Dimensionality Reduction of Longitudinal ‘Omics Data using Modern Tensor Factorizations.

Supplementary discussion

Uria Mor1,2,* Yotam Cohen1,* Rafael Valdés-Mas1,* Denise Kviatkovsky1

Eran Elinav1,3,♯ Haim Avron2,‡

1 Systems Immunology Department, Weizmann Institute of Science, Rehovot, 7610001, Israel.
2 School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997801, Israel.
3 Microbiome and Cancer Division, DKFZ, Heidelberg, Germany.
* Equal first contributors.
♯ Equal last contributors.
‡ Equal last contributors.

All correspondence to:
Haim Avron, Ph.D.
School of Mathematical Sciences, Tel Aviv University
Ramat Aviv,
Tel Aviv, Israel, 6997801
(03) 640-8893 (phone)
haimav@tauex.tau.ac.il
&
Eran Elinav, M.D., Ph.D.
Systems Immunology Department,
Weizmann Institute of Science,
234 Herzl Street,
Rehovot, Israel, 7610001
(08) 934-4014 (phone)
eran.elinav@weizmann.ac.il

Supporting Information

A real order-\(N\) tensor \(\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}\) is an multi-dimensional array of entries\(^1\). Each of the entries of \(\mathbf{A}\) can be referred to by specifying an \(N\)-tuple of numbers \((i_1, \ldots, i_N)\) where \(i_k \in \{1, 2, \ldots, d_k\}\).

\(^1\)In pure mathematics this is actually the definition of a hypermatrix, while a tensor is an algebraic object that describes a multilinear relationship. However, in data science often the term ‘tensor’ is abused to mean a hypermatrix, and we adopt this terminology here as well.
A third-order tensor $\mathcal{A} \in \mathbb{R}^{m \times p \times n}$ can also be viewed as an $m$ elements long list of $p \times n$ matrices, each an horizontal slice of the tensor (Fig. 1a and 1b). This mathematical construct is appealing in the context of longitudinal studies as it enables storing the data in a way that is consistent with the data collection. One might think of $\mathcal{A}$ as a data-structure for holding the results of an experiment during which $n$ samples were collected from $m$ participants, and each sample is characterized by $p$ features. These features may be genes in the case of RNA-seq samples, taxonomic composition of shotgun metagenomics sequencing, etc. The tensor data structure reflects not only the data points but also key relationships between them. For example, let $1 \leq i \leq p$ an integer denoting the index of a certain feature, then variations of this feature across the whole cohort are obtained by fixing the second coordinate of the tensor to $i$: $\mathcal{A}_{:,i,:}^2$. Similarly, tracking this feature in a single timepoint $t$ is done by restriction of the last two coordinates $\mathcal{A}_{:,i,,t}$. The arrangement of the same data in the form of a matrix, which has only two dimensions, would require us to make a somewhat arbitrary choice about which of the two dimensions are to be coalesced into a single dimension. For example, one might consider each individual subject as a single sample, and as such each individual will have a designated row in the data matrix, while repeated measurements of the same feature at several timepoints are treated as entirely different features, resulting in an $m \times np$ matrix. Note that by concatenating the repeated samples of each individual, we form a new feature space in which the temporal context is lost. Another option is to define samples as dimensional entities measured for each subject at all timepoints, resulting in an $nm \times p$ matrix. This formulation breaks the correspondence between data of the same individual across timepoints, as well as the data of all individuals at a single timepoint.

Continuing with the above example, if in addition we were to sample each subject at $k$ body sites (instead of one), then the number of possible ways of rearranging data in the form of a matrix would have increased to four different choices, each of which captures a different aspect of the experimental design. On the other hand, having the data in a format of tensor would only require an addition of a single mode (instead of one), then the number of possible ways of rearranging data in the form of a matrix would have increased to four different choices, each of which captures a different aspect of the experimental design.

The number of terms $r$ in the above factorization denotes the maximal rank of the sought approximate $\mathcal{A}_r$ of the original tensor $\mathcal{A}$. As a convention, each of the $r$ summands $(\sigma u_1^{(1)} \circ \ldots \circ u_i^{(N)})$ is a rank-1 tensor (for alternative definitions of tensor rank see [6]). This factorization provides an intuitive breakdown of the data which is somewhat analogous to that of a PCA for matrix data: the overall contribution of each component $u_i^{(j)}$ to the approximation is determined by the magnitude of the corresponding scaler $\sigma_i$ (larger scaler implies greater contribution), and the components themselves may reflect the different modalities of the data (in the above

\[ \mathcal{A} \approx \mathcal{A}_r = \sum_{i=1}^{r} \sigma_i u_i^{(1)} \circ u_i^{(2)} \circ \ldots \circ u_i^{(N)} \in \mathbb{R}^{d_1 \times \ldots \times d_N}, \] (1)

where $\sigma_i$ are positive scalers, the components $u_i^{(j)}$ are $d_j$ dimensional unit vectors and $\circ$ denotes the outer-product operation $(u_i^{(j)} \circ u_i^{(k)})$ is a $d_j \times d_k$ matrix, while $u_i^{(j)} \circ u_i^{(k)} \circ u_i^{(l)}$ is a tensor in $\mathbb{R}^{d_j \times d_k \times d_l}$. The number of terms $r$ in the above factorization denotes the maximal rank of the sought approximate $\mathcal{A}_r$ of the original tensor $\mathcal{A}$. For brevity, we denote a CP factorization by $\mathcal{A}_r = [\Sigma; \mathbf{U}_1, \ldots, \mathbf{U}_N]$ where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$, and $\mathbf{U}_j$ are $d_j \times r$ matrices of columns that are the unit vectors $\mathbf{u}_1^{(j)}, \ldots, \mathbf{u}_r^{(j)}$. As a convention, each of the $r$ summands $(\sigma u_1^{(1)} \circ \ldots \circ u_i^{(N)})$ is a rank-1 tensor (for alternative definitions of tensor rank see [6]). This factorization provides an intuitive breakdown of the data which is somewhat analogous to that of a PCA for matrix data: the overall contribution of each component $u_i^{(j)}$ to the approximation is determined by the magnitude of the corresponding scaler $\sigma_i$ (larger scaler implies greater contribution), and the components themselves may reflect the different modalities of the data (in the above

\[ 2 \text{Here, we are using MATLAB notation, in which ':' denotes the entire range of a mode, and 'l:k' denotes \{l,l+1, \ldots,k\}.} \]
example, \(u_i^{(1)}\) are associated with the different subjects, while \(u_i^{(2)}, u_i^{(3)}\) components are associated with features and timepoints respectively.

Computing an approximation in the form of a CP decomposition of a given tensor \(A\) is usually accomplished by solving the following optimization problem:

\[
|\Sigma; U_1, ..., U_N| = \arg \min_{\Sigma, U_1, ..., U_N} \|A - \sum_{i=1}^{r} \hat{\sigma}_i u_i^{(1)} \circ ... \circ u_i^{(N)}\|^2_F
\]

where \(\|X\|_F\) denotes the Frobenius norm of the tensor \(X\):

\[
\|X\|_F^2 = \sum_{i_1=1}^{d_1} \cdots \sum_{i_N=1}^{d_N} x_{i_1,i_2,...,i_N}^2
\]

The current gold-standard method for solving this problem is Alternating Least Squares (ALS) [6].

