \quad 3\mathbf{J}_{11} \quad , \quad 3\mathbf{J}_{22} . \tag{125}$$

Normalizing over the columns of $\mathbf{O}_i$ and over the entire $\hat{\mathbf{A}}_i$ gives a matrix implementation of the corresponding EM algorithm. We can write the updates more compactly and generally by invoking the block forms of $\tilde{\mathbf{O}}$ and $\tilde{\mathbf{A}}$:

$$\boxed{\begin{aligned} \tilde{\mathbf{O}} &= \tilde{\mathbf{O}} \odot \frac{\mathbf{X}^T\tilde{\mathbf{O}}\tilde{\mathbf{A}} + \mathbf{X}\tilde{\mathbf{O}}\tilde{\mathbf{A}}^T .}{\mathbf{1}_{LL}\tilde{\mathbf{O}}\left(\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T\right)} , & (126) \\ \hat{\mathbf{A}}_j &= \hat{\mathbf{A}}_j \odot \frac{\mathbf{S}_j^T\mathbf{R}_1\left(\tilde{\mathbf{O}}^T\mathbf{X}\tilde{\mathbf{O}} \odot \mathbf{Q}_j\right)\mathbf{R}_2\mathbf{S}_j}{\mathbf{S}_j^T\mathbf{R}_1\left(\tilde{\mathbf{O}}^T\mathbf{1}_{LL}\tilde{\mathbf{O}} \odot \mathbf{Q}_j\right)\mathbf{R}_2\mathbf{S}_j} , & (127) \end{aligned}}$$

where, for the case of two HMMs:

$$\mathbf{R}_1 = \begin{bmatrix} \mathbf{I} & \mathbf{J}_{12}\hat{\mathbf{A}}_2^T \\ \mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T & \mathbf{I} \end{bmatrix} \quad , \quad \mathbf{R}_2 = \begin{bmatrix} \mathbf{I} & \hat{\mathbf{A}}_1^T\mathbf{J}_{12} \\ \hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T & \mathbf{I} \end{bmatrix} , \tag{128}$$

$$\mathbf{S}_1 = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \quad , \quad \mathbf{S}_2 = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} , \tag{129}$$

$$\mathbf{Q}_1 = \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{12}^T & \mathbf{0} \end{bmatrix} \quad , \quad \mathbf{Q}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{J}_{12} \\ \mathbf{J}_{12}^T & \mathbf{J}_{22} \end{bmatrix} . \tag{130}$$

and for the case of three HMMs:

$$\mathbf{R}_1 = \begin{bmatrix} \mathbf{I} & \mathbf{J}_{12}\hat{\mathbf{A}}_2^T & \mathbf{J}_{13}\hat{\mathbf{A}}_3^T \\ \mathbf{J}_{12}^T\hat{\mathbf{A}}_1^T & \mathbf{I} & \mathbf{J}_{23}\hat{\mathbf{A}}_3^T \\ \mathbf{J}_{13}^T\hat{\mathbf{A}}_1^T & \mathbf{J}_{23}^T\hat{\mathbf{A}}_2^T & \mathbf{I} \end{bmatrix} \quad , \quad \mathbf{R}_2 = \begin{bmatrix} \mathbf{I} & \hat{\mathbf{A}}_1^T\mathbf{J}_{12} & \hat{\mathbf{A}}_1^T\mathbf{J}_{13} \\ \hat{\mathbf{A}}_2^T\mathbf{J}_{12}^T & \mathbf{I} & \hat{\mathbf{A}}_2^T\mathbf{J}_{23} \\ \hat{\mathbf{A}}_3^T\mathbf{J}_{13}^T & \hat{\mathbf{A}}_3^T\mathbf{J}_{23}^T & \mathbf{I} \end{bmatrix} \quad , \tag{131}$$

$$\mathbf{S}_1 = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad , \quad \mathbf{S}_2 = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \\ \mathbf{0} \end{bmatrix} \quad , \quad \mathbf{S}_3 = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{I} \end{bmatrix} \quad , \tag{132}$$

