Statistics in Biosciences (2022) 14:485–513
https://doi.org/10.1007/s12561-021-09331-5

Characterizing Spatiotemporal Transcriptome of the Human Brain Via Low-Rank Tensor Decomposition

Tianqi Liu[1] · **Ming Yuan**[2] · **Hongyu Zhao**[1]

Received: 16 March 2021 / Revised: 30 October 2021 / Accepted: 13 November 2021 /
Published online: 21 January 2022
© The Author(s) under exclusive licence to International Chinese Statistical Association 2021

Abstract

Spatiotemporal gene expression data of the human brain offer insights on the spatial and temporal patterns of gene regulation during brain development. Most existing methods for analyzing these data consider spatial and temporal profiles separately, with the implicit assumption that different brain regions develop in similar trajectories, and that the spatial patterns of gene expression remain similar at different time points. Although these analyses may help delineate gene regulation either spatially or temporally, they are not able to characterize heterogeneity in temporal dynamics across different brain regions, or the evolution of spatial patterns of gene regulation over time. In this article, we develop a statistical method based on low-rank tensor decomposition to more effectively analyze spatiotemporal gene expression data. We generalize the classical principal component analysis (PCA), which is applicable only to data matrices, to tensor PCA that can simultaneously capture spatial and temporal effects. We also propose an efficient algorithm that combines tensor unfolding and power iteration to estimate the tensor principal components efficiently, and provide guarantees on their statistical performance. Numerical experiments are presented to further demonstrate the merits of the proposed method. An application our method to a spatiotemporal brain expression data provides insights on gene regulation patterns in the brain.

Keywords Brain gene expression · Genomics · Principal component analysis · Tensor decomposition

✉ Tianqi Liu
tianqi.terence.liu@gmail.com

Ming Yuan
ming.yuan@columbia.edu

[1] Department of Biostatistics, Yale University, New Haven, USA

[2] Department of Statistics, Columbia University, New York, USA

1 Introduction

Principal component analysis (PCA) is among the most commonly used statistical methods for exploratory analysis of multivariate data [e.g., 10]. By seeking a low-rank approximation to the data matrix, PCA allows us to reduce the dimensionality of the data, and oftentimes serves as a useful first step to capture the essential features in the data. In particular, PCA has been widely used in analyzing gene expression data collected for multiple time points or across different biological conditions [1, 35, 38]. While PCA is appropriate to analyze data matrices, data sometimes come in the format of higher order tensors, or multilinear arrays. In particular, our work here is motivated by characterizing the spatiotemporal gene expression patterns of the human brain based on gene expression profiles collected from multiple brain regions of both developing and adult post-mortem human brains.

The human brain is a sophisticated and complex organ that contains billions of cells with different morphologies, connectivity and functions [e.g., 11]. Different brain regions have specific compositions of cell types, expressing unique combinations of genes at different developmental periods. Recent advances in sequencing and micro-dissection technology have provided us new and powerful tools to take a closer look at this complex system. Many studies have been conducted to date to collect spatiotemporal expression data to identify spatial and temporal signatures of gene regulation in the brain, and to gain insights into various biological processes of interest such as brain development processes, central nervous system formation, and brain anatomical structure shaping, among others [8, 12, 16, 24, 29, 30, 37].

The spatiotemporal expression data may be modeled by a third order multilinear array, or tensor, with one index for gene, one for region, and another one for time. Because the classical PCA can only be applied to data matrices, previous analyses of such data often consider the spatial and temporal patterns separately. To characterize temporal patterns of gene expression, data from different regions are first pooled and treated as replicates, before applying PCA. Similarly, when extracting spatial patterns of gene expression, data from different time points are combined so that PCA could be applied. Such analyses have yielded some useful insights on the gene regulation in spatiotemporal transcriptome [12, 19]. But the data pooling precludes us from understanding the heterogeneity in temporal dynamics across different regions of the brain, or the evolution of spatial gene regulation patterns over time. There is a clear demand to develop statistical methods that can more effectively utilize the tensor structure of spatiotemporal expression data.

To this end, we introduce in this article a higher order generalization, hereafter referred to as tensor PCA, of the classical PCA to better characterize spatial and temporal gene expression dynamics. As in the classical PCA, we seek the best low-rank orthogonal approximation to the data tensor. The orthogonality among the rank-one components is automatically satisfied by the classical PCA but is essential for our purpose. It not only ensures that the components can be

interpreted in the same fashion as the classical PCA, but also is necessary for the low rank approximation to be well-defined. Unlike in the case of matrices, low rank approximations to a higher order tensor without orthogonality is ill-posed and the best approximation may not even exist [e.g., 4]. However, even with orthogonality, low rank approximations to a higher order tensor is still in general NP hard to compute [e.g., 9]. Heuristic or approximation algorithms are often adopted, and they often lead to suboptimal statistical performances [e.g., 25]. It is an active area of research in recent years to achieve a balance between computational and statistical efficiency when dealing with higher order tensor. For our purposes, we propose an efficient algorithm that combines tensor unfolding and power iteration to compute the principal components under the tensor PCA framework. We also show that our estimates are not only easy to compute but also attain the optimal rate of convergence under suitable conditions.

Numerical experiments further demonstrate the merits of our proposed method. We also applied the method to the spatiotemporal expression data from [12]. We found that the proposed tensor PCA approach can effectively reduce the dimensionality of the data while preserving inherent structure among the genes. In particular, through clustering analysis, we show that tensor PCA reveals interesting relationships between gene functions and the spatiotemporal dynamics of gene regulation. To fix ideas, we focus on spatiotemporal expression data in this paper. Our methodology, however, is also readily applicable to other settings where data are in the form of tensor.

The rest of the article is organized as follows. Section 2 introduces the proposed tensor PCA methodology. Section 3 reports the result from simulation studies. Section 4 presents an application of the proposed methodology to a spatiotemporal brain gene expression data set. Finally, we conclude with some remarks and discussions by Sect. 5. All proofs are covered in supplementary materials.

2 Methodology

Denote by x_{gst} an appropriately normalized and transformed expression measurement for gene g, in region s, at time t, where $g = 1, \ldots, d_G$, $s = 1, \ldots, d_S$, and $t = 1, \ldots, d_T$ and d_G, d_S and d_T are the number of genes, regions, and time points, respectively. In many applications, we may also have replicate measurements so that x_{gst} is a vector rather than a scalar. To fix ideas, we shall focus on the case where there is no replicate. In practice, we can average over replicate measurements to convert x_{gst} from a vector to scalar in practice if necessary. Treatment of the more general situation is analogous albeit more cumbersome in notation.

2.1 From Classical PCA to Tensor PCA

As mentioned above, the classical PCA is often applied to estimate spatial and temporal patterns of gene regulation separately. Consider, for example, inferring the spatial patterns of gene regulation. Let

$$\bar{x}_{gs\cdot} = \frac{1}{d_T}\sum_{t=1}^{d_T} x_{gst},$$

be the averaged expression measurements for gene g in region s. The classical PCA then extracts the leading principal components, or equivalently the leading eigenvectors of $d_G \times d_S$ matrix $\boldsymbol{x}_g := (\bar{x}_{g1\cdot}, \ldots, \bar{x}_{gd_S\cdot})^\top$. The principal components can also be interpreted through singular value decomposition of data matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d_G})^\top$. Denote by $\boldsymbol{v}_k := (v_{k1}, \ldots, v_{kd_S})^\top$ the kth leading principal component and $\boldsymbol{u}_k := (u_{k1}, \ldots, u_{kd_G})^\top$ its normalized loadings, that is its ℓ_2 norm $\|\boldsymbol{u}\| = 1$. Then, after appropriate centering, the observed expression measurements can be written as

$$\bar{x}_{gs\cdot} = \sqrt{d_G}\sum_{k=1}^{r} \lambda_k u_{kg} v_{ks} + \bar{\epsilon}_{gs}, \tag{1}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ so that $\sqrt{d_G}\lambda_k$ is the kth largest singular value of the data matrix $(\bar{x}_{gs\cdot})_{1\leq g\leq d_G, 1\leq s\leq d_S}$, and the idiosyncratic noise $\bar{\epsilon}_{gs}$ are iid centered normal random variables. Note that, in (1), the scaling factor $\sqrt{d_G}$ is in place to ensure that λ_k^2 (more precisely $\lambda_k^2 + \mathrm{var}(\bar{\epsilon}_{gs})$) can also be understood as the kth largest eigenvalue of the covariance matrix of $(\bar{x}_{gs\cdot})_{1\leq s\leq d_S}$ when they are viewed as independent random vectors for $g = 1, \ldots, d_G$.

Obviously, because of pooling measurements from different time points, the principal components extracted this way can only be identified with spatial patterns *averaged* over all time points. Therefore it is not able to capture spatial patterns that evolve over time. Similar problem also arises when we pool data from different regions and extract principal components for temporal patterns. In order to model the spatial and temporal dynamics jointly, we now consider a generalization of PCA to specifically account for the tensor structure of the expression data.

The expression data $\boldsymbol{X} = (x_{gst})_{1\leq g\leq d_G, 1\leq s\leq d_S, 1\leq t\leq d_T}$ can be conveniently viewed as a third order tensor of dimension $d_G \times d_S \times d_T$. It is clear that the pooled data matrix

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d_G})^\top = \boldsymbol{X} \times_3 \left(\frac{1}{d_T}\mathbf{1}_{d_T}\right),$$

where $\mathbf{1}_d$ is a d dimensional vector of ones, and $\times_j$ between a tensor and vector stands for multiplication along its jth index, that is,

$$(\boldsymbol{A} \times_3 \boldsymbol{x})_{ij} = \sum_k A_{ijk} x_k.$$

See, e.g., [14] for further discussions on tensor algebra. Instead of seeking a low-rank approximation to the pooled data matrix, we shall work directly with the data tensor $\boldsymbol{X}$. More specifically, with slight abuse of notation, we shall consider the following low rank approximation to $\boldsymbol{X}$:

$$\boldsymbol{X} = \sqrt{d_G}\sum_{k=1}^{r} \lambda_k \left(\boldsymbol{u}_k \otimes \boldsymbol{v}_k \otimes \boldsymbol{w}_k\right) + \boldsymbol{E}, \tag{2}$$

where the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0$, $\boldsymbol{u}_k$s, $\boldsymbol{v}_k$s and $\boldsymbol{w}_k$s are orthonormal basis in $\mathbb{R}^{d_G}$, $\mathbb{R}^{d_S}$ and $\mathbb{R}^{d_T}$ respectively, and the $\boldsymbol{E} = (e_{gst})$ is the residual tensor consisting of independent idiosyncratic noise following a normal distribution $N(0, \sigma^2)$. Here $\otimes$ stands for the outer product so that

$$x_{gst} = \sqrt{d_G} \sum_{k=1}^{r} \lambda_k u_{kg} v_{ks} w_{kt} + e_{gst}, \qquad 1 \leq g \leq d_G, 1 \leq s \leq d_S, 1 \leq t \leq d_T.$$

Conceptually, model (2) can be viewed as a natural multiway generalization of the model for the classical PCA. Similar to the classical PCA, such a tensor decomposition allows us to conveniently capture the spatial dynamics and temporal dynamics by $\boldsymbol{v}_k$s and $\boldsymbol{w}_k$s, respectively. The loading of each gene for a particular interaction of spatial and temporal dynamics is then represented by $\boldsymbol{u}_k$s.