It is generally hard to solve Problem (2) as it is non-convex, thus potentially having many local minimizers which are not global minima. Indeed, even the gold-standard method (ALS) is not guaranteed to find a global minimum. Moreover, given a data tensor \(A\) (subjects, features and time), and its CP form approximate \(A_\epsilon = [\Sigma; U_1, U_2, U_3]\), suppose that we are provided with data for a new participant \(X \in \mathbb{R}^{1 \times p \times n}\), then the task of extending the current factorization to the new data tensor, which is \(A\) augmented with the new sample \(X\), is far from a trivial one. Such out-of-sample extension capability is a fundamental requirement from any embedding algorithm that we wish to use as a step in a ML pipeline.

The aim of TCAM is to provide a tensor-based PCA-like tool that is better than CP, in terms of ease of interpretation of the model’s outcomes and mathematical properties that ensure safe application to downstream statistical analysis and ML workflows.

A PCA

Since we strive to have to claim that TCAM is ‘PCA-like’, it is useful to first present a brief definition of PCA, and discuss it characteristics. All constructions, definitions and properties presented in this section are taken from [7].

Let \(A \in \mathbb{R}^{m \times p}\) be a data matrix with rows \(a_1, ..., a_m\) corresponding to \(m\) samples, and assume that \(A\) has been centered (so the column means are zero). PCA is defined by an orthogonal linear transformation \(W\) transforming the rows of \(A\) to a new coordinates system, in which, the largest portion of variation in the data lies on the first axis (called the first principal component of \(A\)), i.e. \(z_i = Aw_i\) where \(w_i \in \mathbb{R}^p\) is the maximizer of \(w^TA^TAw\) subjected to \(\|w\|_2^2 = 1\). The \(k\)th largest portion of variation lies on the \(k\)th principal component of \(A\), i.e.

\[
z_k = Aw_k, \quad w_k = \arg \max_{\|w\|_2^2=1} w^TA^{(k)}TA^{(k)}w
\]

where given the first \(k-1\) principal components of \(A\) we define \(A^{(k)} = A(I - \sum_{i=1}^{k-1} w_i w_i^T)\). The complete factorization yields the expression \(Z = AW\), where rows of the matrix \(Z \in \mathbb{R}^{m \times p}\) are called the sample PCA scores (the \(i\)th row of \(Z\) contains the sample PC scores of the \(i\)th sample \(a_i\)). The orthogonal \(p \times p\) matrix \(W\) is called the weight or coefficients matrix. The rank \(k\)-truncated PCA of a \(m \times p\) matrix \(A\) is defined by

\[
Z_k = AW_k \in \mathbb{R}^{m \times k}
\]

where \(W_k := [w_1, ..., w_k]\) is an \(p \times k\) matrix with orthonormal rows; \(W_k^TW_k = I_k\). The \(k\)-truncated PCA \(Z_k\) consists of only the first \(k\) principal components.

The sample variance-covariance matrix of the collection \(\{a_i\}_{i=1}^m\) of \(m\), (centered) \(p\)-dimensional samples \(S\) is given by \(S = (m - 1)^{-1}A^TA\).

**Property A.1** ([7, Chapter 2, Properties A1 and A2]). *For any integer \(q = 1, ..., p\), consider the transformation \(y_i = B_\Sigma a_{i},\) where \(y \in \mathbb{R}^q\) and \(B\) is a \(p \times q\) matrix with \(q\) orthonormal columns. Define \(S_y = B_\Sigma^T S_B \in \mathbb{R}^{q \times q}\) the variance-covariance matrix for \(y\). The variance component of \(S_y\), quantified by \(Tr(S_y)\), is maximized when taking \(B = W_q\) and minimized when \(B = [w_{p-q,1}, ..., w_{p-q,p}]\).*

**Property A.2** ([7, Chapter 3, Eq. 3.1.4, Property A3]). *\(S = \sum_{i=1}^{p} \lambda_i w_i w_i^T\)*

**Property A.3** ([7, Chapter 3, Property G4]). *Let \(\hat{A} = [a_1; ..., a_m]\) be a \(m \times p\) matrix of \(m\), \(p\)-dimensional observations. Define \(A\) to be the \(m \times p\) matrix whose \(i\)th row is \(a_i - \bar{a}\) where \(\bar{a} = \frac{1}{m} \sum_{j=1}^{m} a_j\) and consider the matrix \(AA^T \in \mathbb{R}^{m \times m}\). The \(i\)th diagonal element of \(AA^T\) is the squared Euclidean distance of the sample \(a_i\) from the point \(\bar{a}\) that is the center of gravity of the points \(a_1, ..., a_m\). Also, the \(i, j\) entry of \(AA^T\), given by \((a_i - \bar{a}, a_j - \bar{a})\), is the cosine of the angle between the lines joining the points \(a_i\) and \(a_j\), to \(\bar{a}\), scaled by the distances of \(a_i\) and \(a_j\) from \(\bar{a}\). Suppose that \(a_1, ..., a_m\) are projected to a \(q\)-dimensional subspace using a linear orthogonal transformation \(y_i = B_\Sigma^T a_{i},\) and let \(\bar{y} = [\bar{y}_1; ..., \bar{y}_m]\) be the projection of \(y\) for \(\bar{y}_i = m^{-1} \sum_{j=1}^{m} y_{ij} \in \mathbb{R}^q\). Then the choice \(B = W_q\) minimizes the distortion \(\|YY^T - AA^T\|_F^2\).*
Singular Value Decomposition. One practical way of describing and computing PCA for a matrix \( \mathbf{A} \in \mathbb{R}^{m \times p} \), is using the singular value decomposition of its column centered form. Let again \( \mathbf{A} \) denote the column centered form of \( \hat{\mathbf{A}} \), and let \( \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \) be its Singular Value Decomposition (SVD). That is, \( \mathbf{U} \in \mathbb{R}^{m \times m} \) and \( \mathbf{V} \in \mathbb{R}^{p \times p} \) are orthonormal matrices, and \( \Sigma \in \mathbb{R}^{m \times p} \) is a matrix with non-negative elements \( \sigma_1 \geq \sigma_2 \geq \ldots \) on its main diagonal and zeros elsewhere. The sample variance-covariance matrix \( \mathbf{S} \) can rewritten as \( \mathbf{S} = (m-1)^{-1} \mathbf{V} \Sigma \Sigma \mathbf{V}^T \), thus the columns of \( \mathbf{V} \) are the eigenvectors of \( \mathbf{S} \) and, according to Property A.2, the coefficients in the PCA decomposition of \( \hat{\mathbf{A}} \).

Let us denote the rank-\( r \) truncated SVD of \( \mathbf{A} \) by \( \mathbf{A}_r = \mathbf{U}_r \Sigma_r \mathbf{V}_r^T \) where \( \mathbf{U} \) and \( \mathbf{V} \) are the matrices obtained be taking the first \( r \) columns of \( \mathbf{U} \) and \( \mathbf{V} \) respectively, and \( \Sigma_r = \text{diag}(\sigma_1, \ldots, \sigma_r) \). Suppose that the rank of \( \mathbf{A} \) is greater or equal to \( r \), then the rank of \( \mathbf{A}_r \) is exactly \( r \). This truncation of the SVD enjoys many algebraic properties, one such major result is the Eckart-Young-Mirsky Theorem stating that the \( r \)-rank truncated SVD of a matrix is in a sense the best approximate of rank lower or equal to \( r \):

\[
\mathbf{A}_r = \arg \min_{\mathbf{A}, \text{rank}(\mathbf{A}) \leq r} \| \mathbf{A} - \hat{\mathbf{A}} \|^2_F
\]

This result directly implies Properties A.1 and A.3 of the PCA that are concerned with the maximization of the variance and minimizing the distortion of the projected configuration. Thus, we see that the SVD can be used not only as an alternative construction algorithm for the PCA, but also, due to Eckart-Young-Theorem, can serve as the mathematical justification to some of the PCA’s key properties.