$$\mathbf{Q}_1 = \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} & \mathbf{J}_{13} \\ \mathbf{J}_{12}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{J}_{13}^T & \mathbf{0} & \mathbf{0} \end{bmatrix} \quad , \quad \mathbf{Q}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{J}_{12} & \mathbf{0} \\ \mathbf{J}_{12}^T & \mathbf{J}_{22} & \mathbf{J}_{23} \\ \mathbf{0} & \mathbf{J}_{23}^T & \mathbf{0} \end{bmatrix} \quad , \quad \mathbf{Q}_3 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{J}_{13} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{23} \\ \mathbf{J}_{13}^T & \mathbf{J}_{23}^T & \mathbf{J}_{33} \end{bmatrix} \quad . \tag{133}$$

## 5.2 3<sup>RD</sup> MOMENT FACTORIZATION

In this section, we derive the factorization of the 3th moment tensor in terms of the FHMM parameters. We also derive the corresponding multiplicative updates.

### 5.2.1 EXPRESSING MOMENT MATRIX IN TERMS OF PARAMETERS

We can decompose the third-order joint emission density as:

$$P\left(\mathbf{x}_t, \mathbf{x}_{t+1} = j, \mathbf{x}_{t+2}\right) \tag{134}$$

$$= \sum_{r^{1:2}_{t:t+2}} P\left(\mathbf{x}_{t+2} | r^{1:2}_{t+2}\right) P\left(r^{1:2}_{t+2} | r^{1:2}_{t+1}\right) P\left(\mathbf{x}_{t+1} = j | r^{1:2}_{t+1}\right) P\left(r^{1:2}_{t+1} | r^{1:2}_{t}\right) P\left(r^{1:2}_{t}\right) P\left(\mathbf{x}_t | r^{1:2}_{t}\right) \tag{135}$$

$$= \frac{1}{8} \sum_{r^{1:2}_{t:t+2}} \left[\sum_a P\left(\mathbf{x}_{t+2} | r^{a}_{t+2}\right)\right] P\left(r^{1:2}_{t+2} | r^{1:2}_{t+1}\right) \left[\sum_b P\left(\mathbf{x}_{t+1} = j | r^{b}_{t+1}\right)\right] P\left(r^{1:2}_{t+1} | r^{1:2}_{t}\right) P\left(r^{1:2}_{t}\right) \left[\sum_c P\left(\mathbf{x}_t | r^{c}_{t}\right)\right] . \tag{136}$$

Following the same procedure as for the second-order case, we arrive at a matrix expression for the third-order moment tensor factorization. We omit these tedious manipulations for the sake of clarity. The factorization is given as:

$$
\begin{aligned}
\mathbf{P}_{:j:} \propto\ & \mathbf{O}_1\mathbf{A}_1\, diag\left[\mathbf{O}_1\left(j\,,:\right)\right]\mathbf{A}_1\, diag\left(\boldsymbol{v}_1\right)\mathbf{O}_1^T \\
& + \left(\mathbf{O}_1\mathbf{A}_1\, diag\left[\mathbf{O}_1\left(j\,,:\right)\right]\mathbf{A}_1\boldsymbol{v}_1\right)\left(\mathbf{O}_2\boldsymbol{v}_2\right)^T \\
& + \left(\mathbf{O}_2\left(j\,,:\right)\mathbf{A}_2\boldsymbol{v}_2\right)\left(\mathbf{O}_1\mathbf{A}_1\mathbf{A}_1\, diag\left(\boldsymbol{v}_1\right)\mathbf{O}_1^T\right) \\
& + \left(\mathbf{O}_1\mathbf{A}_1\mathbf{A}_1\boldsymbol{v}_1\right)\left(\mathbf{O}_2\left(j\,,:\right)\mathbf{A}_2\, diag\left(\boldsymbol{v}_2\right)\mathbf{O}_2^T\right) \\
& + \left(\mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\boldsymbol{v}_2\right)\left(\mathbf{O}_1\left(j\,,:\right)\mathbf{A}_1\, diag\left(\boldsymbol{v}_1\right)\mathbf{O}_1^T\right) \\
& + \left(\mathbf{O}_1\left(j\,,:\right)\mathbf{A}_1\boldsymbol{v}_1\right)\left(\mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\, diag\left(\boldsymbol{v}_2\right)\mathbf{O}_2^T\right) \\
& + \left(\mathbf{O}_2\mathbf{A}_2\, diag\left[\mathbf{O}_2\left(j\,,:\right)\right]\mathbf{A}_2\boldsymbol{v}_2\right)\left(\mathbf{O}_1\boldsymbol{v}_1\right)^T \\
& + \mathbf{O}_2\mathbf{A}_2\, diag\left[\mathbf{O}_2\left(j\,,:\right)\right]\mathbf{A}_2\, diag\left(\boldsymbol{v}_2\right)\mathbf{O}_2^T\ .
\end{aligned}
\tag{137}
$$

