A APPENDIX

In this supplementary material, we provide the details of the example derivative computations from Section 2.2, give a justification of the NCPD expansion formula exploited in Section 2.1, and provide further experimental results that we were not able to include in Section 3.

EXAMPLE DERIVATIONS

Here, we justify the derivations provided in the example in Section 2.2. We note that Anaissi et al. (2020); Kolda & Hong (2019) provide similar derivations for the CP tensor decomposition, but their decompositions do not attempt to further decompose the CP factor matrices, and thus, their results are not sufficient for providing derivatives with respect to the $\boldsymbol{A}$ and $\boldsymbol{S}$ matrices. Consider the full reconstruction loss function for the order-k tensor $\boldsymbol{\mathsf{X}}$,

$$C = \|\boldsymbol{\mathsf{X}} - [\![\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k]\!]\|_F^2,$$

where for some fixed $1 \leq i \leq k$, $\boldsymbol{X}_i = \boldsymbol{ABC}$ and consider the gradient $\dfrac{\partial C}{\partial \boldsymbol{B}}$. Let

$$\boldsymbol{H}_i = \boldsymbol{X}_k \odot \ldots \odot \boldsymbol{X}_{i+1} \odot \boldsymbol{X}_{i-1} \odot \ldots \odot \boldsymbol{X}_1.$$

Now, we let $\widetilde{\boldsymbol{\mathsf{X}}} = [\![\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k]\!]$. Then, if $\widetilde{\boldsymbol{X}}_{(i)}$ is the mode-i matricization of $\widetilde{\boldsymbol{\mathsf{X}}}$ (see e.g., (Kolda & Bader, 2009)), we have

$$\widetilde{\boldsymbol{X}}_{(i)} = \boldsymbol{X}_i(\boldsymbol{X}_k \odot \ldots \odot \boldsymbol{X}_{i+1} \odot \boldsymbol{X}_{i-1} \odot \ldots \odot \boldsymbol{X}_1)^\top] = \boldsymbol{X}_i \boldsymbol{H}_i^\top.$$

Thus, if we let $\boldsymbol{X}_{(i)}$ be the mode-i matricization of $\boldsymbol{\mathsf{X}}$, we have that

$$C = \|\boldsymbol{X}_{(i)} - \widetilde{\boldsymbol{X}}_{(i)}\|_F^2 = \|\boldsymbol{X}_{(i)} - (\boldsymbol{ABC})(\boldsymbol{X}_k \odot \ldots \odot \boldsymbol{X}_{i+1} \odot \boldsymbol{X}_{i-1} \odot \ldots \odot \boldsymbol{X}_1)^\top\|_F^2.$$

Now, we compute the desired gradient through a series of applications of the chain rule. We then see that

$$\begin{aligned}
\frac{\partial C}{\partial \boldsymbol{B}} &= \frac{\partial C}{\partial \boldsymbol{X}_i \boldsymbol{H}_i^\top} \frac{\partial \boldsymbol{X}_i \boldsymbol{H}_i^\top}{\partial \boldsymbol{ABC}} \frac{\partial \boldsymbol{ABC}}{\partial \boldsymbol{B}} \\
&= \boldsymbol{A}^\top \Big(\frac{\partial C}{\partial \boldsymbol{X}_i \boldsymbol{H}_i^\top} \frac{\partial \boldsymbol{X}_i \boldsymbol{H}_i^\top}{\partial \boldsymbol{X}_i} \Big) \boldsymbol{C}^\top \\
&= 2\boldsymbol{A}^\top \big(\boldsymbol{X}_{(i)} - \boldsymbol{X}_i \boldsymbol{H}_i^\top \big) \boldsymbol{H}_i \boldsymbol{C}^\top.
\end{aligned}$$