2.2 Estimation for Tensor PCA

Clearly, any interpretation of the data based on the tensor PCA model (2) depends on our ability to estimate the principal components $\boldsymbol{v}_k$s and $\boldsymbol{w}_k$s from the expression data $\boldsymbol{X}$. Naturally, we can consider estimating them via maximum likelihood, leading to the problem of computing the best rank r approximation to data tensor $\boldsymbol{X}$. In the case of the usual PCA, such a task can be accomplished by applying SVD to the data matrix. But for the tensor PCA model, this is a more delicate issue because low rank approximation to a generic tensor could be hard to compute at least in the worst case. To address this challenge, we introduce here an approach that combines tensor unfolding and power iteration and show that we can estimate the tenor principal components in an efficient way, both computationally and statistically.

2.2.1 Tensor Unfolding

A commonly used heuristic to overcome this problem is through tensor unfolding. In particular, in our case, we may collapse the second and third indices of $\boldsymbol{X}$ to unfold into a $d_G \times (d_S \cdot d_T)$ matrix $\mathcal{M}(\boldsymbol{X})$ by collapsing the second and third indices, that is,

$$[\mathcal{M}(\boldsymbol{X})]_{i,(j-1)d_T+k} = X_{ijk}, \qquad 1 \leq i \leq d_G, 1 \leq j \leq d_S, 1 \leq k \leq d_T.$$

It is clear that

$$\mathcal{M}(\boldsymbol{X}) = \sqrt{d_G} \sum_{k=1}^{r} \lambda_k \boldsymbol{u}_k \otimes \mathrm{vec}(\boldsymbol{v}_k \otimes \boldsymbol{w}_k) + \mathcal{M}(\boldsymbol{E}),$$

where $\mathrm{vec}(\cdot)$ vectorizes a matrix into a vector of appropriate dimension. This suggests that $\{\mathrm{vec}(\boldsymbol{v}_k \otimes \boldsymbol{w}_k) : 1 \leq k \leq r\}$ are the top right singular vectors of $\mathbb{E}[\mathcal{M}(\boldsymbol{X})]$ and can therefore be estimated by applying singular value decomposition to $\mathcal{M}(\boldsymbol{X})$. Denote by $\sqrt{d_G}\widehat{\lambda}_k$ the kth leading singular value of $\mathcal{M}(\boldsymbol{X})$, and $\widehat{\boldsymbol{h}}_k$ its corresponding right singular vector. We can reshape $\widehat{\boldsymbol{h}}_k$ into a $d_S \times d_T$ matrix $\mathrm{vec}^{-1}(\widehat{\boldsymbol{h}}_k)$, that is

$$[\mathrm{vec}^{-1}(\widehat{\boldsymbol{h}}_k)]_{ij} = (\widehat{\boldsymbol{h}}_k)_{(i-1)d_T+j}, \qquad \forall 1 \le i \le d_S, 1 \le j \le d_T.$$

An estimate of $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ can then be obtained by the leading left and right singular vectors, denoted by $\widehat{\boldsymbol{v}}_k$ and $\widehat{\boldsymbol{w}}_k$ respectively, of $\mathrm{vec}^{-1}(\widehat{\boldsymbol{h}}_k)$. It turns out that this simple approach can yield a consistent estimate of λ_ks, $\boldsymbol{v}_k$s and $\boldsymbol{w}_k$s. More specifically, we have

Theorem 1 *There exists an absolute constant $C > 0$ such that for any simple eigenvalue λ_k ($1 \le k \le r$) under the tensor PCA model (2), if the eigen-gap*

$$g_k := \min\left\{\lambda_{k-1}^2 - \lambda_k^2, \lambda_k^2 - \lambda_{k+1}^2\right\} \ge C\left(\sigma^2 + \sigma\lambda_1\right)\left(d_S d_T/d_G\right)^{1/2},$$

with the convention that $\lambda_0 = \infty$ and $\lambda_{r+1} = 0$, then

$$\max\left\{\widehat{\lambda}_k^2 - \lambda_k^2, 1 - |\langle\widehat{\boldsymbol{v}}_k, \boldsymbol{v}_k\rangle|, 1 - |\langle\widehat{\boldsymbol{w}}_k, \boldsymbol{w}_k\rangle|\right\} \le C\left(\sigma^2 + \sigma\lambda_1\right)g_k^{-1}(d_S d_T/d_G)^{1/2},$$

with probability tending to one as $d_G \to \infty$.

Theorem 1 indicates that the eigenvalue λ_k and its associated eigenvectors $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ can be estimated consistently whenever the eigen-gap

$$g_k \gg \sigma^2(d_S d_T/d_G)^{1/2}.$$

In the context of spatiotemporal expression data, the number of genes d_G is typically much larger than $d_S d_T$. Therefore, even if the eigen-gap is constant, the spatial and temporal PCA can still be consistently estimated.

2.2.2 Power Iteration

Although Theorem 1 suggests that the eigenvalue and eigenvector estimates obtained via our tensor folding scheme is consistent under fairly general conditions, they can actually be further improved. We can indeed use them as the initial value for power iteration or altering least squares to yield estimates that converge to the truth at faster rates.

Power iteration is perhaps the most commonly used algorithm for tensor decomposation [39]. We assume the standard deviation of noise is known and denoted as σ. In practice, when σ is unknown, one can estimate it by the sample variance of the residual tensor with the initial estimate. Specifically, let $\boldsymbol{b}^{[0]}$ and $\boldsymbol{c}^{[0]}$ be initial values for $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$. Let $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ be the estimates of $\boldsymbol{u}_k$, $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$, respectively. Then at the mth ($m \ge 1$) iteration, we update $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ as follows:

- Let $\boldsymbol{a}^{[m]} = \boldsymbol{a}/\|\boldsymbol{a}\|$ where

$$\boldsymbol{a} = \boldsymbol{X} \times_2 \boldsymbol{b}^{[m-1]} \times_3 \boldsymbol{c}^{[m-1]}; \tag{3}$$

- Let $\boldsymbol{b}^{[m]} = \boldsymbol{b}/\|\boldsymbol{b}\|$ where

$$\boldsymbol{b} = \boldsymbol{X} \times_1 \boldsymbol{a}^{[m]} \times_3 \boldsymbol{c}^{[m-1]} - \sigma^2 \boldsymbol{b}^{[m-1]}; \tag{4}$$

- Let $\boldsymbol{c}^{[m]} = \boldsymbol{c}/\|\boldsymbol{c}\|$ where

$$\boldsymbol{c} = \boldsymbol{X} \times_1 \boldsymbol{a}^{[m]} \times_2 \boldsymbol{b}^{[m]} - \sigma^2 \boldsymbol{c}^{[m-1]}. \tag{5}$$

The following theorem shows that the algorithm, after a certain number of iterations, yields estimates of the tensor principal components at an optimal convergence rate.

Theorem 2 *Let $\boldsymbol{b}^{[m]}$ and $\boldsymbol{c}^{[m]}$ be the estimates of $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ from the mth modified power iteration with initial values $\boldsymbol{b}^{[0]} = \widehat{\boldsymbol{v}}_k$ and $\boldsymbol{c}^{[0]} = \widehat{\boldsymbol{w}}_k$ obtained by tensor unfolding as described before. Suppose that the conditions of Theorem 1 hold. Then there exist absolute constants $C_1, C_2 > 0$ such that if*

$$\lambda_k^2 g_k \geq C_1(\sigma^2 + \lambda_1\sigma)\lambda_1^2\sqrt{\frac{d_S d_T}{d_G}},$$

then for any

$$m \geq -C_2 \log\left(\lambda_k^{-2}(\sigma^2 + \lambda_1\sigma)\sqrt{\frac{d_S + d_T}{d_G}}\right),$$

we have

$$\max\left\{1 - |\langle \boldsymbol{b}^{[m]}, \boldsymbol{v}_k\rangle|, 1 - |\langle \boldsymbol{c}^{[m]}, \boldsymbol{w}_k\rangle|\right\} = O_p\left(\lambda_k^{-2}(\sigma^2 + \lambda_1\sigma)\sqrt{\frac{d_S + d_T}{d_G}}\right), \quad \text{as } d_G \to \infty.$$

Note that we only require that the number of genes d_G diverges in Theorem 2, which is the most relevant setting in spatiotemporal expression data. If the singular values $\lambda_1, \ldots, \lambda_r$ are simple and finite, as typically the case in practice, then Theorem 2 indicates that the spatial and temporal PCAs can be estimated at the rate of convergence $\sqrt{(d_S + d_T)/d_G}$. This is to be compared with the unfolding estimates which converge at the rate of $\sqrt{d_S d_T/d_G}$.

It is also worth noting, assuming that λ_ks and σ are finite, the rate of convergence given by Theorem 2 is optimal in the following sense. Suppose that $\boldsymbol{v}_k$ is known in advance, it is not hard to see that $\boldsymbol{X} \times_2 \boldsymbol{v}_k$ is a sufficient statistics for $\boldsymbol{w}_k$. Because $\boldsymbol{w}_k$ is the usual principal component of $\boldsymbol{X} \times_2 \boldsymbol{v}_k$, following classical theory for principal components [see, e.g., 26], we know that the optimal rate of convergence for estimating $\boldsymbol{w}_k$ is of the order $\sqrt{d_T/d_G}$. Similarly, even if $\boldsymbol{w}_k$ is known apriori, the optimal rate of convergence for estimating $\boldsymbol{v}_k$ would be of the order $\sqrt{d_S/d_G}$. Obviously, not knowing either $\boldsymbol{v}_k$ or $\boldsymbol{w}_k$ only makes their estimation more difficult. Therefore, the rate of convergence established in Theorem 2 is the best attainable.

A key difference between the power iteration described above and the usual ones is that subtract $\sigma^2\boldsymbol{b}^{[m-1]}$ and $\sigma^2\boldsymbol{c}^{[m-1]}$ when updating $\boldsymbol{b}$ and $\boldsymbol{c}$ at each iteration. This modification is motivated by a careful examination of the effect of noise $\boldsymbol{E}$ on the

power iteration. Although not essential for the performance of the final estimate, this adjustment allows for faster convergence of the power iterations. A careful inspection of the proof of Theorem 2 suggests that the results continue to hold in this case because of the consistency of the initial value.

Our approach is developed for effectively modeling brain spatiotemporal gene expression data, which is a 3-order tensor. For tensor with higher orders, we can naturally generalize our algorithm. Assume the tensor is $\boldsymbol{X} = \lambda \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} \otimes \boldsymbol{d} + \boldsymbol{E}$. For the tensor unfolding part, we can recursively apply the singular value decomposition on the unfolded tensor with first dimension fixed. For example, we can first estimate $\boldsymbol{a}$ and $\boldsymbol{b} \otimes \boldsymbol{c} \otimes \boldsymbol{d}$ by applying SVD on unfolded tensor with the last three dimensions flattened. Then we can estimate the $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{d}$ according to our tensor unfolding algorithm. For power iteration, we can first add $\times_4 \boldsymbol{d}^{[m-1]}$ on the first term of Eqs. 3, 4, and 5. The we follow the Eq. 5 to update $\boldsymbol{d}$:

$$\boldsymbol{d} = \boldsymbol{X} \times_1 \boldsymbol{a}^{[m]} \times_2 \boldsymbol{b}^{[m]} \times_3 \boldsymbol{c}^{[m]} - \sigma^2 \boldsymbol{d}^{[m-1]}. \tag{6}$$

We can do the same approach for even higher order tensors. The theoretical and numerical study of the algorithm on higher order tensors are beyond the scope of this paper.