We can introduce a row vector $\mathbf{b}_j^T$ that selects the $j^{\text{th}}$ row of $\mathbf{O}_i$, i.e. $\mathbf{O}\left(j\,,:\right) = \mathbf{b}_j^T\mathbf{O}_i$. Grouping terms into corresponding pairs, we have:

$$
\begin{aligned}
\mathbf{P}_{:j:} \propto\ & \mathbf{O}_1\mathbf{A}_1\, diag\left[\mathbf{b}_j^T\mathbf{O}_1\right]\mathbf{A}_1\, diag\left(\boldsymbol{v}_1\right)\mathbf{O}_1^T + \mathbf{O}_2\mathbf{A}_2\, diag\left[\mathbf{b}_j^T\mathbf{O}_2\right]\mathbf{A}_2\, diag\left(\boldsymbol{v}_2\right)\mathbf{O}_2^T \\
& + \mathbf{O}_1\mathbf{A}_1\, diag\left[\mathbf{b}_j^T\mathbf{O}_1\right]\mathbf{A}_1\boldsymbol{v}_1\boldsymbol{v}_2^T\mathbf{O}_2^T + \mathbf{O}_2\mathbf{A}_2\, diag\left[\mathbf{b}_j^T\mathbf{O}_2\right]\mathbf{A}_2\boldsymbol{v}_2\boldsymbol{v}_1^T\mathbf{O}_1^T \\
& + \left(\mathbf{b}_j^T\mathbf{O}_1\mathbf{A}_1\boldsymbol{v}_1\right)\mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\, diag\left(\boldsymbol{v}_2\right)\mathbf{O}_2^T + \left(\mathbf{b}_j^T\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{O}_1\mathbf{A}_1\mathbf{A}_1\, diag\left(\boldsymbol{v}_1\right)\mathbf{O}_1^T \\
& + \mathbf{O}_1\mathbf{A}_1\mathbf{A}_1\boldsymbol{v}_1\mathbf{b}_j^T\mathbf{O}_2\mathbf{A}_2\, diag\left(\boldsymbol{v}_2\right)\mathbf{O}_2^T + \mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\boldsymbol{v}_2\mathbf{b}_j^T\mathbf{O}_1\mathbf{A}_1\, diag\left(\boldsymbol{v}_1\right)\mathbf{O}_1^T\ .
\end{aligned}
\tag{138}
$$

### 5.2.2 DERIVING UPDATES FROM FACTORIZATION

The KL error is:

$$
D \propto \sum_j tr\left[-\mathbf{V}_{:j:}^T \log\left(\mathbf{P}_{:j:}\right) + \mathbf{1}_{LL}\,\mathbf{P}_{:j:}\right]\ , \tag{139}
$$

and the partial derivatives with respect to the first chain's parameters are:

$$
\begin{aligned}
\frac{\partial D}{\partial \mathbf{O}_1} = -\sum_j \Big[ & \mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_1\right]\mathbf{A}_1^T + \mathbf{b}_j\mathbf{1}_{1L}\left[\mathbf{X}_j^T\mathbf{O}_1\mathbf{A}_1 \odot \mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\right] \\
& + \mathbf{X}_j^T\mathbf{O}_1\mathbf{A}_1\, \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_1\right]\mathbf{A}_1\, \mathit{diag}(\boldsymbol{v}_1) + \mathbf{X}_j\mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_1\right]\mathbf{A}_1^T \\
& + \mathbf{b}_j\mathbf{1}_{1L}\left[\left(\mathbf{X}_j^3\right)^T\mathbf{O}_1\mathbf{A}_1 \odot \mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T\mathbf{A}_1^T\right] + \left(\mathbf{X}_j^4\right)^T\mathbf{O}_2\mathbf{A}_2\, \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_2\right]\mathbf{A}_2\boldsymbol{v}_2\boldsymbol{v}_1^T \\
& + \mathbf{b}_j\mathbf{1}_{1L}\left[\mathbf{X}_j \odot \mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{O}_2^T\right]\mathbf{1}_{L1}\boldsymbol{v}_1^T\mathbf{A}_1^T + \left(\mathbf{b}_j^T\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\mathbf{A}_1^T \\
& + \left(\mathbf{b}_j^T\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{X}_j^T\mathbf{O}_1\mathbf{A}_1\mathbf{A}_1\, \mathit{diag}(\boldsymbol{v}_1) + \mathbf{X}_j\mathbf{O}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{b}_j\boldsymbol{v}_1^T\mathbf{A}_1^T\mathbf{A}_1^T \\
& + \mathbf{b}_j\boldsymbol{v}_2^T\mathbf{A}_2^T\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T + \mathbf{X}_j^T\mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\boldsymbol{v}_2\mathbf{b}_j^T\mathbf{O}_1\mathbf{A}_1\, \mathit{diag}(\boldsymbol{v}_1)\Big] \\
& + \Big[\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_1\right]\mathbf{A}_1^T + \mathbf{1}_{L1}\left(\mathbf{1}_{1L}\mathbf{O}_1\mathbf{A}_1 \odot \mathbf{1}_{1L}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\right) \\
& + \mathbf{1}_{LL}\mathbf{O}_1\mathbf{A}_1\, \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_1\right]\mathbf{A}_1\, \mathit{diag}(\boldsymbol{v}_1) + \mathbf{1}_{LL}\mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_1\right]\mathbf{A}_1^T \\
& + \mathbf{1}_{L1}\left(\mathbf{1}_{1L}\mathbf{O}_1\mathbf{A}_1 \odot \mathbf{1}_{1L}\mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T\mathbf{A}_1^T\right) + \mathbf{1}_{LL}\mathbf{O}_2\mathbf{A}_2\, \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_2\right]\mathbf{A}_2\boldsymbol{v}_2\boldsymbol{v}_1^T \\
& + \mathbf{1}_{LL}\mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{O}_2^T\mathbf{1}_{L1}\boldsymbol{v}_1^T\mathbf{A}_1^T + \left(\mathbf{1}_{1L}\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\mathbf{A}_1^T \\
& + \left(\mathbf{1}_{1L}\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{1}_{LL}\mathbf{O}_1\mathbf{A}_1\mathbf{A}_1\, \mathit{diag}(\boldsymbol{v}_1) + \mathbf{1}_{LL}\mathbf{O}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{1}_{L1}\boldsymbol{v}_1^T\mathbf{A}_1^T\mathbf{A}_1^T \\
& + \mathbf{1}_{L1}\boldsymbol{v}_2^T\mathbf{A}_2^T\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T + \mathbf{1}_{LL}\mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\boldsymbol{v}_2\mathbf{1}_{1L}\mathbf{O}_1\mathbf{A}_1\, \mathit{diag}(\boldsymbol{v}_1)\Big] ,
\end{aligned}
\tag{140}
$$