Now, using the calculations above we can proceed in calculating $\dfrac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}}$. Gao et al. (2019) show that if $\left(\dfrac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}}\right)^{\boldsymbol{S}}$ denotes derivative of C with respect to $\boldsymbol{A}_i^{(\ell_j)}$, holding the $\boldsymbol{S}$ matrices constant, then we have

$$\frac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}} = \left(\frac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_1)}}\right)^{\boldsymbol{S}} + \sum_{\substack{\ell_1 \leq \ell_2 \leq \mathcal{L}-2 \\ 1 \leq j \leq r}} \boldsymbol{U}_i^{(\ell_1,\ell_2),j}$$

where $\boldsymbol{U}_i^{(\ell_1,\ell_2),j}$ relates C to $\boldsymbol{A}_i^{(\ell_1)}$ through $\boldsymbol{S}_i^{(\ell_2)}$ and $\boldsymbol{S}_i^{(\ell_1)}$, is defined column-wise (j), and depends upon $\left(\dfrac{\partial C}{\partial \boldsymbol{S}_i^{(\ell_2)}}\right)^*$, the derivative of C with respect to $\boldsymbol{S}_i^{(\ell_2)}$ holding $\boldsymbol{S}_i^{(\ell_2+1)}, \ldots, \boldsymbol{S}_i^{(\mathcal{L}-2)}$ constant. Thus, $\left(\dfrac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}}\right)^{\boldsymbol{S}}$ and $\left(\dfrac{\partial C}{\partial \boldsymbol{S}_i^{(\ell_j)}}\right)^*$ are sufficient to calculate $\dfrac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}}$. We calculate the gradient $\left(\dfrac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}}\right)^{\boldsymbol{S}}$ where

$$C = \|\boldsymbol{\mathsf{X}} - [\![\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2, \ldots, \widetilde{\boldsymbol{X}}_k]\!]\|_F^2$$

and $\widetilde{\boldsymbol{X}}_i = \boldsymbol{A}_i^{(0)}\boldsymbol{A}_i^{(1)}...\boldsymbol{A}_i^{(\mathcal{L}-2)}\boldsymbol{S}_i^{(\mathcal{L}-2)}$. Since we can assume that $\boldsymbol{A}_i^{(\ell_1)}$ is independent of all other $\boldsymbol{A}$'s and $\boldsymbol{S}$'s, we have that

$$\left(\frac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}}\right)^{\boldsymbol{S}} = 2\left(\boldsymbol{A}_i^{(0)}\boldsymbol{A}_i^{(1)}...\boldsymbol{A}_i^{(\ell_j-1)}\right)^\top \left(\boldsymbol{X}_{(i)} - \boldsymbol{X}_i\boldsymbol{H}_i^\top\right)\boldsymbol{H}_i\left(\boldsymbol{A}_i^{(\ell_{j+1})}...\boldsymbol{A}_i^{(\mathcal{L}-2)}\boldsymbol{S}_i^{(\mathcal{L}-2)}\right)^\top. \tag{17}$$

Now, we calculate $\left(\frac{\partial C}{\partial \boldsymbol{S}_i^{(\ell_j)}}\right)^*$. Since we can assume that $\boldsymbol{S}_i^{(\ell_j)}$ is independent of all other $\boldsymbol{A}$'s and $\boldsymbol{S}$'s, we have that

$$\left(\frac{\partial C}{\partial \boldsymbol{S}_i^{(\ell_j)}}\right)^* = 2\left(\boldsymbol{A}_i^{(0)}\boldsymbol{A}_i^{(1)}...\boldsymbol{A}_i^{(\ell_j)}\right)^\top \left(\boldsymbol{X}_{(i)} - \boldsymbol{X}_i\boldsymbol{H}_i^\top\right)\boldsymbol{H}_i. \tag{18}$$

Thus, we have the required derivatives to evaluate $\frac{\partial C}{\partial \boldsymbol{A}_i^{(\ell_j)}}$.

HNCPD EXPANSION

We now provide brief justification of the expansion of the NCPD in terms of later factorizations used in Section 2.1; that is,

$$[\![\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2, \cdots, \widetilde{\boldsymbol{X}}_k]\!] = \sum_{1\le j_1,j_2,...j_k\le r^{(0)}} \alpha_{j_1,j_2,...j_k}\left((\boldsymbol{A}_1^{(0)})_{:,j_1} \otimes (\boldsymbol{A}_2^{(0)})_{:,j_2} \otimes \ldots \otimes (\boldsymbol{A}_k^{(0)})_{:,j_k}\right)$$

where $\alpha_{j_1,j_2,...j_k} = \sum_{p=1}^r (\boldsymbol{S}_1^{(0)})_{j_1,p}(\boldsymbol{S}_2^{(0)})_{j_2,p}\ldots(\boldsymbol{S}_k^{(0)})_{j_k,p}$.