3 Numerical Experiments

To demonstrate the merits of the tensor PCA method described in the previous section, we conducted several sets of simulations.

3.1 Convergence of Power Iteration

To gain further insights into the operating characteristics of the power iteration, we examine how the estimation error changes from iteration to iteration for 50 typical

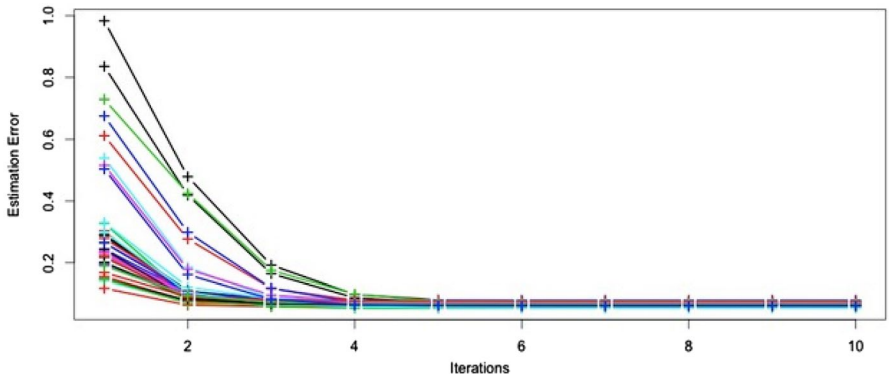

Fig. 1 Estimation error as a function of iterations for 50 typical simulated datasets with $\lambda = 4$ and $d = 200$

simulation runs with $\lambda = 4$ and $d = 200$ in Fig. 1. First, it is evident to see the estimation error reduces quickly with the iterations. It is also worth noting that the algorithm converges in only several iterations. This has great practical implication as computation is often a significant issue when dealing with tensor data.

3.2 Principal Components Estimation Accuracy

3.2.1 Synthetic Data Generation

We begin with a simple simulation setup designed to investigate the effect of dimensionality and signal strength on the estimation of tensor accuracy. In particular, we simulated data tensor from the following rank-one tensor PCA model:

$$\boldsymbol{X} = \sqrt{d}\lambda \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} + \boldsymbol{E}. \tag{7}$$

To assess the effect of dimensionality, we consider cubic tensors of dimension $\mathbb{R}^{d\times d\times d}$ where $d = 25, 50, 100$. We set $\lambda = 4$. The principal components $\boldsymbol{v}$ and $\boldsymbol{w}$, as well as the loadings $\boldsymbol{u}$ were uniformly sampled from the unit sphere in $\mathbb{R}^d$. We recall that a uniform sample from the unit sphere in $\mathbb{R}^d$ can be obtained by $Z/\|Z\|$ where $Z \sim N(0, I_d)$. The noise tensor $\boldsymbol{E}$ is a Gaussian ensemble whose entries are independent standard normal variables.

3.2.2 Baseline Approaches and Metrics

For each simulated data tensor $\boldsymbol{X}$, we compared our proposed approach (TPCA) with the following baseline approaches:

- *Tensor unfolding (UFD)* The baseline approach is described in Sect. 2.2.1. This baseline is to study the effect of power iteration.
- *Power iteration (PI1, PI5, or PI10)* We conduct power iteration (described in Sect. 2.2.2) from random initial state. We repeat the power iteration with different starting states 1, 5, or 10 times and denote them as PI1, PI5, or PI10, respectively. This is to study the efficiency of using tensor unfolding as initial state.

We use $2 \cdot \max\{1 - |\langle \hat{\boldsymbol{v}}, \boldsymbol{v}\rangle|, 1 - |\langle \hat{\boldsymbol{w}}, \boldsymbol{w}\rangle|\}$ as the estimation error, which is equivalent to $\max(\|\hat{\boldsymbol{v}} - \boldsymbol{v}\|^2, \|\hat{\boldsymbol{w}} - \boldsymbol{w}\|^2)$.

3.2.3 Results

For each simulation setting, we repeat it for 200 times and report the metrics in Table 1. Compared with UFD, it is evident from the comparison that TPCA improves the quality of estimates, especially for situations with high dimensionality. These observations are in agreement with the theoretical analysis presented in Theorems 1 and 2. Compared with (PI1, PI5, PI10), TPCA achieves the best performance. As the dimension goes higher, it requires more repetitions in power iteration from random states. During the simulation, we chose the smallest error among

Table 1 Principal components estimation errors comparison for rank 1 tensor

d	UFD	PI1	PI5	PI10	TPCA
25	0.083 (0.021)	0.293 (0.591)	**0.076 (0.017)**	**0.076 (0.017)**	**0.076 (0.017)**
50	0.092 (0.017)	0.760 (0.877)	0.082 (0.12)	**0.073 (0.012)**	**0.073 (0.012)**
100	0.133 (0.088)	1.160 (0.894)	0.299 (0.564)	0.096 (0.193)	**0.072 (0.009)**

Approaches are tensor unfolding (UFD), power iteration with 5 repetitions (PI5), power iteration with 10 repetitions (PI10), our proposed approach (TPCA). We report means and standard deviations (in parenthesis) averaged over 200 simulation runs

Bold values indicate the best metric among all methods

repetitions, which is infeasible in real applications since we don't know the ground truth. In all other cases, TPCA significantly improves upon the power iteration from random. PI1 performs worse than UFD, which suggests that pure power iteration cannot yield good results even compared to tensor unfolding. The improvement is least significant in the easiest case with $d = 25$ when PI5 estimate already appears to be quite accurate.

3.3 Synthetic Gene Expression Data

Our development was motivated by the analysis of spatiotemporal expression data. To better assess the performance of our method in such a context, we now consider a simulation setting designed to mimic it. More specifically, we simulated a spatiotemporal gene expression data tensor with $d_G = 2000$ genes, $d_S = 10$ spatial regions, $d_T = 13$ temporal regions. We assume the following tensor PCA model of rank three:

$$\boldsymbol{X} = \sqrt{d_G}\lambda \sum_{k=1}^{3} \frac{4-k}{3} \cdot \boldsymbol{u}_k \otimes \boldsymbol{v}_k \otimes \boldsymbol{w}_k + \boldsymbol{E},$$

where we fix $\sigma = 1$ and $\lambda = 3$. The eigenvectors $\boldsymbol{u} \in \mathbb{R}^{d_G}$, $\boldsymbol{v} \in \mathbb{R}^{d_S}$ and $\boldsymbol{w} \in \mathbb{R}^{d_T}$ were uniformly sampled from the Grassmannian of conformable dimensions. This simulation setting allows us to appreciate the effect of eigengap and eigenvalue, as well as the unequal dimensions on the accuracy of our estimates.

Usually, spatial-temporal gene expression data are heterogeneous. It could be the case that the variance differs across genes, locations, and time periods. To study the effect of heterogeneity along dimension d_G, we apply linear increase of standard deviation as $\sigma_i = i/d_G$ for $i = 1, \ldots, d_G$. Similarly, we can apply on the heterogeneous noise on spatial and temporal dimension.

3.3.1 Principal Components Estimation Accuracy

We compare the proposed tensor PCA approach with the classical PCA approach for estimating each of the flattened spatiotemporal principal component. We add homogeneous noise and heterogeneous noise across gene, spatial, temporal dimensions

Table 2 Principal components estimation errors comparison for synthetic rank 3 noisy spatiotemporal gene expression tensor with different noise types

Noise type	Principal component	PCA	TPCA
Homogeneous	PC1	0.090 (0.006)	**0.001 (0.000)**
Homogeneous	PC2	0.141 (0.009)	**0.001 (0.000)**
Homogeneous	PC3	0.351 (0.025)	**0.008 (0.003)**
Gene-wise heterogeneous	PC1	0.051 (0.004)	**0.000 (0.000)**
Gene-wise heterogeneous	PC2	0.079 (0.005)	**0.000 (0.000)**
Gene-wise heterogeneous	PC3	0.187 (0.012)	**0.002 (0.001)**
Spatial heterogeneous	PC1	0.065 (0.006)	**0.001 (0.000)**
Spatial heterogeneous	PC2	0.117 (0.011)	**0.002 (0.001)**
Spatial heterogeneous	PC3	0.448 (0.092)	**0.021 (0.010)**
Temporal heterogeneous	PC1	0.065 (0.005)	**0.001 (0.000)**
Temporal heterogeneous	PC2	0.117 (0.012)	**0.002 (0.001)**
Temporal heterogeneous	PC3	0.433 (0.074)	**0.017 (0.008)**

The data tensor is of 2000 genes by 10 spatial regions by 13 time periods. We report means and standard deviations (in parenthesis) averaged over 200 simulation runs

Bold values indicate the best metric among all methods

for each simulation. For each principal component, we use ($\|x - \boldsymbol{v} \otimes \boldsymbol{w}\|$ as metrics, where $x = \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{w}}$ for TPCA and $x =$ "right singular vector" for PCA. The results reported in Table 2 confirm our theoretical findings and suggests the superior performance of the proposed approach over the classical PCA agnostic to the heterogeneity of the noise. It is worth noting that gene wise heterogeneous noise has smaller effect on the principal component estimation, while spatial and temporal heterogeneity can make the estimation more challenging.

3.3.2 Signal Tensor Estimation Accuracy

We compared TPCA with classical PCA, tensor unfolding (UFD), Higher Order Orthogonal Iteration of Tensors (HOOI) [3] on signal tensor estimation. HOOI is a specific orthogonal Tucker decomposition algorithm that generalizes the matrix singular value decomposition. It is an iterative approach that computes the singular values for each mode fixing others. See Sheehan and Saad [33] for details of the algorithm. We simulated the data in the same way as described early in the section. For signal tensor $\boldsymbol{T}$ and estimated signal tensor $\hat{\boldsymbol{T}}$, we compute the relative error as $\|\hat{\boldsymbol{T}} - \boldsymbol{T}\|_F / \|\boldsymbol{T}\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm of a tensor. The relative errors are reported in Table 3. TPCA again shows the best performance among all approaches.

3.4 Clustering Based on Tensor PCA

Oftentimes in practice, PCA is not the final goal of data analysis. It is commonly used as an initial step to reduce the dimensionality before further analysis.

Table 3 Signal tensor estimation comparisons among the classical PCA, Higher Order Orthogonal Iteration of Tensors, tensor unfolding, and tensor PCA, in terms of relative errors averaged over 200 simulation runs

Noise type	PCA	HOOI	UFD	TPCA
Homogeneous	0.496 (0.005)	0.469 (0.005)	0.469 (0.005)	**0.467 (0.005)**
Gene-wise heterogeneous	0.283 (0.003)	**0.270 (0.003)**	0.271 (0.003)	**0.270 (0.003)**
Spatial heterogeneous	0.357 (0.022)	0.350 (0.039)	0.331 (0.022)	**0.329 (0.022)**
Temporal heterogeneous	0.354 (0.020)	0.351 (0.040)	0.329 (0.020)	**0.327 (0.021)**

Numbers in parentheses are the standard deviations

Bold values indicate the best metric among all methods

For example, PCA based clustering is often performed when dealing with gene expression data. See, e.g., [38]. Similarly, our tensor PCA can serve the same purpose. To investigate the utility of our approach in this capacity, we conducted a set of simulation studies where for each simulated dataset, we first estimated the loadings $\boldsymbol{u}_k$s and then applied clustering to the loadings. To fix ideas, we adopted the popular k-means technique for clustering although other alternatives could also be employed.