B Tensor Component Analysis

A recent work by Kilmer et. al. introduced a tensor version of the Eckart-Young Theorem [1], stating that the truncated tensor SVD is the best low rank approximate within a specified tensor-tensor product framework based on the *\( \ast_m \)-product. The *\( \ast_m \)-product operation is defined by an invertible matrix \( \mathbf{M} \), and the best low rank approximation results from [1] were established for matrices \( \mathbf{M} \) that are nonzero multiple of a unitary matrix \( (\mathbf{M}^T \mathbf{M} = \mathbf{M} \mathbf{M}^T = c^2 \mathbf{I}_n \text{ for some constant } c \neq 0) \). In this work, we only consider \( \mathbf{M} \) that are unitary matrices.

We construct the tcam on top the tensor SVD introduced by Kilmer et. al., and utilize Kilmer’s Eckart-Young-like result to establish tensor analogs of Properties A.1 and A.3, in the same way that matrix SVD can be used for deriving the same properties for matrices. As a prelude to our discussion, we present some key notions and operations necessary for our derivation. Elaborated introduction and discussion can be found in [1].

B.1 Tensor-tensor \( \ast_m \)-product framework

We begin with some definitions. Let \( \mathbf{M} \) be an \( n \times n \) orthogonal matrix \( (\mathbf{M}^T \mathbf{M} = \mathbf{I}_n = \mathbf{M} \mathbf{M}^T) \), and a tensor \( \mathbf{A} \in \mathbb{R}^{m \times p \times n} \). We define the domain transform specified by \( \mathbf{M} \) as \( \hat{\mathbf{A}} := \mathbf{A} \times_3 \mathbf{M} \), where \( \times_3 \mathbf{M} \) denotes the tensor-multiplication of applying \( \mathbf{M} \) to each of the tensor \( n \) dimensional tubal fibers \( (\mathbf{A}_{i,j,:}) \). The transpose of a real \( m \times p \times n \) tensor \( \mathbf{A} \) with respect to \( \mathbf{M} \), denoted by \( \mathbf{A}^T \), is a \( p \times m \times n \) tensor for which \( [\hat{\mathbf{A}}^T]_{i,:,:} = [\hat{\mathbf{A}}]^T_{i,:,:} = [\hat{\mathbf{A}}]^T_{i,:,:} \). Given two tensors \( \mathbf{A} \in \mathbb{R}^{m \times p \times n} \) and \( \mathbf{B} \in \mathbb{R}^{p \times l \times n} \), the facewise tensor-product of \( \mathbf{A} \) and \( \mathbf{B} \), denoted by \( \mathbf{A} \times \mathbf{B} \), is the \( m \times l \times n \) tensor for which \( [\mathbf{A} \times \mathbf{B}]_{i,:,:} = \mathbf{A}_{i,:,:} \mathbf{B}_{i,:,:} \).

The tensor-tensor \( \ast_m \)-product of \( \mathbf{A} \in \mathbb{R}^{m \times p \times n} \) and \( \mathbf{B} \in \mathbb{R}^{p \times l \times n} \) is defined by \( \mathbf{A} \ast_m \mathbf{B} := (\hat{\mathbf{A}} \ast \hat{\mathbf{B}}) \times_3 \mathbf{M}^{-1} \in \mathbb{R}^{m \times l \times n} \). A few definitions now naturally follow. The \( p \times p \times n \) identity tensor with respect to \( \ast_m \) is the tensor \( \mathbf{I} \) such that for any tensor \( \mathbf{E} \in \mathbb{R}^{p \times p \times n} \) it holds that \( \mathbf{I} \ast_m \mathbf{E} = \mathbf{E} = \mathbf{E} \ast_m \mathbf{I} \). In situations where the dimensions are unclear from context, we use \( \mathbf{I}_m \) to denote the \( m \times m \times n \) identity tensor. Two tensors \( \mathbf{A}, \mathbf{B} \in \mathbb{R}^{1 \times m \times n} \) are called \( \ast_m \)-orthogonal slices if \( \mathbf{A}^T \ast_m \mathbf{B} = 0 \), where \( 0 \in \mathbb{R}^{1 \times 1 \times n} \) is the zero tube fiber, while \( \mathbf{Q} \in \mathbb{R}^{m \times n} \) is called \( \ast_m \)-diagonal matrices with entries that are either ones or zeros.

B.2 The TSVDM

With the \( \ast_m \)-product framework set-up, it is now possible to introduce the tensor singular value decomposition (TSVDM). Let \( \mathbf{A} \in \mathbb{R}^{m \times p \times n} \) be a real tensor, then is possible to write the full tubal singular value decomposition of \( \mathbf{A} \) as \( \mathbf{A} = \mathbf{U} \ast_m \mathbf{S} \ast_m \mathbf{V}^T \), where \( \mathbf{U}, \mathbf{V} \) are \( m \times m \times n \) and \( p \times p \times n \) \( \ast_m \)-unitary tensors respectively, and \( \mathbf{S} \in \mathbb{R}^{m \times p \times n} \) is an \( f \)-diagonal tensor, that is, a tensor whose frontal slices \( (\mathbf{S}_{i,:,:}) \) are
matrices with zeros outside their main diagonal (see [1] for additional details). We use the notation $\hat{\sigma}^{(i)}_j$ to denote the $j^{th}$ largest singular value on the $i^{th}$ lateral face of $\hat{S}$: $\hat{\sigma}^{(i)}_j := \hat{S}_{j,i,i}$.

The tsvdm construction makes it possible to expand the concept of tensor-rank discussed earlier. The t-rank of $A$ is the number of nonzero tubes of $A$: $r = |\{i = 1, \ldots, n : \|S_{i,:}\|_F^2 > 0\}|$. Additionally, the multi-rank of $A$ under $*_{\text{max}}$, denoted by the vector $\rho \in \mathbb{N}^n$ whose $i^{th}$ entry is $\rho_i = \text{rank}(\hat{A}_{:,i,i})$, and the implicit rank under $*_{\text{max}}$ of a tensor $A$ with multi-rank $\rho$ under $*_{\text{max}}$ is $r = \sum_{i=1}^n \rho_i$.

The definitions of tensor t-rank and multi-rank under $*_{\text{max}}$ also make it possible to define tsvdm rank truncation with respect to these ranks. The tensor $A^{(q)} = U_{:,q} *_{\text{max}} S_{q,q,q} V_{:,q,q}^T$ denotes the t-rank $q$ truncation of $A$ under $*_{\text{max}}$. Let the tuples $\rho \in \mathbb{N}^n$, where $\sigma$ implied by target explicit rank $r$, are the ‘best’ $t$/multi-rank $q$, $\rho$ approximations of $\mathcal{A}$ respectively, where ‘best’ refers to entrywise squared error, i.e. the Frobenius norm of the error. In other words, $\mathcal{A}^{(q)}$ and $\mathcal{A}_\rho$ are the global minimizers of $\|\mathcal{A} - \mathcal{B}\|_F^2$ for $\mathcal{B}$ with t-rank $q$ (respectively, multi-rank $\rho$) under $*_{\text{max}}$ of the same dimensions as $\mathcal{A}$.

Let $\mathcal{A} = U *_{\text{max}} S *_{\text{max}} V^T \in \mathbb{R}^{m \times p \times n}$, we will use $j_1, \ldots, j_{np}$ and $i_1, \ldots, i_{np}$ to denote the indexes of the non-zeros of $\mathcal{S}$ ordered in decreasing order. That is

$$\hat{\sigma}_\ell := \hat{\sigma}^{(i)}_j$$

(4)

where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \cdots \geq \hat{\sigma}_n$.