We have that by definition,

$$[\![\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2, \cdots, \widetilde{\boldsymbol{X}}_k]\!] = \sum_{p=1}^r \left((\widetilde{\boldsymbol{X}}_1)_{:,p} \otimes (\widetilde{\boldsymbol{X}}_2)_{:,p} \otimes \ldots \otimes (\widetilde{\boldsymbol{X}}_k)_{:,p}\right).$$

We also have that $\widetilde{\boldsymbol{X}}_i = \boldsymbol{A}_i^{(0)}\boldsymbol{S}_i^{(0)}$ for $1 \le i \le k$, so we have that for each column $p, 1 \le p \le r$ of $\widetilde{\boldsymbol{X}}_i$,

$$(\widetilde{\boldsymbol{X}}_i)_{:,p} = \sum_{j=1}^{r^{(0)}} (\boldsymbol{S}_i^{(0)})_{j,p}(\boldsymbol{A}_i^{(0)})_{:,j}.$$

Thus, by the linearity of the outer product we have that

$$\left((\widetilde{\boldsymbol{X}}_1)_{:,p} \otimes (\widetilde{\boldsymbol{X}}_2)_{:,p} \otimes \ldots (\widetilde{\boldsymbol{X}}_k)_{:,p}\right) = \sum_{1\le j_1,j_2,...j_k\le r^{(0)}} \alpha_{p,j_1,j_2,...j_k}\left((\boldsymbol{A}_1^{(0)})_{:,j_1} \otimes (\boldsymbol{A}_2^{(0)})_{:,j_2} \otimes \ldots \otimes (\boldsymbol{A}_k^{(0)})_{:,j_k}\right)$$

where $\alpha_{p,j_1,j_2,...j_k} = (\boldsymbol{S}_1^{(0)})_{j_1,p}(\boldsymbol{S}_2^{(0)})_{j_2,p}\ldots(\boldsymbol{S}_k^{(0)})_{j_k,p}$. Now, by noting that

$$\alpha_{j_1,j_2,...,j_k} = \sum_{p=1}^r \alpha_{p,j_1,j_2,...,j_k}$$

we arrive at the original statement.

SYNTHETIC EXPERIMENT

In this section, we provide the additional views of the synthetic tensor and computed approximations from Section 3.1. In Figure 2 in the main text, for visualization we displayed the projection of each tensor onto the third mode. In Figure 9, we display the projections of these tensors onto all three modes. We see that due to the simple block structure used to produce the synthetic data tensor, the three modes all tell a similar story; that is, Neural NCPD is able to recover meaningful structure along all three modes.