$$
\begin{aligned}
\frac{\partial D}{\partial \mathbf{A}_1} = -\sum_j \Big[ & \mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_1\right] + \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_1\right]\mathbf{A}_1^T\mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1) \\
& + \mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_1\right] + \mathit{diag}\left[\mathbf{b}_j^T\mathbf{O}_1\right]\mathbf{A}_1^T\mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T \\
& + \mathbf{O}_1^T\mathbf{b}_j\mathbf{1}_{1L}\left[\mathbf{X}_j \odot \mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{O}_2^T\right]\mathbf{1}_{L1}\boldsymbol{v}_1^T \\
& + \left(\mathbf{b}_j^T\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T + \left(\mathbf{b}_j^T\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{A}_1^T\mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1) \\
& + \mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{b}_j\boldsymbol{v}_1^T\mathbf{A}_1^T + \mathbf{A}_1^T\mathbf{O}_1^T\mathbf{X}_j\mathbf{O}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{b}_j\boldsymbol{v}_1^T \\
& + \mathbf{O}_1^T\mathbf{b}_j\boldsymbol{v}_2^T\mathbf{A}_2^T\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{X}_j\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\Big] \\
& + \Big[\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_1\right] + \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_1\right]\mathbf{A}_1^T\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1) \\
& + \mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T\mathbf{A}_1^T\, \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_1\right] + \mathit{diag}\left[\mathbf{1}_{1L}\mathbf{O}_1\right]\mathbf{A}_1^T\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2\boldsymbol{v}_2\boldsymbol{v}_1^T \\
& + \mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2\mathbf{A}_2\mathbf{A}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{O}_2^T\mathbf{1}_{L1}\boldsymbol{v}_1^T \\
& + \left(\mathbf{1}_{1L}\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\mathbf{A}_1^T + \left(\mathbf{1}_{1L}\mathbf{O}_2\mathbf{A}_2\boldsymbol{v}_2\right)\mathbf{A}_1^T\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1) \\
& + \mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{1}_{L1}\boldsymbol{v}_1^T\mathbf{A}_1^T + \mathbf{A}_1^T\mathbf{O}_1^T\mathbf{1}_{LL}\mathbf{O}_2\, \mathit{diag}(\boldsymbol{v}_2)\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{1}_{L1}\boldsymbol{v}_1^T \\
& + \mathbf{O}_1^T\mathbf{1}_{1L}\boldsymbol{v}_2^T\mathbf{A}_2^T\mathbf{A}_2^T\mathbf{O}_2^T\mathbf{1}_{LL}\mathbf{O}_1\, \mathit{diag}(\boldsymbol{v}_1)\Big] ,
\end{aligned}
\tag{141}
$$