In this work, we consider truncation with respect to the explicitly given implicit rank under $*_{\text{max}}$. For $q = 1, \ldots, pm$, the explicit rank-$q$ truncation under $*_{\text{max}}$ of $A$ is the tensor $A_\rho$ of multi-rank $\rho$ under $*_{\text{max}}$ where

$$\rho_i = \max\{j = 1, \ldots, p \mid (j,i) \in \{(j_1, i_1), \ldots, (j_q, i_q)\}\}.$$  

(5)

In words, we keep the $q$ top singular values of any frontal slice of $\hat{S}$, and zero out the rest. Note that the explicit rank-$q$ truncation is not always uniquely defined by $q$ since ties between singular values result in multiple possible choices for the multi-rank $\rho$. However, all explicit-rank $q$ truncations are equivalent in that they produce identical reconstruction errors. Indeed, let $A_\rho$ be a multi-rank $\rho$ truncation of $A$, implied by target explicit rank $q$, then $\|A - A_\rho\|_F^2 = \sum_{\ell=q+1}^n \hat{\sigma}_\ell^2$, meaning that the reconstruction error remains the same for any choice of multi-rank $\rho$ in eq. (5). Furthermore, given the definition of explicit rank truncation, we get the following.

Claim B.1. Suppose that $M$ is a unitary matrix. Let $A \in \mathbb{R}^{m \times p \times n}$, with a full tsvdm $A = U *_{\text{max}} S *_{\text{max}} V^T$, and $\rho = [\rho_1, \ldots, \rho_n]$ be the multi-rank defined by explicit rank-$q$ truncation of $A$ in eq. (5). Let $A_\rho = U_{\rho} *_{\text{max}} S_{\rho} *_{\text{max}} V_{\rho}^T \in \mathbb{R}^{m \times p \times n}$, the multi-rank $\rho$ truncation of $A$, then $A_\rho$ is the best implicit rank-$q$ approximation of $A$.

Proof. Let $\varphi = [\varphi_1, \ldots, \varphi_n]$ be a multi-rank such that $\sum_{i=1}^n \varphi_i = q$. Then $A_\varphi = U_{\varphi} *_{\text{max}} S_{\varphi} *_{\text{max}} V_{\varphi}^T \in \mathbb{R}^{m \times p \times n}$ is the best multi-rank $\varphi$ approximation of $A$. Let the tuples $(j_1, i_1), \ldots, (j_q, i_q)$ denote an ordering of the singular values $\{\hat{\sigma}^{(i)}_j\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, \varphi_i$, such that $\hat{\sigma}_1 \geq \hat{\sigma}_2 \cdots \geq \hat{\sigma}_q$ where $\hat{\sigma}_1 = \hat{\sigma}^{(i)}_j$. Then, by construction of $\hat{\sigma}_\ell$ in eq. (4), we have that $\hat{\sigma}_1 \geq \hat{\sigma}_2 \cdots \geq \hat{\sigma}_q$ for all $\ell = 1, \ldots, q$.

Now, we have that

$$\|A - A_\varphi\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p (\hat{\sigma}^{(i)}_j)^2 - \sum_{i=1}^n \sum_{j=1}^p (\hat{\sigma}^{(i)}_j)^2$$

$$= \sum_{h=1}^{\rho_n} \hat{\sigma}_h^2 - \sum_{h=1}^{q} \hat{\sigma}_h^2$$

$$\geq \sum_{h=1}^{\rho_n} \hat{\sigma}_h^2 - \sum_{h=1}^{q} \hat{\sigma}_h^2$$

$$= \|A - A_\varphi\|_F^2$$

Thus, we established that the best implicit rank $q$ approximation of $A$, is the tensor $A_\rho$ with multi-rank $\rho$ implied by explicit rank $q$ truncation of $A$. □
We also note that for the explicit rank-\(q\) truncation of \(\mathcal{A}\), it holds that
\[
\|\mathcal{A}_\rho\|_F^2 = \sum_{h=1}^{q} \hat{\sigma}_h^2
\]  
(6)
where \(\hat{\sigma}_h\) is defined by eq. (4).

### B.3 A PCA-like tensor decomposition

The following construction was first presented in [8]. Short presentation of the objects we are concerned with will be followed by novel results regarding their algebraic and geometric properties. For elaborate definitions and discussion, see [8, 1].

Let \(\mathcal{A} = U \ast_{SM} S \ast_{SM} Y^T \in \mathbb{R}^{m \times p \times n}\) in mean deviation form (see Supporting Information D section), and consider the following expressions [8]:
\[
\mathcal{Z} = \mathcal{A} \ast_{SM} Y
\]  
(7)
and
\[
\mathcal{Z}_{\rho} := \mathcal{A} \ast_{SM} Y_{\rho}
\]  
(8)
where the tensor \(Y_{\rho} \in \mathbb{R}^{p \times p \times n}\) is obtained from explicit rank \(q\) truncated tsvdm of \(\mathcal{A}\). The expressions in eq. (8) makes use in the tsvdm to form a construction of high resemblance to the truncated PCA shown in eq. (3). We also point that \(Y_{\rho}\) is a member of particular family - it is a pseudo \(\ast_{SM}\)-orthogonal tensor of implicit rank \(q\).

Let \(\mathcal{B} \in \mathbb{R}^{p \times p \times n}\) be a pseudo \(\ast_{SM}\)-orthogonal tensor of implicit rank \(q\), and notice that \((B^T \ast_{SM} B)_{\times_{3} M}\) is an f-diagonal tensor with diagonals consists of zeros and ones, and for which the number of non-zero entries is \(q\). Define
\[
\mathbf{y} = \mathcal{A} \ast_{SM} \mathcal{B}
\]  
(9)

In Property A.1 the formulation is concerned with maximizing (or minimizing) the variance component of a random variable \(\mathbf{y}\), that is quantified using the trace of the variance-covariance matrix of that random variable. The objective of maximizing (minimizing) the sample variance of \(\mathbf{y}\), that is proportional to \(\text{Tr}(B^T \ast_{SM} B)\), which, by definition of the Frobenius norm, results in \(\|AB\|_F^2\). In absence of analog definitions for variance of a tensor valued random variable and the trace of a tensor, we resort to discuss an algebraically similar trait.

Define the sample variance-covariance tensor [8] for \(\mathbf{y}\) as
\[
E_{\mathbf{y}} := B^T \ast_{SM} E \ast_{SM} B
\]  
(10)
where \(E = (m-1)^{-1}A^T \ast_{SM} A \in \mathbb{R}^{p \times p \times n}\) is the sample variance-covariance tensor for \(\mathcal{A}\) (and \(\mathcal{Z}\)).

**Property B.2.** Suppose that \(\mathcal{M}\) is a unitary matrix. Let \(\mathcal{B} \in \mathbb{R}^{p \times p \times n}\) a pseudo \(\ast_{SM}\)-orthogonal tensor of implicit rank \(q\). Then the quantity \(\|A \ast_{SM} B\|_F^2\) is maximized when \(\mathcal{B} = Y_{\rho}\), where \(\mathcal{A}_\rho = U \ast_{SM} S \ast_{SM} Y^T\) is the explicit rank \(q\) truncation of \(\mathcal{A}\).