Motivated by the dataset from [12], which we shall discuss in further details in the next section, we simulated a data tensor of size $\mathbb{R}^{1087\times 10\times 13}$ from the following model:

$$\boldsymbol{X} = \sum_{k=1}^{3} \lambda_k \boldsymbol{u}_k \otimes \boldsymbol{v}_k \otimes \boldsymbol{w}_k + \sigma^2 \boldsymbol{E}. \tag{8}$$

where $\lambda_1 = 337.8$, $\lambda_2 = 27.1$, $\lambda_3 = 9.0$, and $\sigma = 0.2$. These values, along with the principal components $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ are based on estimates when fitting a tensor PCA model to the data from [12]. The clusters, induced by the loadings $\boldsymbol{u}_k$, were generated as follows. For a given number K of clusters, we first generated the cluster centroids $C \in \mathbb{R}^{K\times 3}$ from right singular vector matrix of K by 3 Gaussian random matrix. We then assigned clusters among 1087 observations and generated the observed tensor with $\sigma = 1, 5, 10, 20$, representing different levels of signal-to-noise ratio.

For comparison purposes, we also considered using the classical PCA based approach to reduce the dimensionality. For each method, we took the loadings

Table 4 Clustering performance comparison between the classical PCA and tensor PCA, in terms of Adjusted Rand index averaged over 200 simulation runs

Noise	PCA	TPCA
20	0.525 (0.219)	**0.930 (0.106)**
10	0.619 (0.254)	**0.963 (0.074)**
5	0.718 (0.296)	**0.965 (0.080)**
1	0.942 (0.124)	**1.000 (0.000)**

Numbers in parentheses are the standard deviations

Bold values indicate the best metric among all methods

from the first four directions and then applied k-means to infer the cluster membership. We used the adjusted Rand Index as a means of measuring the clustering quality. The results for each method and a variety of combinations of dimension, averaged over 200 runs, are reported in Table 4. The results suggest that tensor PCA based clustering is superior to that based on the classical PCA.

4 Application to Human Brain Expression Data

We now turn to the spatiotemporal expression data from Kang et al. [12] that we alluded to earlier.

4.1 Dataset Description and Preprocessing

4.1.1 Dataset Description

[12] reported the generation and analysis of exon-level transcriptome and associated genotyping data from multiple brain regions and neocortical areas of developing and adult post-mortem human brains. The dataset was also analyzed by Liu et al. [22] on selecting ultrahigh dimensional feature and Lin et al. [21] on modeling spatial temporal pattern with Markov Random Field. It consists of spatiotemporal gene expression data of post mortem human brains with each from a time period with all neocortex regions. It has 11 areas and 15 time periods. The areas include orbitofrontal cortex (OFC), dorsolateral prefrontal cortex (DFC), ventral frontal cortex (VFC), primary motor cortex (M1C), primary somatosensory cortex (S1C), posterior inferior parietal cortex (IPC), primary auditory (A1) cortex (A1C), superior temporal cortex (STC), medial prefrontal cortex (MFC), inferior temporal cortex (ITC), and primary visual cortex (V1C). The time periods span from embryonic (period 1) to late adulthood (period 15), we refer readers to Table 5 for details. We ignore the first two time periods (period 1 and 2) and one neocortex region (V1C) due to the high variations. For one time period with more than one brains, we aggregate over samples for each time and region combination. We refer the readers to those papers for more dataset description.

4.1.2 Dataset Preprocessing

Following [8], we selected genes with reproducible spatial patterns across individuals according to their correlations between samples, leading to a total of 1087 genes. To reduce individual variations, we first take average across subjects for each (gene, location, time period). Then we get a data tensor of size $d_G = 1087$, $d_S = 10$ and $d_T = 13$.

Before applying the tensor PCA, we first centered the gene expression measurements by subtracting the mean expression level for each gene because we are primarily interested in the spatial and temporal dynamics of the expression levels. To

Table 5 Periods of human development and adulthood as defined by [12]: *M* postnatal months; *PCW* post-conceptional weeks; *Y* postnatal years

Period	Description	Age
1	Embryonic	4PCW≤Age<8PCW
2	Early fetal	8PCW≤Age<10PCW
3	Early fetal	10PCW≤Age<13PCW
4	Early mid-fetal	13PCW≤Age<16PCW
5	Early mid-fetal	16PCW≤Age<19PCW
6	Late mid-fetal	19PCW≤Age<24PCW
7	Late fetal	24PCW≤Age<38PCW
8	Neonatal and early infancy	0M (birth)≤Age<6M
9	Late infancy	6M≤Age<12M
10	Early childhood	1Y≤Age<6Y
11	Middle and late childhood	6Y≤Age<12Y
12	Adolescence	12Y≤Age<20Y
13	Young adulthood	20Y≤Age<40Y
14	Middle adulthood	40Y≤Age<60Y
15	Late adulthood	60Y≤Age

remove the mean level, however it is more subtle than the classical PCA, we want to remove both mean spatial effect and mean temporal effect. More specifically, we applied tensor PCA to $\tilde{\boldsymbol{X}} \in \mathbb{R}^{d_G \times d_T \times d_S}$ where

$$\tilde{x}_{gst} = x_{gst} - \bar{x}_{g\cdot t} - \bar{x}_{gs\cdot} + \bar{x}_{g\cdot\cdot}$$

and $\boldsymbol{X}$ is the original data tensor.

4.2 Analysis Based on Tensor PCA

4.2.1 Choose the Number of Components

We first conduct tensor decomposition by our proposed algorithm. As in the classical PCA, we can look at the scree plot to examine the contribution of each component in the tensor PCA model. We can see that the contribution from the principal components quickly tapers off (Fig. 2). We choose the top three components according to the scree plot. Notice that choosing the number of components is trickier for clustering analysis, we use the scree plot here to fix ideas. For more discussion on how to choose optimal number of components, we refer readers to Yeung and Ruzzo [38].

4.2.2 Biological Interpretations of the Spatial and Temporal Factors

To gain insights, the top three spatial and temporal principal components are given in Fig. 3. And the top three spatial factors are mapped to brain neocortex regions in Fig. 4, where the color represents value, the darker the higher. L1 to L8 denote the different physical slice coordinates of brains. The first factor increases from L1 and

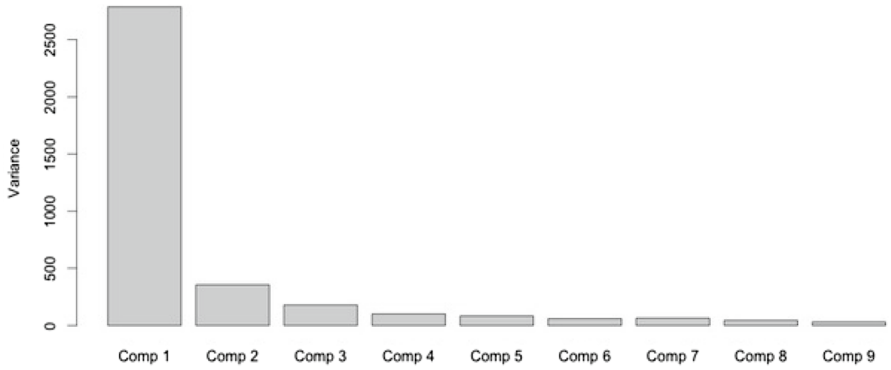

Fig. 2 Scree plot of the tensor PCA for the dataset from [12]

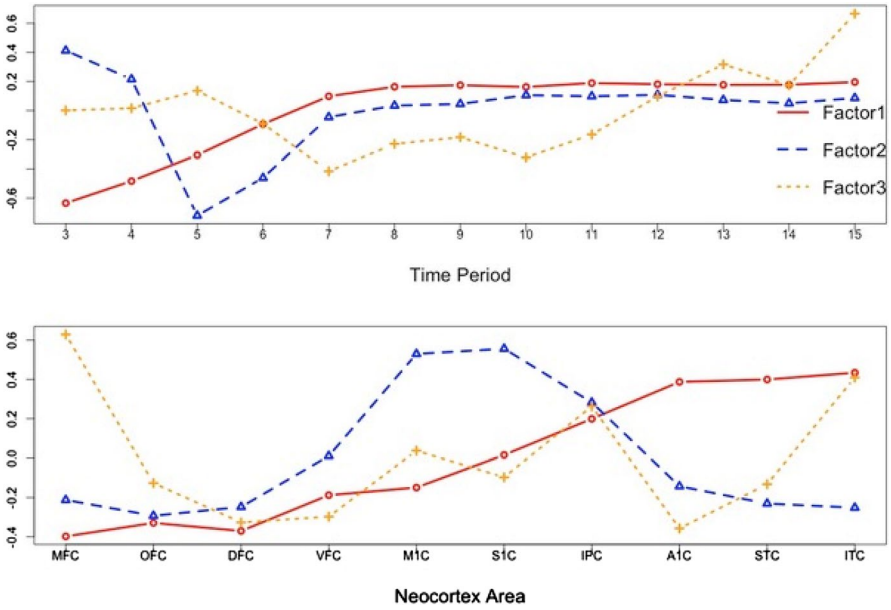

Fig. 3 Temporal and spatial factors of tensor PCA for the dataset from [12]

L2 to L4 and L5. The second factor achieves maximum at M1C and S1C and decays over distance from the above two regions. The third factor shows strong signals in MFC and ITC.

To better understand these three factors, we conducted gene set enrichment analysis based on Gene Ontology (http://geneontology.org) for each factor. We calculated the relative weight of factor i for each gene by $|u_i|/\sum_{j=1}^{3}|u_j|$, where $u \in \mathbb{R}^3$ is one row of gene factors. For each factor, we chose the top 15% quantile genes to form the gene sets. The results are presented in Table 6. Factor 1 relates with anatomical structure development, and this result is consistent with its spatial gradient pattern

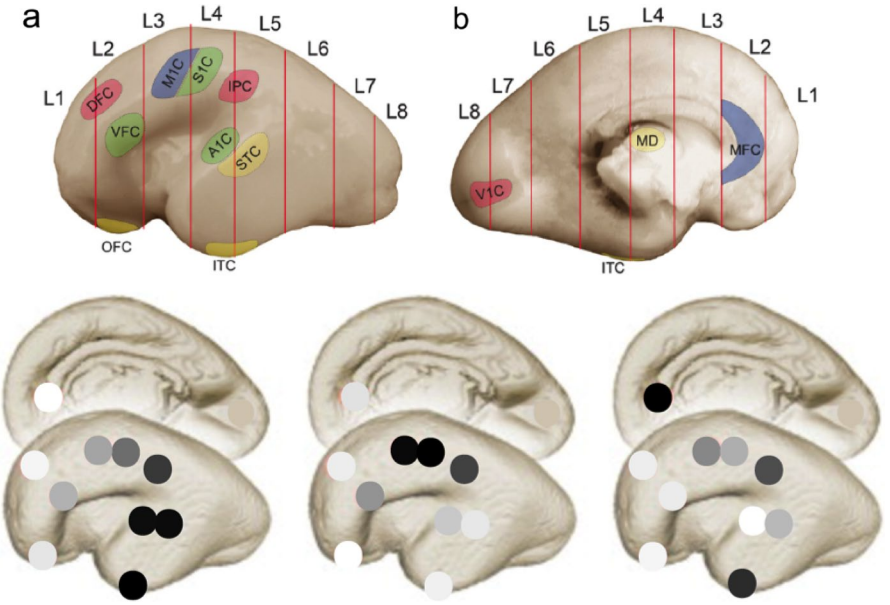

Fig. 4 Spatial factors on locations of neocortex

Table 6 Gene enrichment analysis results on factors

Factor	Enriched term	*P*-value with Bonferroni correction
1	Anatomical structure development	4.65E−04
	Developmental process	2.93E−03
2	Nervous system development	4.20E−04
	Sensory organ development	1.09E−03
	Positive regulation of signal transduction	1.36E−02
	Generation of neurons	1.98E−02
3	Chemical synaptic transmission	3.23E−06
	Multicellular organismal response to stress	7.32E−04
	Nucleic acid metabolic process	9.09E−04
	Ion transmembrane transport	8.02E−04
	Innervation	1.62E−02
	Startle response	2.79E−02

and decrease in magnitude of temporal pattern. Factor 2 has enriched term in sensory organ development, and this agrees with its huge magnitude in S1C. Besides, regulation of anatomical structure morphogenesis term supports the smooth spatial pattern from S1C and M1C to MFC and ITC. Factor 3 is enriched in innervation