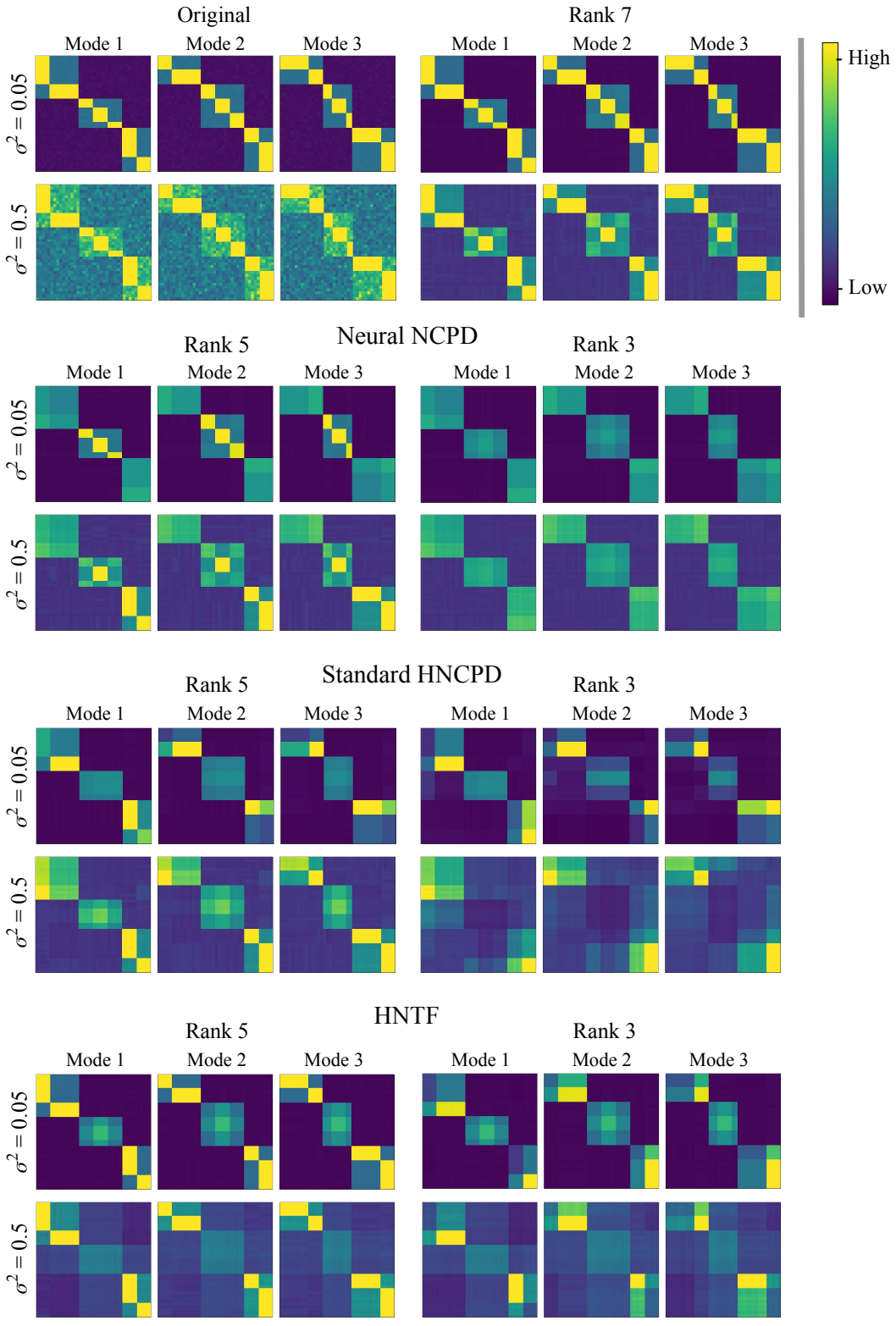
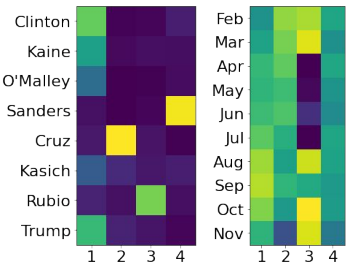


Figure 9: Here we display the projections onto all three modes for the original data tensor $\mathbf{X}$ and approximations of $\mathbf{X}$ at ranks $r = 7$, $r^{(0)} = 5$, and $r^{(1)} = 3$ produced by Neural NCPD, Standard HNCPD, and HNTF at two levels of noise.

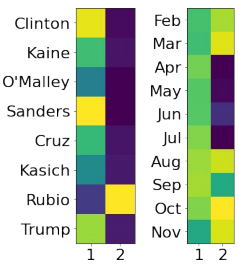
TEMPORAL DOCUMENT ANALYSIS EXPERIMENT

HNTF-1

Topic 1	Topic 2
hillary	tedcruz
trump	cruz
realdonaldtrump	ted
crooked	internet
hillaryclinton	choosecruz
Topic 3	**Topic 4**
marcorubio	berniesanders
teammarco	people
vote	bernie
flsen	must
click	change