**Proof of Property B.2.** We have that \(\|A \ast_{SM} B\|_F^2 = \|\hat{A} \Delta \hat{B}\|_F^2 = \sum_{i=1}^{n} \text{Tr}(\hat{Z}_{i,\ldots,i} \hat{A}^T \hat{Z}_{i,\ldots,i} \hat{B}^T)\). Recall that \(\mathcal{A} = U \ast_{SM} S \ast_{SM} Y^T\), and write \(\mathcal{B} = V \ast_{SM} C\) where \(C = V Y^T \ast_{SM} B\). Note that \(C\) is also an implicit rank \(q\) pseudo \(\ast_{SM}\)-orthogonal tensor. So, \(\|A \ast_{SM} B\|_F^2 = \sum_{i=1}^{n} \text{Tr}(C_{i,\ldots,i}^T \hat{S}_{i,\ldots,i} \hat{S}_{i,\ldots,i}^T)\). Thus \(\|A \ast_{SM} B\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \sigma_{j,i}^2 c_{j,i}\), which may be re-ordered according to eq. (4) to obtain
\[
\|A \ast_{SM} B\|_F^2 = \sum_{h=1}^{p} \sigma_{h}^2 c_{j_h,i_h}
\]

Since \(0 \leq c_{j_h,i_h} \leq 1\) and \(\|C\|_F^2 = \sum_{j=1}^{p} c_{j,i} = q\), we have that \(\sum_{h=1}^{p} \sigma_{h}^2 c_{j_h,i_h} \leq \sum_{h=1}^{q} \sigma_{h}^2\). Hence, \(\|A \ast_{SM} B\|_F^2 \leq \sum_{h=1}^{q} \hat{\sigma}_{h}^2\), which according to eq. (6), equals to \(\|A_\rho\|_F^2\) where \(\rho\) is the multi-rank implied by explicit rank \(q\) truncation of \(\mathcal{A}\) (See eq. (5)).

Note that taking \(\mathcal{B} = Y_{\rho}\) yield \(c_{j,i} = (\widehat{V}_{j,i}^T \hat{V}_{\rho}, \ldots, i)(\widehat{V}_{j,i}^T \hat{V}_{\rho}, \ldots, i)^T\), that in turn results in
\[
c_{j,i} = \begin{cases} 1 & \exists h \leq q \text{ s.t. } (j_h, i_h) = (j, i) \\ 0 & \text{otherwise} \end{cases}
\]
thus, the upper bound is achieved for \(\|A \ast_{SM} Y_{\rho}\|_F^2 = \sum_{h=1}^{q} \hat{\sigma}_{h}^2\).  
\[\square\]
The tensor equivalent result for Property A.3 says the minimizer of the distortion under pseudo-orthogonality constraints is again $V_\rho$.

**Property B.3.** Suppose that $M$ is a unitary matrix. Let $A \in \mathbb{R}^{m \times p \times n}$ in mean deviation form and define $Y = A *_m B \in \mathbb{R}^{m \times p \times n}$ where $B \in \mathbb{R}^{p \times p \times n}$ is a pseudo $*_m$-orthogonal tensor of implicit rank $q$. Then the distortion formed by $B$, measured as $\|Y *_m Y^T - A *_m A^T\|^2_F$, is minimized when $B = V_\rho$.

**Proof of Property B.3.** Note that $\hat{A}_{:,i} A_{:,i}^T$ is symmetric positive semidefinite, thus, $\hat{A}_{:,i} A_{:,i}^T = \sum_{j=1}^{\rho_j} (\hat{\sigma}_j^{(i)})^2 u_j^i (u_j^i)^T$ where $\{u_j^i\}_{j=1}^{\rho_j} = \{u_j^i\}_{j=1}^{\rho_j}$ are unit orthogonal vectors, $(\hat{\sigma}_1^{(i)})^2 \geq (\hat{\sigma}_2^{(i)})^2 \geq \cdots \geq (\hat{\sigma}_q^{(i)})^2$ are the eigenvalues of $A_{:,i} A_{:,i}^T$, and $\theta = \{\theta_1, \ldots, \theta_n\}$ is the multi-rank of $A$ under $*_m$. Similarly, we have that $\tilde{Y}_{:,i} \tilde{Y}_{:,i}^T = \sum_{j=1}^{\varphi_i} (\tilde{\sigma}_j^{(i)})^2 p_j^i (p_j^i)^T$ where $\{p_j^i\}_{j=1}^{\varphi_i} = \{p_j^i\}_{j=1}^{\varphi_i}$ are unit orthogonal vectors, $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_q$ are the eigenvalues of $\tilde{Y}_{:,i} \tilde{Y}_{:,i}^T$, and $\varphi_i$ denote the rank of $\tilde{Y}_{:,i} \tilde{Y}_{:,i}^T$.

Note that $\tilde{Y}_{:,i} \tilde{Y}_{:,i}^T = \sum_{j=1}^{\varphi_i} \mu_j^i \tilde{B}_{:,i}^T \tilde{B}_{:,i}^T$ (with $\tilde{B}_{:,i}^T \tilde{B}_{:,i}^T$ being a projection matrix), therefore, it must hold that $\varphi_i \leq \theta_i$ for all $i = 1, \ldots, n$. Combining the last two, we have

$$\|Y *_m Y^T - A *_m A^T\|^2_F \geq \sum_{i=1}^{n \rho} (\hat{\sigma}_i^{(i)})^4 - \sum_{i=1}^{n \varphi} (\tilde{\sigma}_i^{(i)})^4 \geq \sum_{h=1}^{n \rho} (\hat{\sigma}_h)^4 - \sum_{h=1}^{n \varphi_i} (\tilde{\sigma}_h)^4$$

where $\hat{\sigma}_h = \hat{\sigma}_{j_h,i_h}$ (See eq. (4)). Simple calculations reveals that setting $B = V_\rho$ results in exactly

$$\|A *_m V_\rho *_m V_\rho^T *_m A^T - A *_m A^T\|^2_F = \sum_{h=1}^{n \rho} (\hat{\sigma}_h)^4 - \sum_{h=1}^{n \varphi_i} (\tilde{\sigma}_h)^4$$

so we conclude that $V_\rho$ is a global minimizer of the distortion as it reaches the global lower bound.

**B.4 Explicit rank truncated tensors in vector representation**

Having established optimality properties with respect to variance and distortion for explicit rank truncated tensors (Properties B.2 and B.3), we turn to do the same for vector representation of these truncations. The vector form of truncated tensors is obtained by mode-1 unfolding. Let $X \in \mathbb{R}^{1 \times p \times n}$, we define $\text{vec}(X) \in \mathbb{R}^{pn}$ as

$$\text{vec}(X) = [X_{1,1,1}, X_{1,2,1}, \ldots, X_{1,p,1}, X_{1,1,2}, X_{1,2,2}, \ldots, X_{1,p,n}]$$

For tensors $X \in \mathbb{R}^{m \times p \times n}$, we slightly abuse the above notation by letting vec($X$) $\in \mathbb{R}^{m \times np}$ denote the matrix whose $\ell^{th}$ row is vec($X_{:,\ell}$) (so, even if $m = 1$ we will view vec($X$) as a row vector).

Consider the family $Q$ of tensor-vector mappings $\mathbb{R}^{p \times p \times n}$ to $\mathbb{R}^{m \times np}$ whose members are of the form

$$Q_B(X) = \text{vec}((X *_m B) \times_3 M)$$

where $B \in \mathbb{R}^{p \times p \times n}$ is a pseudo $*_m$-orthogonal tensor. We define the rank of a member $Q_B$ in $Q$ as the implicit rank of $B$.

In analogy to Supporting Information A and B.3 we present the following properties.