$$
\begin{aligned}
\frac{\partial D}{\partial \boldsymbol{v}_1} = -\sum_j \Bigg[ & \left(\mathbf{A}_1^T \, diag\left[\mathbf{b}_j^T \mathbf{O}_1\right] \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{X}_j \odot \mathbf{O}_1^T\right) \mathbf{1} + \mathbf{A}_1^T \, diag\left[\mathbf{b}_j^T \mathbf{O}_1\right] \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{X}_j \mathbf{O}_2 \boldsymbol{v}_2 \\
& + \mathbf{O}_1^T \mathbf{X}_j^T \mathbf{O}_2 \mathbf{A}_2 \, diag\left[\mathbf{b}_j^T \mathbf{O}_2\right] \mathbf{A}_2 \boldsymbol{v}_2 + \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{b}_j \mathbf{1}_{1L} \left[\mathbf{X}_j \odot \mathbf{O}_2 \mathbf{A}_2 \mathbf{A}_2 \, diag\left(\boldsymbol{v}_2\right) \mathbf{O}_2^T\right] \mathbf{1}_{L1} \\
& + \left(\mathbf{b}_j^T \mathbf{O}_2 \mathbf{A}_2 \boldsymbol{v}_2\right) \left(\mathbf{A}_1^T \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{X}_j \odot \mathbf{O}_1^T\right) \mathbf{1}_{L1} + \mathbf{A}_1^T \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{X}_j \mathbf{O}_2 \, diag\left(\boldsymbol{v}_2\right) \mathbf{A}_2^T \mathbf{O}_2^T \mathbf{b}_j \\
& + \left(\mathbf{A}_1^T \mathbf{O}_1^T \mathbf{b}_j \boldsymbol{v}_2^T \mathbf{A}_2^T \mathbf{A}_2^T \mathbf{O}_2^T \mathbf{X}_j \odot \mathbf{O}_1^T\right) \mathbf{1}_{L1} \Bigg] \\
& + \Bigg[ \left(\mathbf{A}_1^T \, diag\left[\mathbf{1}_{1L} \mathbf{O}_1\right] \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{1}_{L1} \odot \mathbf{O}_1^T \mathbf{1}_{L1}\right) + \mathbf{A}_1^T \, diag\left[\mathbf{1}_{1L} \mathbf{O}_1\right] \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{1}_{LL} \mathbf{O}_2 \boldsymbol{v}_2 \\
& + \mathbf{O}_1^T \mathbf{1}_{LL} \mathbf{O}_2 \mathbf{A}_2 \, diag\left[\mathbf{1}_{1L} \mathbf{O}_2\right] \mathbf{A}_2 \boldsymbol{v}_2 + \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{1}_{LL} \mathbf{O}_2 \mathbf{A}_2 \mathbf{A}_2 \, diag\left(\boldsymbol{v}_2\right) \mathbf{O}_2^T \mathbf{1}_{L1} \\
& + \left(\mathbf{1}_{1L} \mathbf{O}_2 \mathbf{A}_2 \boldsymbol{v}_2\right) \left(\mathbf{A}_1^T \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{1}_{L1} \odot \mathbf{O}_1^T \mathbf{1}_{L1}\right) + \mathbf{A}_1^T \mathbf{A}_1^T \mathbf{O}_1^T \mathbf{1}_{LL} \mathbf{O}_2 \, diag\left(\boldsymbol{v}_2\right) \mathbf{A}_2^T \mathbf{O}_2^T \mathbf{1}_{L1} \\
& + \left(\mathbf{A}_1^T \mathbf{O}_1^T \mathbf{1}_{L1} \boldsymbol{v}_2^T \mathbf{A}_2^T \mathbf{A}_2^T \mathbf{O}_2^T \mathbf{1}_{L1} \odot \mathbf{O}_1^T \mathbf{1}_{L1}\right) \Bigg] \ ,
\end{aligned}
\tag{142}
$$

where $\mathbf{X}_j = \frac{\mathbf{V}_{:j:}}{\mathbf{P}_{:j:}}$. Multiplicative update rules can be derived from these derivatives by placing the negative and positive portions in the numerator and denominator of the multiplicative factor, respectively. Taking advantage of normalization properties of the parameter matrices, the denominator terms simplify to:

$$
4\mathbf{1}_{L1}\left[\left(\mathbf{A}_1^2 + \mathbf{A}_1 + \mathbf{I}\right)\boldsymbol{v}_1\right]^T \quad , \quad \mathbf{1}_{J_1,1}\left[\left(4\mathbf{A}_1 + 6\mathbf{I}\right)\boldsymbol{v}_1\right]^T \quad , \quad 7\mathbf{1}_{J_1,1} \ . \tag{143}
$$

The derivatives for the second chain's parameters are analogous, with $\mathbf{A}_1$ swapped with $\mathbf{A}_2$ and so on for the other parameters. After each iteration, all matrices should be normalized column-wise.

# REFERENCES

[Ding et al.(2008)Ding, Li, and Peng] Ding, C., Li, T., and Peng, W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52(8):3913–3927, 2008.

[Hofmann(1999)] Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. ACM, 1999.

[Lee & Seung(2001)Lee and Seung] Lee, D and Seung, H S. Algorithms for non-negative matrix factorization. *Conference on Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[Magnus & Neudecker(1999)Magnus and Neudecker] Magnus, J. R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, 2nd edition, 1999.