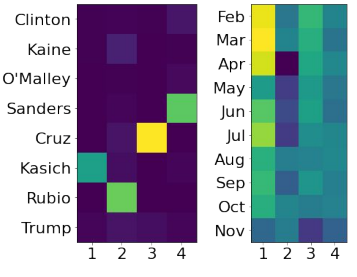


Topic 1
trump
hillary
tedcruz
people
berniesanders
Topic 2
marcorubio
teammarco
vote
flsen
click

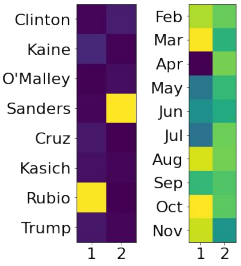


HNTF-2

Topic 1	Topic 2
johnkasich	marcorubio
kasich	teammarco
ohio	vote
john	florida
gov	flsen
Topic 3	**Topic 4**
tedcruz	berniesanders
cruz	people
ted	bernie
internet	must
choosecruz	change

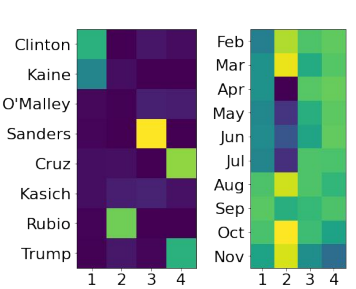


Topic 1
marcorubio
teammarco
vote
florida
flsen
Topic 2
berniesanders
people
bernie
must
change



HNTF-3

Topic 1	Topic 2
trump	marcorubio
hillary	teammarco
donald	vote
president	flsen
timkaine	click
Topic 3	**Topic 4**
berniesanders	tedcruz
people	cruz
bernie	ted
must	crooked
country	thank



Topic 1
tedcruz
cruz
ted
hillary
crooked
Topic 2
berniesanders
people
bernie
must
vote

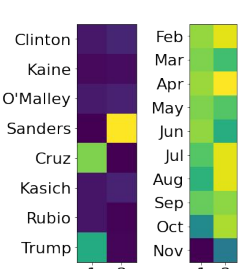


Figure 10: Here we display a three-layer HNTF on the Twitter dataset from Section 3.2 at ranks $r = 8$, $r^{(0)} = 4$, and $r^{(1)} = 2$, run separately for each of the possible ordering on the data tensor. We display the top keywords and heatmaps of topics in the candidate and temporal modes at ranks 4 (left) and 2 (right). We note that the rank 8 factorization is identical to that of Neural NCPD, so we do not re-display it here (see Section 3.2).

In Figure 10, we display the results from running HNTF on the Twitter dataset in Section 3.2, excluding the topics at rank 8 because they are identical to those learned by Neural NCPD (see Section 3.2). We see that while the factorization for the first possible ordering is similar to that of Neural NCPD and contains significant meaningful topic modeling information, the other two orderings lose significant information by the last layer and, and have topic presence and from only 2 or 3 of the eight candidates.