**Property B.4.** Suppose that $M$ is a unitary matrix. Let $A \in \mathbb{R}^{m \times p \times n}$ be in mean deviation form and $Y := Q_B(A) \in \mathbb{R}^{m \times np}$ for some pseudo $*_m$-orthogonal $B$. As in Supporting Information A, we define $S_Y := (m - 1)^{-1} Y Y^T$ the sample variance-covariance matrix of $Y$. Then, the rank $q$ member of $Q$ for which the variance component $(m - 1)^{-1} \text{Tr}(Y Y^T)$ is maximized, is given by $Q_{V_\rho}$ where $A_\rho = U_\rho *_m S_\rho *_m V_\rho^T$ is the multi-rank $\rho$ truncation $A$ under $*_m$ implied by explicit rank $q$ truncation of $A$.  

7
Proof of Property B.4. Note that
\[ \text{Tr}(S_Y) \propto \text{Tr}(YY^T) \]
\[ = \text{Tr}\left(\sum_{h=1}^{np} Y_{:,h}(Y_{:,h})^T\right) \]
\[ = \sum_{h=1}^{np} \text{Tr}(Y_{:,h}(Y_{:,h})^T) \]
\[ = \sum_{i=1}^{n} \sum_{j=1}^{p} \text{Tr}(\tilde{y}_{:,j,i}(\tilde{y}_{:,j,i})^T) \]
\[ = \sum_{i=1}^{n} \sum_{j=1}^{p} (\tilde{y}_{:,j,i})^T \tilde{y}_{:,j,i} \]
\[ = \sum_{i=1}^{n} \|\tilde{y}_{:,j,i}\|^2_F = \|y\|^2_F. \]

where \( y = A \ast_m B \). So we conclude that \( \text{Tr}(S_Y) = (m-1)^{-1}\|A \ast_m B\|^2_F \). Now, by Property B.2, we have that the implicit rank \( q \) for which this quantity is maximal is \( V_{\rho}^q \).

Property B.4 provides a variance maximization result for flattened explicit rank \( q \) truncations that is similar to Properties A.1 and B.2 which state variance maximization for traditional rank \( q \) truncations of matrices (PCA) and explicit rank-\( q \) truncations of tensors respectively.

We proceed with presenting a slightly modified version of Properties A.3 and B.3. Recall that Properties A.3 and B.3 discuss the minimization of the Frobenius norm of the difference between configuration matrices (or tensors) of the truncated and of the complete representation:

\[ \|YY^T - AA^T\|^2_F, \quad \|y \ast_m y^T - A \ast_m A^T\|^2_F. \]

Formulating analog property for flattened rank \( q \) truncated representations, requires a definition of the complete flattened representation of a tensor. Given \( A \in \mathbb{R}^{m \times p \times n} \) in mean deviation form, we let

\[ A := QV(A) = \text{vec}(\widehat{z}) \quad (11) \]

denote the complete flattened representation of \( A \), where \( A = U \ast_m S \ast_m V^T = Z \ast_m V^T \). We argue that the definition of the complete flattened representation stated above is rather natural since it holds all information regarding the original tensor \( A \), and reconstructing \( A \) from \( A \) is possible by applying an inverse vec operation to \( A \), followed by a \( \times_3 M^{m-1} \) tensor-matrix product, and a right \( \ast_m \) product with \( y^T \). Additional argument in favor of the choice in eq. (11), is that the sample variance component in \( S_A \) is identical to that of the original tensor, as could be deduced from the proof of Property B.4.

Therefore, we will use \( A \) (eq. (11)) as the reference point for quantifying the distortion in configuration of flattened truncated rank representations of \( A \). Note that in contrast to the fact that the expressions \( \|y\|^2_F \) and \( \|Y\|^2_F \) are identical (a fact that was used to show Property B.4), the quantities \( \|y \ast_m y^T\|^2_F \) and \( \|YY^T\|^2_F \) are not equal in general. This forces a slight modification in the definition of the concept of distortion itself. Instead of using the traditional formulation \( \|YY^T - AA^T\|^2_F \) as in [7], we replace the Frobenius norm with the nuclear norm. Given a matrix \( X \), the nuclear norm of \( X \), denoted by \( \|X\|_n \), is defined as the sum of \( X \)'s singular values. In the special case where \( X \) is a positive semi-definite matrix (a case which we briefly describe as \( X \succeq 0 \)), we have that \( \|X\|_n = \text{Tr}(X) \). We now show the following.

Property B.5. Suppose that \( M \) is a unitary matrix. Let \( A \in \mathbb{R}^{m \times p \times n} \) be in mean deviation form, \( Y := Q_B(A) \in \mathbb{R}^{m \times np} \) and \( A \in \mathbb{R}^{m \times pn} \) defined by eq. (11). Then, the rank \( q \) member of \( Q \) for which the distortion \( \|YY^T - AA^T\|_n \) is minimized, is given by \( V_{\rho}^q \), where \( A_\rho = U_\rho \ast_m S_\rho \ast_m V^T_\rho \) is the multi-rank \( \rho \) truncation \( A \) under \( \ast_m \) implied by explicit rank \( q \) truncation of \( A \).

Proof of Property B.5. Begin by noting that for any \( \ell \) and \( k \) between 1 and \( m \) , we have

\[ [YY^T]_{\ell,k} = \text{vec}(\tilde{y}_{\ell,:})^T \text{vec}(\tilde{y}_{k,:}) \]
\[ = \sum_{i=1}^{n} \tilde{y}_{\ell,:)i} \tilde{y}_{k,i}^T \]
\[ = \sum_{i=1}^{n} \tilde{y}_{\ell,:)i} \tilde{y}_{k,:i}^T. \]
and similarly, we get \([\mathbf{A}\mathbf{A}^T]_{\ell,k} = (\sum_{i=1}^{n} \hat{z}_{\ell,i} \hat{z}_i^T)_{\ell,k}\), thus, \(\mathbf{Y}\mathbf{Y}^T - \mathbf{A}\mathbf{A}^T = \sum_{i=1}^{n} \hat{y}_{\ell,i} \hat{y}_i^T - \hat{z}_{\ell,i} \hat{z}_i^T\).

We write \(\mathbf{Q} = \mathbf{V}^T \otimes_{\mathbf{M}} \mathbf{B}\), and note that \(\mathbf{Q}\) is a pseudo \(\otimes_{\mathbf{M}}\)-orthogonal tensor of implicit rank \(q\), then

\[
\hat{y}_{\ell,i} \hat{y}_i^T - \hat{z}_{\ell,i} \hat{z}_i^T = (\hat{Q}_{\ell,i} \hat{Q}_i^T - \mathbf{I}) \hat{z}_i^T
\]

where \(\hat{Q}_{\ell,i} \hat{Q}_i^T - \mathbf{I}\) is a \(p \times p\) symmetric matrix. Since \(\mathbf{I} - \hat{Q}_{\ell,i} \hat{Q}_i^T\) is a projection matrix, we have that \(\mathbf{A}\mathbf{A}^T - \mathbf{Y}\mathbf{Y}^T \preceq 0\), and hence

\[
\|\mathbf{A}\mathbf{A}^T - \mathbf{Y}\mathbf{Y}^T\|_* = \text{Tr}(\mathbf{A}\mathbf{A}^T - \mathbf{Y}\mathbf{Y}^T)
\]

Note that

\[
\text{Tr}(\mathbf{A}\mathbf{A}^T - \mathbf{Y}\mathbf{Y}^T) = \text{Tr}(\sum_{i=1}^{n} \hat{z}_{\ell,i} \hat{z}_i^T - \hat{y}_{\ell,i} \hat{y}_i^T)
\]

\[
= \|\mathbf{Z}\|_F^2 - \|\mathbf{Y}\|_F^2
\]

a quantity that is that is minimized when \(\|\mathbf{Y}\|_F^2\) is maximized. Since \(\|\mathbf{Y}\|_2^2\) is proportional to the sample variance component of \(\mathbf{y}\), according to the proof of Property B.2, \(\|\mathbf{A}\mathbf{A}^T - \mathbf{Y}\mathbf{Y}^T\|_*^2\) is minimized when \(\mathbf{Y} = Q_{\mathbf{V}_\rho}(\mathbf{A})\), that is, \(\mathbf{B} = \mathbf{V}_\rho\).