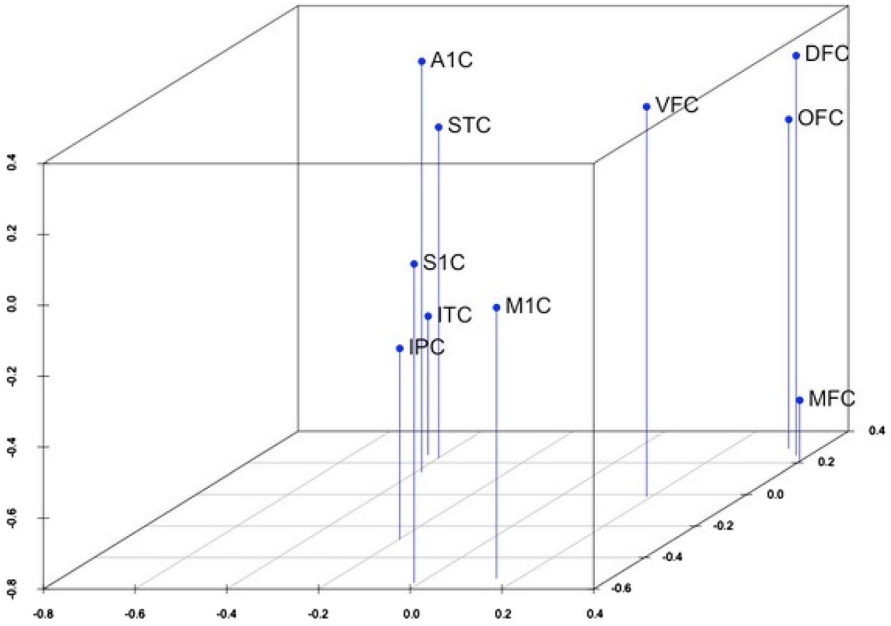

Fig. 5 Loadings on the top three spatial factors for each of the ten neocortex regions

related with aging [2, 18], startle response associated with ITC [32], and chemical synaptic transmission related with aging [23].

To further examine the meaning of the spatial factors, we use the three spatial factors as the coordinates for each of the 10 locations in a 3D plot as shown in Fig. 5. Remarkably the spatial patterns of these locations are fairly consistent with the physical locations of these neocortex regions in the brain.

It is interesting to note, from the temporal trajectories, that the first two factors show clear signs of prenatal development (until Period 7) while the third factor exhibits increasing influence from young childhood (from Period 11). Factor 1 shows a spatial gradient effect that expression level tapers off from ITC to MFC or the other way. Remarkably, the same effect was reported in [24], which is explained by intrinsic signaling controlled partly by graded expression of transcription factors. Some representative genes such as FGFR3 and CBLN2 were found to preserve in both human and mouse neocortex. Taking temporal effect into consideration, factor 1 indicates that the gradient effect diminishes from early fetal (Period 3) to late fetal (Period 7), and almost vanishes after early infancy. Same effects were observed in [30] that areal transcriptional become more synchronized during postnatal development. Factor 2 suggests the importance of prenatal development of M1C and S1C. Both areas are well represented in the second factor while essentially absent from the other factors. This observation based on our analysis seems to agree with recent findings in neuroscience that activation patterns of extremely preterm infants' primary somatosensory cortex area are predictive of future development outcome. See, e.g., [27]. Factor 3 distinguishes middle adulthood (Period 14) and late adulthood

(Period 15) with different value in ITC and MFC comparing other 8 regions. This effect was reported in [30] that MFC and ITC have much higher number of neocortical interareal differentially expressed (DEX) genes. In term of aging, declining metabolism in MFC correlates with declining cognitive function [5–7, 28], and shrinkage of ITC increases with age [31]. When we consider 3 factors together, we can validate the temporal hourglass pattern observed in [30] that huge number of DEX genes exist before infancy (Period 8), and areal differences almost vanish from infancy to adulthood (Period 14) and reappear in late adulthood (Period 15).

4.2.3 Clustering Analysis

Finally, we used the factors estimated based on our tensor PCA model as the basis for clustering. In particular, we applied k-means clustering with $k = 5$ clusters to the three dimensional factor loadings. The resulting cluster sizes are 156, 167, 332, 280, and 152, respectively. Gene set enrichment analysis based on Gene ontology was performed for each group with the results presented in Table 7.

These results show a clear separation among different functional groups. This further indicates that the spatiotemporal pattern of a gene informs its functionality. Moreover, enriched terms such as anatomical structure development, forebrain development are highly associated with the spatial areas of neocortex, which again suggests the meaningfulness of the tensor principal components.

5 Conclusions

In this paper, we have introduced a generalization of the classical PCA that can be applied to data in the form of tensors. We also proposed efficient algorithms to estimate the principal components using a novel combination of power iteration and tensor unfolding. Both theoretical analysis and numerical experiments point to the efficacy of our method. Although the methodology is generally applicable to other applications, our development was motivated by the analysis of spatiotemporal expression data which in recent years have become a common place in studying brain development among other biological processes. An application of our method to one such example further demonstrates its potential usefulness.

6 Software

Software in the form of R package with complete documentation. It is available at https://github.com/TerenceLiu4444/tensorpca.

Table 7 Gene enrichment analysis results. We apply TPCA algorithm to reduce the dimensionality of spatiotemporal pattern to 3d for each gene

Cluster	Enriched term	*P*-value after Bonferroni correction
1	Nervous system development	8.58E−11
	Anatomical structure development	3.43E−09
	Neurogenesis	1.63E−05
	Regulation of developmental process	3.12E−05
	Cell communication	9.93E−05
2	Chemical synaptic transmission	8.38E−08
	Inorganic ion transmembrane transport	2.98E−04
	Nucleic acid metabolic process	6.57E−04
	Regulation of postsynaptic membrane potential	8.45E−04
	Multicellular organismal response to stress	1.24E−02
3	Single-organism process	1.81E−10
	Regulation of localization	9.92E−04
	Single organism signaling	1.06E−03
	Response to stimulus	1.59E−03
	Regulation of multicellular organismal process	6.18E−03
4	Single-organism process	4.13E−06
	Anatomical structure development	2.72E−04
	Nervous system development	4.05E−04
	Signal transduction	4.84E−02
5	Single-organism developmental process	6.18E−05
	Forebrain development	4.77E−03
	Chemical synaptic transmission	1.18E−03
	Neuron projection morphogenesis	9.42E−03
	Axon development	9.65E−03
	Regulation of neuron differentiation	3.23E−02
	Regulation of smooth muscle cell migration	3.88E−02

Then we apply kmean clustering algorithm to cluster genes into 5 clusters according their spatiotemporal patterns. We conduct gene enrichment analysis for each cluster of genes and identify the most salient biological process associated with each cluster

Appendix: Proofs

Proof (Proof of Theorem 1) Write

$$\boldsymbol{T} = \sqrt{d_G} \sum_{k=1}^{r} \lambda_k \left(\boldsymbol{u}_k \otimes \boldsymbol{v}_k \otimes \boldsymbol{w}_k \right).$$

Then $\boldsymbol{X} = \boldsymbol{T} + \boldsymbol{E}$. Denote by

$$X_g = (x_{gst})_{1 \le s \le d_S, 1 \le t \le d_T}.$$

Let T_g, E_g be similarly defined. Then

$$\begin{aligned}\frac{1}{d_G}\mathcal{M}(\boldsymbol{X})^\top\mathcal{M}(\boldsymbol{X}) =& \frac{1}{d_G}\sum_{g=1}^{d_G}\mathrm{vec}(X_g)\otimes\mathrm{vec}(X_g)\\ =&\mathcal{M}\left(\frac{1}{d_G}\sum_{g=1}^{d_G}X_g\otimes X_g\right)\\ =&\mathcal{M}\left(\frac{1}{d_G}\sum_{g=1}^{d_G}T_g\otimes T_g+\frac{1}{d_G}\sum_{g=1}^{d_G}E_g\otimes E_g+\frac{1}{d_G}\sum_{g=1}^{d_G}\left(T_g\otimes E_g+E_g\otimes T_g\right)\right).\end{aligned}$$

Hereafter, with slight abuse of notation, we use $\mathcal{M}$ to denote the matricization operator that collapses the first two, and remaining two indices of a fourth order tensor respectively. Observe that

$$T_g=\sqrt{d_G}\sum_{k=1}^{r}\lambda_k u_{kg}\left(\boldsymbol{v}_k\otimes\boldsymbol{w}_k\right).$$

Therefore

$$T_g\otimes T_g=d_G\sum_{k_1,k_2=1}^{r}\lambda_{k_1}\lambda_{k_2}u_{k_1g}u_{k_2g}\left(\boldsymbol{v}_{k_1}\otimes\boldsymbol{w}_{k_1}\otimes\boldsymbol{v}_{k_2}\otimes\boldsymbol{w}_{k_2}\right).$$

Because of the orthogonality among $\boldsymbol{u}_k$s, we get

$$\frac{1}{d_G}\sum_{g=1}^{d_G}T_g\otimes T_g=\sum_{k=1}^{r}\lambda_k^2\left((\boldsymbol{v}_k\otimes\boldsymbol{w}_k)\otimes(\boldsymbol{v}_k\otimes\boldsymbol{w}_k)\right).$$

On the other hand, note that

$$\mathcal{M}\left(\frac{1}{d_G}\sum_{g=1}^{d_G}E_g\otimes E_g\right)=\frac{1}{d_G}\sum_{g=1}^{d_G}\left(\mathrm{vec}(E_g)\otimes\mathrm{vec}(E_g)\right).$$

In other words, $\mathcal{M}(d_G^{-1}\sum_{g=1}^{d_G}E_g\otimes E_g)$ is the sample covariance matrix of independent Gaussian vectors

$$\mathrm{vec}(E_g)\sim N(0,I_{d_S\cdot d_T}),\qquad 1\le g\le d_G.$$

Therefore, there exists an absolute constant $C_1>0$ such that

$$\left\|\mathcal{M}\left(\frac{1}{d_G}\sum_{g=1}^{d_G}E_g\otimes E_g\right)-I_{d_S\cdot d_T}\right\|\le C_1\sigma^2\sqrt{\frac{d_Sd_T}{d_G}}.$$

with probability tending to one as $d_G\to\infty$. See, e.g., [34].