VIDEO DATA EXPERIMENT

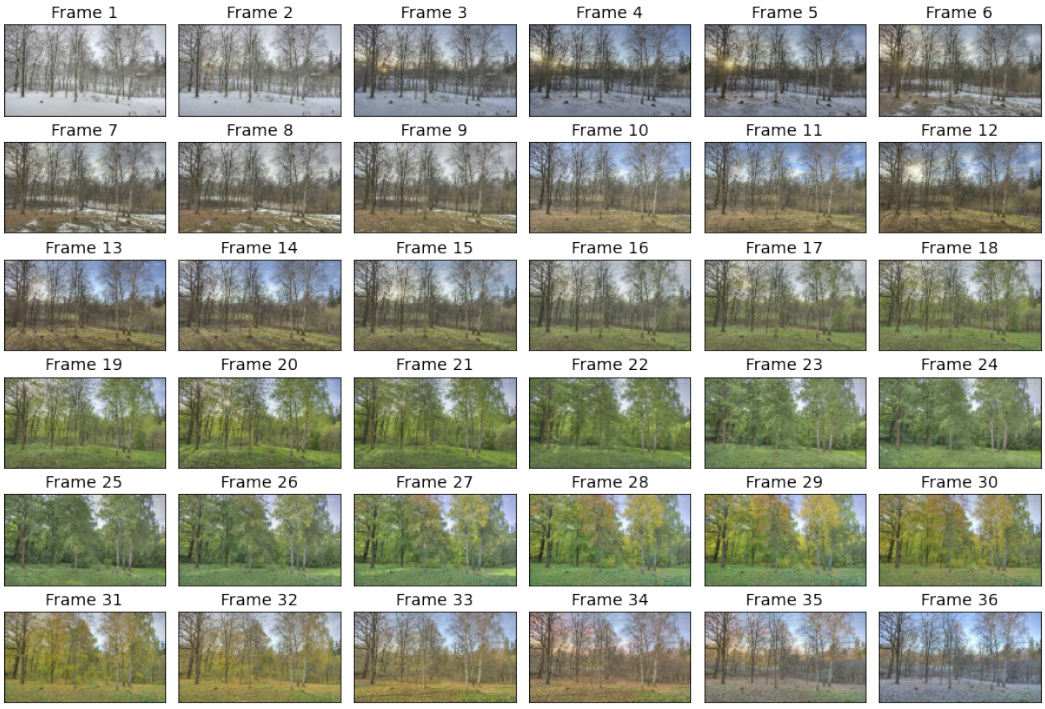


Figure 11: Here we display the first 36 of 37 frames of the time lapse video dataset from Section 3.3 (The 37th frames is included in Figure 6

In Figure 11, we display the first 36 of 37 frames of the time lapse video dataset from Section 3.3 (the 37th frame is included in Figure 6) in order to make it clear how seasons progress throughout the frames. We see that the video begins in the white winter months, transitions to spring at around frame 16, and stays green until it transitions to fall around frame 28.

In Figure 12, we display the $\boldsymbol{S}_3^{(0)}$ matrix (top) and $\boldsymbol{S}_3^{(1)}$ matrix (bottom) produced by Neural NCPD on the time-lapse video tensor described in Section 3.3. By examining the $\boldsymbol{S}$ matrices from our Neural NCPD algorithm, we are also able to see the hierarchical relationship between the topics from different ranks. In the $\boldsymbol{S}_3^{(0)}$ matrix, we see the hierarchical relationship between the rank 6 and rank 8 topics. In the $\boldsymbol{S}_3^{(1)}$ matrix, we see the hierarchical relationship between the rank 3 and rank 8 topics. We note that the $\boldsymbol{S}_3^{(0)}$ matrix (top) illustrates that topic one of rank 6 NCPD is closely related to topic eight of rank 8 NCPD, and $\boldsymbol{S}_3^{(1)}$ (bottom) similarly illustrates that topic two of rank 3 NCPD is closely related to topic eight in rank 8 NCPD; these relationships are unsurprising because, as seen in Figure 7 in the main text, these topics are present temporally during winter and fall and spatially in the sky behind the trees.

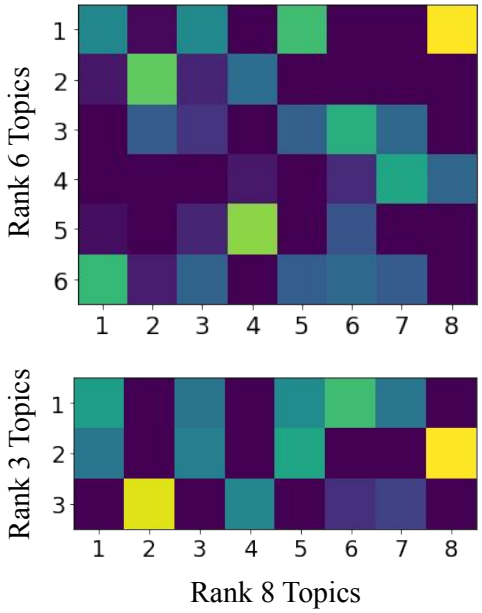


Figure 12: The $\boldsymbol{S}_3^{(0)}$ matrix (top) and $\boldsymbol{S}_3^{(1)}$ matrix (bottom) produced by Neural NCPD on the time-lapse video tensor described in Section 3.3.