\(\square\)

### C  TCAM

A closer inspection of \(\mathbf{V}_\rho\) reveals that the flattened rank \(q\) truncated representation \(Q_{\mathbf{V}_\rho}\), which was shown in Supporting Information B.4 to have optimality guarantees with respect to distortion and variance, can be considerably compressed. Consider the \(i^{th}\) frontal face of \(\mathbf{V}_\rho\), then we have that

\[
\|\langle \mathbf{V}_{\rho} \rangle_{:,j,i} \|_F^2 = \begin{cases} 1 & \exists h \leq q \text{ s.t. } (j_h, i_h) = (j, i) \\ 0 & \text{otherwise} \end{cases}
\]

where we used the ordering of frontal and lateral indices defined in eq. (4). Thus, for any tensor \(\mathbf{X} \in \mathbb{R}^{1 \times p \times n}\) we have that \(\mathbf{X}_{:,j,i} = \langle \mathbf{V}_{\rho} \rangle_{:,j,i}\) can obtain nonzero values only for indices \(j = 1, \ldots, p\) such that \((j, i) = (j_h, i_h)\) for some \(h \leq q\) essentially making \(Q_{\mathbf{V}_\rho}\) a mapping from \(\mathbb{R}^{1 \times p \times n}\) to \(\mathbb{R}^q\).

Given \(\mathbf{A} \in \mathbb{R}^{m \times p \times n}\) with explicit rank \(q\) truncation \(\mathbf{A}_\rho = \mathbf{U}_\rho \otimes_{\mathbf{M}} \mathbf{S}_\rho \otimes_{\mathbf{M}} \mathbf{V}_\rho^T\), we let \(Q : \mathbb{R}^{1 \times p \times n} \rightarrow \mathbb{R}^q\) denote the compact representation of \(Q_{\mathbf{V}_\rho}\);

\[
Q(\mathbf{X}) = [x_1, x_2, \ldots, x_q]
\]

where \(x_h = [(\mathbf{X} \otimes_{\mathbf{M}} \mathbf{V}_\rho) \times_3 \mathbf{M}]_{:,j_h,i_h}\).

For any \(\mathbf{X} \in \mathbb{R}^{k \times p \times n}\), let \(\mathbf{X} := Q_{\mathbf{V}_\rho}(\mathbf{X})\) and \(\mathbf{X} := Q(\mathbf{X})\). Clearly, \(\text{Tr}(\mathbf{X}^T \mathbf{X}) = \text{Tr}(\mathbf{X}^T \mathbf{X})\) and also \(\mathbf{X}^T = \mathbf{X}^T\). This gives rise to the following:

**Definition C.1 (TCAM).** Let \(\mathbf{A} \in \mathbb{R}^{m \times p \times n}\) in mean deviation form. The TCAM of \(\mathbf{A}\) consists of the following:

1. The TCAM scores, given by the tensor-to-vector mapping \(Q\) specified above. Note that given a new sample \(\mathbf{X} \in \mathbb{R}^{1 \times p \times n}\), the TCAM scores of \(\mathbf{X}\) may be easily obtained by applying \(Q\) to \(\mathbf{X}\).

2. The TCAM loadings (or coefficients) matrix \(\mathbf{V} \in \mathbb{R}^{p \times q}\) with entries \(v_{h,j} = \dot{V}_{j_h,j,i_h}\), where the index \(h\) corresponds to the ordering of the singular values in eq. (4).

The TCAM may be truncated at any target number of factors \(q\) by taking the first \(q\) rows for the scores matrix, and the first \(q\) rows in the coefficients matrix.

The scores of the \(q\) truncated TCAM enjoy the same properties as \(Q_{\mathbf{V}_\rho}(\mathbf{A})\), hence, these scores make variance maximizing vector representation of samples in \(\mathbf{A}\) (Property B.4) while minimizing the distortion with respect to the nuclear norm (Property B.5).

### D  Data pre-processing

**Baseline normalization.** For experiment with timepoints \(t_1, \ldots, t_n\), where \(t_1, \ldots, t_j\) are considered the baseline samples. Let \(s_{t_1}, \ldots, s_{t_n}\) denote the samples collected from subject \(\ell\). Assuming all data are non-negative (as in count data or relative abundance), the log-folds from baseline (LFB) transformation of the data \(s_{t_1}, \ldots, s_{t_n}\) is defined by \(\dot{s}_{\ell k} := \log_2(s_{\ell k})\) where \(\bar{s}_\ell\) is the mean of subject \(\ell\)'s baseline samples: \(s_{t_1}, \ldots, s_{t_j}\). In cases data are normalized, the deviation from baseline (DFB) transformed data \(s_{t_1}, \ldots, s_{t_n}\) is obtained by \(\dot{s}_{\ell} := s_{\ell k} - \bar{s}_\ell\).
**Post antibiotics reconstitution.** Shotgun metagenomics sequencing data of stool samples was downloaded from ENA (project accession number: PRJEB28097). QC filtration and read trimming were done using fastp, followed by removal of reads mapped to human genome by bowtie2 mapper. MetaPhlan3 was used for taxonomic assignment of the reads. For the analysis, we included the timepoints with missing samples of no more than 2 subjects, allowing for single day deviation in any direction. Subjects without at least one sample in each phase of the experiment (baseline, antibiotics, intervention) were excluded. Relative abundance values were capped at $10^{-6}$ and features with maximal values less than $10^{-6}$ were omitted. For the analysis using tcam, data from each participant were LFB transformed. PERMANOVA was computed using truncated distance matrices reconstructed using the minimal number of components (either tcam or PCA) such that the truncation accounts for at least 20% of the total variation in the data. Feature selection for univariate time-series analysis was done by taking the 0.75 quantile of loadings norm computed for the factors demonstrating significant univariate difference (ANOVA). Univariate time-series analysis was performed using lmer.

**ECAM dataset.** 16S rDNA sequencing data and metadata for [9] were obtained from a repository set up by Martino et. al. [10]. Relative abundance values were capped at $10^{-3}$, i.e. features with maximal values less than $10^{-3}$ were omitted, then LFB transformation was applied to the data. Scikit-learn’s SVC was fitted to the first two resulting tcam factors. Univariate time-series analysis was performed using lmer.

**Dietary fiber intervention.** 16S rDNA sequencing data and metadata for [11] were downloaded from ENA (project accession number: PRJNA560950). Overlapping paired-end FASTQ files of 16S amplicon sequencing data were matched and analyzed using the Qiime2 pipeline (q2cli version 2021.4.0) [12]. Poor quality bases were trimmed, sequences were denoised and binned to amplicon sequence variants (ASVs) using the dada2 plugin for Qiime2 [13]. Taxonomic assignment was performed using naive Bayes feature classifier and Greengenes 13_8 database. Given the number of data points (n=34 participants), and under low rank assumption, CTF [3] was applied to the raw count data with target ranks ranging from one to ten, seeking to optimize the representation to show differences between groups. The analysis using tcam was performed on top of two normalization methods: 1) a DFB transformed RCLR normalized data, where the RCLR was applied to the raw count data with target rank set to 7 (after searching for optimal rank, similarly to the procedure employed for CTF), 2) Relative abundance values were capped at $10^{-3}$, i.e. features with maximal values less than $10^{-3}$ were omitted, then LFB transformation was applied to the data.