Finally, observe that

$$\sum_{g=1}^{d_G} T_g \otimes E_g = \sqrt{d_G} \sum_{k=1}^{r} \lambda_k \left[\boldsymbol{v}_k \otimes \boldsymbol{w}_k \otimes \left(\sum_{g=1}^{d_G} u_{kg} E_g \right) \right] =: \sqrt{d_G} \sum_{k=1}^{r} \lambda_k \big(\boldsymbol{v}_k \otimes \boldsymbol{w}_k \otimes Z_k \big).$$

By the orthogonality of $\boldsymbol{u}_k$s, it is not hard to see that Z_ks are independent Gaussian matrices:

$$\mathrm{vec}(Z_k) \sim N\Big(0, \sigma^2 I_{d_S \cdot d_T}\Big),$$

so that there exists an absolute constant $C_2 > 0$ such that

$$\left\| \mathcal{M}\left(\frac{1}{d_G} \sum_{g=1}^{d_G} \big(T_g \otimes E_g + E_g \otimes T_g \big) \right) \right\| \le \frac{2}{d_G} \left\| \mathcal{M}\left(\sum_{g=1}^{d_G} T_g \otimes E_g \right) \right\| \le C_2 \lambda_1 \sigma \sqrt{\frac{d_S d_T}{d_G}},$$

with probability tending to one.

To sum up, we get

$$\left\| \frac{1}{d_G} \mathcal{M}(\boldsymbol{X})^\top \mathcal{M}(\boldsymbol{X}) - A \right\| \le (C_1 \sigma^2 + C_2 \lambda_1 \sigma) \sqrt{\frac{d_S d_T}{d_G}}.$$

where

$$A = I_{d_S \cdot d_T} + \sum_{k=1}^{r} \lambda_k^2 \big[\mathrm{vec}\big(\boldsymbol{v}_k \otimes \boldsymbol{w}_k \big) \otimes \mathrm{vec}\big(\boldsymbol{v}_k \otimes \boldsymbol{w}_k \big) \big].$$

It is clear that

$$\Big\{ (1 + \lambda_k^2, \mathrm{vec}(\boldsymbol{v}_k \otimes \boldsymbol{w}_k)) \,:\, 1 \le k \le r \Big\}$$

are the leading eigenvalue-eigenvector pairs of A.

Recall that $(\widehat{\lambda}_k^2, \widehat{\boldsymbol{h}}_k)$ is the kth eigenvalue-eigenvector pair of $\mathcal{M}(\boldsymbol{X})^\top \mathcal{M}(\boldsymbol{X})$. By Lidskii's inequality,

$$|\widehat{\lambda}_k^2 - \lambda_k^2| \le (C_1 \sigma^2 + C_2 \lambda_1 \sigma) \sqrt{\frac{d_S d_T}{d_G}}.$$

See, e.g., [13, 20]. Then

$$\begin{aligned} \|\mathrm{vec}^{-1}(\widehat{\boldsymbol{h}}_k) - \boldsymbol{v}_k \otimes \boldsymbol{w}_k\|^2 \le & \|\mathrm{vec}^{-1}(\widehat{\boldsymbol{h}}_k) - \boldsymbol{v}_k \otimes \boldsymbol{w}_k\|_{\mathrm{F}}^2 \\ = & 2 - 2\langle \widehat{\boldsymbol{h}}_k, \mathrm{vec}(\boldsymbol{v}_k \otimes \boldsymbol{w}_k) \rangle \\ \le & 2 \left\| \widehat{\boldsymbol{h}}_k \otimes \widehat{\boldsymbol{h}}_k - \mathrm{vec}(\boldsymbol{v}_k \otimes \boldsymbol{w}_k) \otimes \mathrm{vec}(\boldsymbol{v}_k \otimes \boldsymbol{w}_k) \right\| \\ \le & 8 (C_1 \sigma^2 + C_2 \lambda_1 \sigma) g_k^{-1} \sqrt{\frac{d_S d_T}{d_G}}, \end{aligned}$$

where the last inequality follows from Lemma 1 from [15]. For large enough C, we can ensure that

$$\|\mathrm{vec}^{-1}(\widehat{\boldsymbol{h}}_k) - \boldsymbol{v}_k \otimes \boldsymbol{w}_k\|^2 \le \frac{C}{4}(\sigma^2 + \lambda_1\sigma)g_k^{-1}\sqrt{\frac{d_S d_T}{d_G}} \le \frac{1}{4}.$$

Recall also that $\widehat{\boldsymbol{v}}_k$ and $\widehat{\boldsymbol{w}}_k$ be the leading singular vectors of $\mathrm{vec}^{-1}(\widehat{\boldsymbol{h}}_k)$. By Wedin's perturbation theorem, we obtain immediately that

$$\max\left\{1 - |\langle\widehat{\boldsymbol{v}}_k, \boldsymbol{v}_k\rangle|, 1 - |\langle\widehat{\boldsymbol{w}}_k, \boldsymbol{w}_k\rangle|\right\} \le C(\sigma^2 + \lambda_1\sigma)\sigma^2 g_k^{-1}\sqrt{\frac{d_S d_T}{d_G}}.$$

See, e.g., [25, 36]. □

Proof (Proof of Theorem 2) Denote by

$$\tilde{\boldsymbol{b}} = \left(\frac{1}{d_G}\sum_{g=1}^{d_G} X_g \otimes X_g\right) \times_2 \boldsymbol{c}^{[m-1]} \times_3 \boldsymbol{c}^{[m-1]} \times_4 \boldsymbol{b}^{[m-1]} - \sigma^2 \boldsymbol{b}^{[m-1]}.$$

It is not hard to see that

$$\boldsymbol{b}^{[m]} = \tilde{\boldsymbol{b}}/\|\tilde{\boldsymbol{b}}\|.$$

Let $\mathcal{M}^{-1}$ be the inverse of the matricization operator $\mathcal{M}$ that unfold a fourth order tensor into matrices, that is, $\mathcal{M}^{-1}$ reshapes a $(d_S d_T) \times (d_S d_T)$ matrix into a fourth order tensor of size $d_S \times d_T \times d_S \times d_T$. Observe that

$$\begin{aligned}
\frac{1}{d_G}\sum_{g=1}^{d_G} X_g \otimes X_g =& \frac{1}{d_G}\sum_{g=1}^{d_G} T_g \otimes T_g + \frac{1}{d_G}\sum_{g=1}^{d_G} E_g \otimes E_g + \frac{1}{d_G}\sum_{g=1}^{d_G}\left(T_g \otimes E_g + E_g \otimes T_g\right) \\
=& \lambda_k^2\left((\boldsymbol{v}_k \otimes \boldsymbol{w}_k) \otimes (\boldsymbol{v}_k \otimes \boldsymbol{w}_k)\right) + \sum_{j\neq k}\lambda_j^2\left((\boldsymbol{v}_j \otimes \boldsymbol{w}_j) \otimes (\boldsymbol{v}_j \otimes \boldsymbol{w}_j)\right) \\
&+ \sigma^2\mathcal{M}^{-1}(I_{d_S\cdot d_T}) + \left(\frac{1}{d_G}\sum_{g=1}^{d_G} E_g \otimes E_g - \mathcal{M}^{-1}(I_{d_S\cdot d_T})\right) \\
&+ \frac{1}{d_G}\sum_{g=1}^{d_G}\left(T_g \otimes E_g + E_g \otimes T_g\right) \\
=:& \lambda_k^2\left((\boldsymbol{v}_k \otimes \boldsymbol{w}_k) \otimes (\boldsymbol{v}_k \otimes \boldsymbol{w}_k)\right) + \Delta_1 + \sigma^2\mathcal{M}^{-1}(I_{d_S\cdot d_T}) + \Delta_2 + \Delta_3.
\end{aligned}$$

We get

$$\tilde{\boldsymbol{b}} = \lambda_k^2\langle\boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k\rangle\langle\boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k\rangle^2\boldsymbol{v}_k + (\Delta_1 + \Delta_2 + \Delta_3) \times_2 \boldsymbol{c}^{[m-1]} \times_3 \boldsymbol{c}^{[m-1]} \times_4 \boldsymbol{b}^{[m-1]},$$

where we used the fact that

$$\mathcal{M}^{-1}(I_{d_S \cdot d_T}) \times_2 \boldsymbol{c}^{[m-1]} \times_3 \boldsymbol{c}^{[m-1]} \times_4 \boldsymbol{b}^{[m-1]} = \boldsymbol{b}^{[m-1]}.$$

Therefore

$$\begin{aligned}
|\langle \tilde{\boldsymbol{b}}, \boldsymbol{v}_k \rangle| =& \left| \lambda_k^2 \langle \boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k \rangle \langle \boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k \rangle^2 + \langle \Delta_1 + \Delta_2 + \Delta_3, \boldsymbol{v}_k \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{b}^{[m-1]} \rangle \right| \\
=& \lambda_k^2 |\langle \boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k \rangle| \langle \boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k \rangle^2 + \left| \langle \Delta_2 + \Delta_3, \boldsymbol{v}_k \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{b}^{[m-1]} \rangle \right| \\
\geq& \lambda_k^2 |\langle \boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k \rangle| \langle \boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k \rangle^2 - \|\Delta_2 + \Delta_3\|.
\end{aligned}$$

Denote by

$$\tau_m = \min \left\{ |\langle \boldsymbol{b}^{[m]}, \boldsymbol{v}_k \rangle|, |\langle \boldsymbol{c}^{[m]}, \boldsymbol{w}_k \rangle| \right\}.$$

Then,

$$\left| \langle \tilde{\boldsymbol{b}}, \boldsymbol{v}_k \rangle \right| \geq \lambda_k^2 \tau_{m-1}^3 - \|\Delta_2 + \Delta_3\|.$$

On the other hand, note that

$$\begin{aligned}
\|\tilde{\boldsymbol{b}}\| = \langle \tilde{\boldsymbol{b}}, \boldsymbol{b}^{[m]} \rangle \leq& \lambda_k^2 \langle \boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k \rangle \langle \boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k \rangle^2 \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle \\
&+ \left\langle \Delta_1 + \Delta_2 + \Delta_3, \boldsymbol{b}^{[m]} \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{b}^{[m-1]} \right\rangle.
\end{aligned}$$

Write

$$P_{\boldsymbol{v}_k}^{\perp} = I_{d_S} - \boldsymbol{v}_k \otimes \boldsymbol{v}_k, \quad \text{and} \quad P_{\boldsymbol{w}_k}^{\perp} = (I_{d_T} - \boldsymbol{w}_k \otimes \boldsymbol{w}_k).$$

Then

$$\begin{aligned}
\|\tilde{\boldsymbol{b}}\| =& \lambda_k^2 \langle \boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k \rangle \langle \boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k \rangle^2 \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle \\
&+ \langle \Delta_1, P_{\boldsymbol{v}_k}^{\perp} \boldsymbol{b}^{[m]} \otimes P_{\boldsymbol{w}_k}^{\perp} \boldsymbol{c}^{[m-1]} \otimes P_{\boldsymbol{w}_k}^{\perp} \boldsymbol{c}^{[m-1]} \otimes P_{\boldsymbol{v}_k}^{\perp} \boldsymbol{b}^{[m-1]} \rangle \\
&+ \langle \Delta_2 + \Delta_3, \boldsymbol{b}^{[m]} \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{c}^{[m-1]} \otimes \boldsymbol{b}^{[m-1]} \rangle \\
\leq& \lambda_k^2 \langle \boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k \rangle \langle \boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k \rangle^2 \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle \\
&+ \lambda_1^2 \left(1 - \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle^2\right)^{1/2} \left(1 - \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m-1]} \rangle^2\right)^{1/2} \left(1 - \langle \boldsymbol{w}_k, \boldsymbol{c}^{[m-1]} \rangle^2\right) + \|\Delta_2 + \Delta_3\| \\
\leq& \lambda_k^2 |\langle \boldsymbol{b}^{[m-1]}, \boldsymbol{v}_k \rangle| \langle \boldsymbol{c}^{[m-1]}, \boldsymbol{w}_k \rangle^2 \\
&+ \lambda_1^2 \left(1 - \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle^2\right)^{1/2} \left(1 - \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m-1]} \rangle^2\right)^{1/2} \left(1 - \langle \boldsymbol{w}_k, \boldsymbol{c}^{[m-1]} \rangle^2\right) + \|\Delta_2 + \Delta_3\| \\
\leq& \lambda_k^2 \tau_{m-1}^3 + \lambda_1^2 \left(1 - \tau_{m-1}^2\right)^{3/2} \left(1 - \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle^2\right)^{1/2} + \|\Delta_2 + \Delta_3\|.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
|\langle \boldsymbol{b}^{[m]}, \boldsymbol{v}_k \rangle| =& |\langle \tilde{\boldsymbol{b}}, \boldsymbol{v}_k \rangle| / \|\tilde{\boldsymbol{b}}\| \\
\geq& 1 - \left(\lambda_k^2 \tau_{m-1}^3\right)^{-1} \left[\lambda_1^2 \left(1 - \tau_{m-1}^2\right)^{3/2} \left(1 - \langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle^2\right)^{1/2}\right] \\
& - \left(\lambda_k^2 \tau_{m-1}^3\right)^{-1} \|\Delta_2 + \Delta_3\| \\
\geq& 1 - 4\left(\lambda_k^2 \tau_{m-1}^3\right)^{-1} \left[\lambda_1^2 \left(1 - \tau_{m-1}\right)^{3/2} \left(1 - |\langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle|\right)^{1/2}\right] \\
& - \left(\lambda_k^2 \tau_{m-1}^3\right)^{-1} \|\Delta_2 + \Delta_3\| \\
\geq& 1 - \max\left\{ 8\left(\lambda_k^2 \tau_{m-1}^3\right)^{-1} \left[\lambda_1^2 \left(1 - \tau_{m-1}\right)^{3/2} \left(1 - |\langle \boldsymbol{v}_k, \boldsymbol{b}^{[m]} \rangle|\right)^{1/2}\right], \right. \\
& \left. 2\left(\lambda_k^2 \tau_{m-1}^3\right)^{-1} \|\Delta_2 + \Delta_3\| \right\} \\
\geq& 1 - \max\left\{ 64\left(\lambda_k^2 \tau_{m-1}^3\right)^{-2} \lambda_1^4 \left(1 - \tau_{m-1}\right)^3, 2\left(\lambda_k^2 \tau_{m-1}^3\right)^{-1} \|\Delta_2 + \Delta_3\| \right\}.
\end{aligned}
$$

Assume that

$$
\tau_{m-1} \geq \max\left\{ 1 - \frac{1}{64}\left(\frac{\lambda_k}{\lambda_1}\right)^2, \frac{1}{2} \right\}, \tag{9}
$$

which we shall verify later. Then

$$
1 - |\langle \boldsymbol{b}^{[m]}, \boldsymbol{v}_k \rangle| \leq \max\left\{ \frac{1}{2}\left(1 - \tau_{m-1}\right), 16\lambda_k^{-2} \|\Delta_2 + \Delta_3\| \right\}. \tag{10}
$$

Similarly, we can show that

$$
1 - |\langle \boldsymbol{c}^{[m]}, \boldsymbol{w}_k \rangle| \leq \max\left\{ \frac{1}{2}\left(1 - \tau_{m-1}\right), 16\lambda_k^{-2} \|\Delta_2 + \Delta_3\| \right\}.
$$

Together, they imply that

$$
1 - \tau_m \leq \max\left\{ \frac{1}{2}\left(1 - \tau_{m-1}\right), 16\lambda_k^{-2} \|\Delta_2 + \Delta_3\| \right\}. \tag{11}
$$

It is clear from (11) that if

$$
1 - \tau_{m-1} \leq 16\lambda_k^{-2} \|\Delta_2 + \Delta_3\|, \tag{12}
$$

so is $1 - \tau_m$. Thus (12) holds for any

$$
m \geq -\log_2\left(\frac{16}{1 - \tau_0} \lambda_k^{-2} \|\Delta_2 + \Delta_3\| \right).
$$

We now derive bounds for $\|\Delta_2 + \Delta_3\|$. By triangular inequality $\|\Delta_2 + \Delta_3\| \leq \|\Delta_2\| + \|\Delta_3\|$. By Lemma 1,

$$\|\Delta_2\| \le 6\sigma^2\sqrt{\frac{d_S+d_T}{d_G}}.$$

Next we consider bounding $\|\Delta_3\|$. Recall that

$$\Delta_3 = \frac{1}{d_G}\sum_{g=1}^{d_G} T_g \otimes E_g + \frac{1}{d_G}\sum_{g=1}^{d_G} E_g \otimes T_g.$$

By triangular inequality,

$$\|\Delta_3\| \le \left\|\frac{1}{d_G}\sum_{g=1}^{d_G} T_g \otimes E_g\right\| + \left\|\frac{1}{d_G}\sum_{g=1}^{d_G} E_g \otimes T_g\right\| = \frac{2}{d_G}\left\|\sum_{g=1}^{d_G} T_g \otimes E_g\right\|.$$

Note that

$$\sum_{g=1}^{d_G} T_g \otimes E_g = \sqrt{d_G}\sum_{k=1}^{r}\lambda_k\left[\boldsymbol{v}_k \otimes \boldsymbol{w}_k \otimes \left(\sum_{g=1}^{d_G} u_{kg}E_g\right)\right] =: \sqrt{d_G}\sum_{k=1}^{r}\lambda_k(\boldsymbol{v}_k \otimes \boldsymbol{w}_k \otimes Z_k),$$

where Z_ks are independent $d_S \times d_T$ Gaussian ensembles. By Lemma 2, we get

$$\left\|\sum_{g=1}^{d_G} T_g \otimes E_g\right\| = O_p\Big(\lambda_1\sigma\sqrt{d_G(d_S+d_T)}\Big), \quad \text{as } d_G \to \infty,$$

where we used the fact that $r \le \min\{d_S, d_T\}$. Therefore,

$$\|\Delta_3\| = O_p\left(\lambda_1\sigma\sqrt{\frac{d_S+d_T}{d_G}}\right).$$

Thus, (12) implies that

$$1-\tau_m = O_p\left(\lambda_k^{-2}(2\sigma^2+\lambda_1\sigma)\sqrt{\frac{d_S+d_T}{d_G}}\right), \tag{13}$$

for any large enough m.

It remains to verify condition (9), which we shall do by induction. In the light of Theorem 1 and the assumption on λ_1 and λ_k, we know that it is satisfied when $m=0$, as soon as the numerical constant $C>0$ is taken large enough. Now if τ_{m-1} satisfies (9), then (11) holds. We can then deduct that the lower bound given by (9) also holds for τ_m. □

B. Auxiliary Results

We now derive tail bounds necessary for the proof of Theorem 2.

Lemma 1 *Let $\boldsymbol{E} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ $(d_1 \geq d_2 \geq d_3)$ be a third order tensor whose entries $e_{i_1 i_2 i_3}$ $(1 \leq i_k \leq d_k)$ are independently sampled from the standard normal distribution. Write $E_i = (e_{i_1 i_2 i_3})_{1 \leq i_2 \leq d_2, 1 \leq i_3 \leq d_3}$ its ith (2, 3) slice. Then*

$$\left\| \frac{1}{d_1} \sum_{i=1}^{d_1} \left\{ E_i \otimes E_i - \mathbb{E}\big(E_i \otimes E_i\big) \right\} \right\| \leq 6 \sqrt{\frac{d_2 + d_3}{d_1}}$$

with probability tending to one as $d_1 \to \infty$.

Proof (Proof of Lemma 1) For brevity, denote by

$$\boldsymbol{T}_i = E_i \otimes E_i - \mathbb{E}\big(E_i \otimes E_i\big)$$

and

$$\boldsymbol{T} = \frac{1}{d_1} \sum_{i=1}^{d_1} \boldsymbol{T}_i.$$

Note that $\boldsymbol{T}$ is a $d_2 \times d_3 \times d_3 \times d_2$ tensor obeying

$$T(\omega) = T(\pi_{14}(\omega)) = T(\pi_{23}(\omega)), \qquad \forall \omega \in [d_2] \times [d_3] \times [d_3] \times [d_2],$$

where $\pi_{k_1 k_2}$ permutes the k_1 and k_2 entry of vector. Therefore

$$\boldsymbol{T} = \sup_{\substack{\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^{d_2}, \boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^{d_3} \\ \|\boldsymbol{a}_1\|, \|\boldsymbol{a}_2\|, \|\boldsymbol{b}_1\|, \|\boldsymbol{b}_2\| = 1}} \langle \boldsymbol{T}, \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{a}_2 \rangle = \sup_{\substack{\boldsymbol{a} \in \mathbb{R}^{d_2}, \boldsymbol{b} \in \mathbb{R}^{d_3} \\ \|\boldsymbol{a}\|, \|\boldsymbol{b}\| = 1}} \langle \boldsymbol{T}, \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{b} \otimes \boldsymbol{a} \rangle.$$

Observe that for any $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{S}^{d_2 - 1}$ and $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{S}^{d_3 - 1}$,

$$\begin{aligned}
&\big| \langle \boldsymbol{T}, \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{a}_1 \rangle - \langle \boldsymbol{T}, \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{a}_2 \rangle \big| \\
&\quad \leq \big| \langle \boldsymbol{T}, \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{a}_1 \rangle - \langle \boldsymbol{T}, \boldsymbol{a}_2 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{a}_2 \rangle \big| \\
&\qquad + \big| \langle \boldsymbol{T}, \boldsymbol{a}_2 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{a}_2 \rangle - \langle \boldsymbol{T}, \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{a}_2 \rangle \big| \\
&\quad \leq \big| \langle \boldsymbol{T}, (\boldsymbol{a}_1 - \boldsymbol{a}_2) \otimes \boldsymbol{b}_1 \otimes \boldsymbol{b}_1 \otimes (\boldsymbol{a}_1 + \boldsymbol{a}_2) \rangle \big| \\
&\qquad + \big| \langle \boldsymbol{T}, \boldsymbol{a}_2 \otimes (\boldsymbol{b}_1 - \boldsymbol{b}_2) \otimes (\boldsymbol{b}_1 + \boldsymbol{b}_2) \otimes \boldsymbol{a}_2 \rangle \big| \\
&\quad \leq 2 \|\boldsymbol{T}\| \big(\|\boldsymbol{a}_1 - \boldsymbol{a}_2\| + \|\boldsymbol{b}_1 - \boldsymbol{b}_2\| \big).
\end{aligned}$$