In all cases, PERMANOVA was applied to between-subjects distance matrices, computed in the embedding space using the minimal number of components such that the explained variance ratio is greater or equal to 0.2. Feature selection for univariate time-series analysis was done by taking the 0.75 quantile of loadings norm computed for the factors demonstrating significant univariate difference (ANOVA). Univariate time-series analysis was performed using lmer.

**Identification of insulin resistance using longitudinal proteomics.** Proteomics data and metadata were downloaded from https://figshare.com/articles/dataset/Multi_Omics_Seasonal_RData/12376508. Three years duration was stratified to trimesters, and repeated samples of the same participant within a trimester were median aggregated. Subjects lacking measurements in more than a single trimester were omitted from the analysis. Missing timepoints were imputed via linear interpolation or forward/backward filled in case of missing last/first trimester. The first year was considered as “baseline”. Since proteomics data contain negative values, the data for each subject were shifted by the median baseline measurements for that subject.

**Pediatric ulcerative colitis.** Sample-specific metadata and final microbial OTU relative abundances were acquired from [14]. Subjects without at least one sample in each timepoint of the experiment (0, 4, 12, 52) were excluded. Relative abundance values were capped at $10^{-4}$, i.e. features with maximal values less than $10^{-4}$ were omitted, and data from each participant were LFB transformed. A multilayer perceptron (MLP) with single 1000 neuron wide hidden layer was trained to predict remission state using the minimal number of tcam factors such that at least 80% of the variation in the data is explained by the factors. The same MLP architecture was trained using PCA for the log2-fold changes of weeks 12 and 52 from the baseline of each subject. Rocket based classifier was obtained by applying Rocket on top of the same pre-processing scheme, followed by logistic model (RidgeClassifier), where the hyper-parameters of both steps were optimized via a grid-search.
References

[1] Misha E. Kilmer, Lior Horesh, Haim Avron, and Elizabeth Newman. Tensor-tensor algebra for optimal representation and compression of multiway data. *Proceedings of the National Academy of Sciences*, 118(28):e2015851118, jul 2021.

[2] Alex H. Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I. Ryu, Krishna V. Shenoy, Mark Schnitzer, Tamara G. Kolda, and Surya Ganguli. Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron*, 98(6):1099–1115.e8, jun 2018.

[3] Cameron Martino, Liat Shenhav, Clarisse A. Marotz, George Armstrong, Daniel McDonald, Yoshiki Vázquez-Baeza, James T. Morton, Lingjing Jiang, Maria Gloria Dominguez-Bello, Austin D. Swafford, Eran Halperin, and Rob Knight. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nature Biotechnology*, pages 1–4, aug 2020.

[4] Omar Delannoy-Bruno, Chandani Desai, Arjun S. Raman, Robert Y. Chen, Matthew C. Hibberd, Jiyi Cheng, Nathan Han, Juan J. Castillo, Garret Coutre, Carlito B. Lebrilla, Ruteja A. Barve, Vincent Lombard, Bernard Henrisatt, Seinen A. Leyn, Dmitry A. Rodionov, Andrei L. Osterman, David K. Hayashi, Alexandra Meynier, Sophie Vinoy, Kyliegh Kirbach, Tara Wilmot, Andrew C. Heath, Samuel Klein, Michael J. Barratt, and Jeffrey I. Gordon. Evaluating microbiome-directed fibre snacks in gnotobiotic mice and humans. *Nature*, 595(7865):91–95, jul 2021.

[5] Frank L. Hitchcock. The Expression of a Tensor or a Polyadic as a Sum of Products. *Journal of Mathematics and Physics*, 6(1-4), 1927.

[6] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, August 2009.

[7] Ian T Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

[8] Ning Hao, Misha E. Kilmer, Karen Braman, and Randy C. Hoover. Facial recognition using tensor-tensor decompositions. *SIAM Journal on Imaging Sciences*, 6(1):437–463, January 2013.

[9] Nicholas A. Bokulich, Jennifer Chung, Thomas Battaglia, Nora Henderson, Melanie Jay, Huilin Li, Arnon D. Lieber, Fen Wu, Guillermo I. Perez-Perez, Yu Chen, William Schweizer, Xuhui Zheng, Monica Contreras, Maria Gloria Dominguez-Bello, and Martin J. Blaser. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine*, 8(343), June 2016.

[10] Cameron Martino*, Liat Shenhav*, Clarisse Marotz, George Armstrong, Daniel McDonald, Yoshiki Vázquez-Baeza, James T. Morton, Lingjing Jiang, Maria Gloria Dominguez-Bello, Austin D. Swafford, Eran Halperin, and Rob Knight. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. https://doi.org/10.24433/CO.5938114.v1, 2020.

[11] Edward C. Deechan, Chen Yang, Maria Elisa Perez-Muñoz, Nguyen K. Nguyen, Christopher C. Cheng, Lucila Triador, Zhengxiao Zhang, Jeffrey A. Bakal, and Jens Walter. Precision Microbiome Modulation with Discrete Dietary Fiber Structures Directs Short-Chain Fatty Acid Production. *Cell Host and Microbe*, 27(3):389–404.e6, mar 2020.

[12] Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J Brislawn, C Titus Brown, Benjamin J Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily Cope, Ricardo Da Silva, Pieter C Dorrestein, Gavin M Douglas, Daniel M Durall, Claire Duvallet, Christian F Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M Gauglitz, Deanna L Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiaron Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin Huttley, Stefan Janssen, Alan K Jarmusch, Lingjing Jiang, Benjamin Kaehler, Kyo Bin Kang, Christopher R Keefe, Paul Keim, Scott T Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan GI Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erika Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan Martin, Daniel McDonald, Lauren J McIver, Alexey V Melnik, Jessica L Metcalf, Sydney C Morgan, Janie Morton, Ahmad Taran Naimey, Jose A Navas-Molina, Louis Felix Nothias, Stephanie B Orchanian, Talima Pearson, Samuel L Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam
Rivers, Michael S Robeson, II, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R Spear, Austin D Swafford, Luke R Thompson, Pedro J Torres, Pauline Trinh, Anupriya Tripathi, Peter J Turnbaugh, Sabah Ul-Hasan, Justin JJ van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C Weber, Chase HD Williamson, Amy D Willis, Zhenjiang Zech Xu, Jesse R Zaneveld, Yilong Zhang, Rob Knight, and J Gregory Caporaso. Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints, 6:e27295v1, oct 2018.

[13] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. DADA2: High-resolution sample inference from illumina amplicon data. Nature Methods, 13(7):581–583, May 2016.

[14] Melanie Schirmer, Lee Denson, Hera Vlamakis, Eric A. Franzosa, Sonia Thomas, Nathan M. Gotman, Paul Rufo, Susan S. Baker, Cary Sauer, James Markowitz, Marian Pfefferkorn, Maria Oliva-Hemker, Joel Rosh, Anthony Otley, Brendan Boyle, David Mack, Robert Baldassano, David Keljo, Neal LeLeiko, Melvin Heyman, Anne Griffiths, Ashish S. Patel, Joshua Noe, Subra Kugathasan, Thomas Walters, Curtis Huttenhower, Jeffrey Hyams, and Ramnik J. Xavier. Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis Patients Are Linked to Disease Course. Cell Host & Microbe, 24(4):600–610.e4, oct 2018.