In particular, if $\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|, \|\boldsymbol{b}_1 - \boldsymbol{b}_2\| \leq 1/8$, then

$$\big| \langle \boldsymbol{T}, \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{a}_1 \rangle - \langle \boldsymbol{T}, \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{a}_2 \rangle \big| \leq \frac{1}{2} \|\boldsymbol{T}\|. \tag{14}$$

We can find a 1/8 cover set $\mathcal{N}_1$ of $\mathbb{S}^{d_2-1}$ such that $|\mathcal{N}_1| \le 9^{d_2}$. Similarly, let $\mathcal{N}_2$ be a 1/8 covering set of $\mathbb{S}^{d_3-1}$ such that $|\mathcal{N}_2| \le 9^{d_3}$. Then by (14)

$$\|\boldsymbol{T}\| \le \sup_{\boldsymbol{a}\in\mathcal{N}_1,\boldsymbol{b}\in\mathcal{N}_2} \langle \boldsymbol{T}, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{b}\otimes\boldsymbol{a}\rangle + \frac{1}{2}\|\boldsymbol{T}\|,$$

suggesting

$$\|\boldsymbol{T}\| \le 2 \sup_{\boldsymbol{a}\in\mathcal{N}_1,\boldsymbol{b}\in\mathcal{N}_2} \langle \boldsymbol{T}, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{b}\otimes\boldsymbol{a}\rangle.$$

Now note that for any $\boldsymbol{a} \in \mathcal{N}_1$ and $\boldsymbol{b} \in \mathcal{N}_2$,

$$\langle \boldsymbol{T}_i, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{b}\otimes\boldsymbol{a}\rangle = \langle E_i, \boldsymbol{a}\otimes\boldsymbol{b}\rangle^2 - \mathbb{E}\langle E_i, \boldsymbol{a}\otimes\boldsymbol{b}\rangle^2 = \langle E_i, \boldsymbol{a}\otimes\boldsymbol{b}\rangle^2 - 1 \sim \chi_1^2 - 1.$$

Therefore

$$\langle \boldsymbol{T}, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{b}\otimes\boldsymbol{a}\rangle \sim \frac{1}{d_1}\chi_{d_1}^2 - 1.$$

An application of the χ^2 tail bound from [17] leads to

$$\mathbb{P}\{\langle \boldsymbol{T}, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{b}\otimes\boldsymbol{a}\rangle \ge x\} \le \exp(-d_1x^2/4),$$

for any $x < 1$. By union bound,

$$\mathbb{P}\left\{\sup_{\boldsymbol{a}\in\mathcal{N}_1,\boldsymbol{b}\in\mathcal{N}_2} \langle \boldsymbol{T}, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{b}\otimes\boldsymbol{a}\rangle \ge x\right\} \le 9^{d_2+d_3}\exp(-d_1x^2/4),$$

so that

$$\|\boldsymbol{T}\| \le 6\sqrt{\frac{d_2+d_3}{d_1}}$$

with probability tending to one as $d_1 \to \infty$. □

Lemma 2 *Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{d_1}\}$ be an orthonormal basis of $\mathbb{R}^{d_1}$, and $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{d_2}\}$ an orthonormal basis of $\mathbb{R}^{d_2}$. Let $Z_1, \ldots, Z_r$ be independent $d_3 \times d_4$ Gaussian random matrix whose entries are independently drawn from the standard normal distribution. Then for any sequence of nonnegative numbers $\lambda_1, \ldots, \lambda_r \le 1$:*

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{r}\lambda_k(\boldsymbol{v}_k\otimes\boldsymbol{w}_k\otimes Z_k)\right\| \ge \sqrt{d_3}+\sqrt{d_4}+\sqrt{2\log r}+t\right\} \le \exp(-t^2/2).$$

Proof (Proof of Lemma 2) Observe that

$$\begin{aligned}\left\|\sum_{k=1}^{r}\lambda_k\left(\boldsymbol{v}_k\otimes\boldsymbol{w}_k\otimes Z_k\right)\right\| &= \sup_{\boldsymbol{a}\in\mathbb{S}^{d_1-1},\boldsymbol{b}\in\mathbb{S}^{d_2-1}}\left\|\sum_{k=1}^{r}\lambda_k\langle\boldsymbol{a},\boldsymbol{v}_k\rangle\langle\boldsymbol{b},\boldsymbol{w}_k\rangle Z_k\right\| \\ &= \sup_{\boldsymbol{a}\in\mathbb{S}^{r-1},\boldsymbol{b}\in\mathbb{S}^{r-1}}\left\|\sum_{k=1}^{r}\lambda_k a_k b_k Z_k\right\| \\ &\le \sup_{\boldsymbol{a}\in\mathbb{S}^{r-1},\boldsymbol{b}\in\mathbb{S}^{r-1}}\sum_{k=1}^{r}\lambda_k a_k b_k\|Z_k\| \\ &\le \left(\max_{1\le k\le r}\lambda_k\|Z_k\|\right)\left(\sup_{\boldsymbol{a}\in\mathbb{S}^{r-1},\boldsymbol{b}\in\mathbb{S}^{r-1}}\sum_{k=1}^{r}a_k b_k\right) \\ &\le \max_{1\le k\le r}\|Z_k\|.\end{aligned}$$

By concentration bounds for Gaussian random matrices,

$$\mathbb{P}\left\{\|Z_k\|\ge\sqrt{d_3}+\sqrt{d_4}+t\right\}\le\exp(-t^2/2).$$

See, e.g., [34]. □

References

1. Alter O, Brown P, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci 97:10101–10106
2. Coyle JT, Price DL, Delong MR (1983) Alzheimer's disease: a disorder of cortical cholinergic innervation. Science 219 (4589):1184–1190
3. De Lathauwer L, De Moor B, Vandewalle J (2000) A multilinear singular value decomposition. SIAM J Matrix Anal Appl 21 (4):1253–1278
4. de Silva V, Lim LH (2008) Tensor rank and the ill-posedness of the best low-rank approximation problem. SIAM J Matrix Anal Appl 30 (3):1084–1127
5. Donoso M, Collins AG, Koechlin E (2014) Foundations of human reasoning in the prefrontal cortex. Science 344 (6191):1481–1486
6. Fjell AM, Westlye LT, Amlien I, Espeseth T, Reinvang I, Raz N, Agartz I, Salat DH, Greve DN, Fischl B et al (2009) High consistency of regional cortical thinning in aging across multiple samples. Cereb cortex 19:2001–2012
7. Gutchess AH, Kensinger EA, Schacter DL (2007) Aging, self-referencing, and medial prefrontal cortex. Soc Neurosci 2 (2):117–133
8. Hawrylycz M, Miller JA, Menon V, Feng D, Dolbeare T, Guillozet-Bongaarts AL, Jegga AG, Aronow BJ, Lee CK, Bernard A et al (2015) Canonical genetic signatures of the adult human brain. Nat Neurosci 18 (12):1832
9. Hillar C, Lim L (2013) Most tensor problems are np-hard. J ACM 60 (6):45
10. Jolliffe I (2002) Principal component analysis. Springer, Berlin
11. Kandel ER, Schwartz JH, Jessell TM et al (2000) Principles of neural science, vol 4. McGraw-Hill, New York
12. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G et al (2011) Spatio-temporal transcriptome of the human brain. Nature 478 (7370):483–489
13. Kato T (1982) A short introduction to perturbation theory for linear operators. Springer, New York
14. Koldar TG, Bader BW (2009) Tensor decompositions and applications. SIAM Rev 51:455–500
15. Koltchinskii V, Lounici K (2014) Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. arXiv:14084643

16. Landel V, Baranger K, Virard I, Loriod B, Khrestchatisky M, Rivera S, Benech P, Féron F (2014) Temporal gene profiling of the 5xfad transgenic mouse model highlights the importance of microglial activation in Alzheimer's disease. Mol Neurodegener 9 (1):1–18
17. Laurent B, Massart P (1998) Adaptive estimation of a quadratic functional by model selection. Ann Stat 28 (5):1303–1338
18. Lauria G, Holland N, Hauer P, Cornblath DR, Griffin JW, McArthur JC (1999) Epidermal innervation: changes with aging, topographic location, and in sensory neuropathy. J Neurol Sci 164 (2):172–178
19. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ et al (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445 (7124):168–176
20. Lidskii V (1950) The proper values of the sum and product of symmetric matrices. Dokl Akad Nauk SSSR 75:769–772
21. Lin Z, Sanders SJ, Li M, Sestan N, State MW, Zhao H (2015) A markov random field-based approach to characterizing human brain development using spatial-temporal transcriptome data. Ann Appl Stat 9 (1):429
22. Liu T, Lee KY, Zhao H (2016) Ultrahigh dimensional feature selection via kernel canonical correlation analysis. arXiv:160407354
23. Luebke J, Chang YM, Moore T, Rosene D (2004) Normal aging results in decreased synaptic excitation and increased synaptic inhibition of layer 2/3 pyramidal cells in the monkey prefrontal cortex. Neuroscience 125 (1):277–288
24. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K et al (2014) Transcriptional landscape of the prenatal human brain. Nature 508 (7495):199–206
25. Montanari A, Richard E (2014) A statistical model for tensor pca. NIPS
26. Muirhead RJ (2009) Aspects of multivariate statistical theory, vol 197. Wiley, Hoboken
27. Nevalainen P, Lauronen L, Pihko E (2014) Development of human somatosensory cortical functions—what have we learned from magnetoencephalography: a review. Front Hum Neurosci 8:158
28. Pardo JV, Lee JT, Sheikh SA, Surerus-Johnson C, Shah H, Munch KR, Carlis JV, Lewis SM, Kuskowski MA, Dysken MW (2007) Where the brain grows old: decline in anterior cingulate and medial prefrontal function with normal aging. Neuroimage 35 (3):1231–1237
29. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell 155 (5):1008–1021
30. Pletikos M, Sousa AM, Sedmak G, Meyer KA, Zhu Y, Cheng F, Li M, Kawasawa YI, Šestan N (2014) Temporal specification and bilaterality of human neocortical topographic gene expression. Neuron 81 (2):321–332
31. Raz N, Lindenberger U, Rodrigue KM, Kennedy KM, Head D, Williamson A, Dahle C, Gerstorf D, Acker JD (2005) Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. Cereb Cortex 15 (11):1676–1689
32. Sabatinelli D, Bradley MM, Fitzsimmons JR, Lang PJ (2005) Parallel amygdala and inferotemporal activation reflect emotional intensity and fear relevance. Neuroimage 24 (4):1265–1270
33. Sheehan BN, Saad Y (2007) Higher order orthogonal iteration of tensors (hooi) and its relation to pca and glram. In: Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, pp 355–365
34. Vershynin R (2012) Introduction to the non-asymptotic analysis of random matrices. In: Compressed Sensing. Cambridge University Press, Cambridge pp 210–268
35. Wall M, Dyck P, Brettin T (2001) Singular value decomposition analysis of microarray data. Bioinformatics 17:566–568
36. Wedin P (1972) Perturbation bounds in connection with singular value decomposition. BIT Num Math 12 (1):99–111
37. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci 95 (1):334–339
38. Yeung KY, Ruzzo WL (2001) Principal component analysis for clustering gene expression data. Bioinformatics 17 (9):763–774
39. Zhang T, Golub GH (2001) Rank-one approximation to high order tensors. SIAM J Matrix Anal Appl 23 (2):534–550