

Supplementary Note 1

This note proves two propositions regarding the marginal mean tensor.

Preliminaries: Given a data tensor $X \in \mathbb{R}^{T \times N \times C}$, let $\mathbf{x} = \text{vec}(X) \in \mathbb{R}^{TNC}$ be the vectorized tensor ($\text{vec}(\cdot)$ is vectorization in the order from the first to the last tensor mode). Define the matrix $H_T = \frac{1}{NC}(\mathbf{1}_C \otimes \mathbf{1}_N \otimes I_T) \in \mathbb{R}^{TNC \times T}$. H_T computes the marginal mean of the tensor X along the temporal mode (i.e., $\mu_T = H_T^\top \mathbf{x} \in \mathbb{R}^T$). Similarly, define the matrices $H_N = \frac{1}{TC}(\mathbf{1}_C \otimes I_N \otimes \mathbf{1}_T) \in \mathbb{R}^{TNC \times N}$ and $H_C = \frac{1}{TN}(I_C \otimes \mathbf{1}_N \otimes \mathbf{1}_T) \in \mathbb{R}^{TNC \times C}$, which map \mathbf{x} to its other marginal means. Throughout, \otimes is the Kronecker product, I_D is the identity matrix of size $D \times D$, and $\mathbf{1}_D$ is the ones vector of size D .

We define a marginal mean tensor $M \in \mathbb{R}^{T \times N \times C}$ as any tensor that, when subtracted from the data X , results in a tensor with zero marginal means; that is, $\bar{X} = X - M$ has $H_T^\top \bar{\mathbf{x}} = 0$, $H_N^\top \bar{\mathbf{x}} = 0$, $H_C^\top \bar{\mathbf{x}} = 0$ or equivalently $\mathbf{m} = \text{vec}(M)$ has $H_T^\top \mathbf{m} = \mu_T$, etc. The subspace $\mathcal{M} = \{M \in \mathbb{R}^{T \times N \times C} : H_T^\top \mathbf{m} = \mu_T, H_N^\top \mathbf{m} = \mu_N, H_C^\top \mathbf{m} = \mu_C\}$ has dimension $TNC - (T + N + C)$. A procedure for creating a marginal mean tensor is sequential mean subtraction: applying H_T, H_N, H_C in a specified order, say:

$$\begin{aligned}\bar{\mathbf{x}}^{(1)} &= \mathbf{x} - (NCH_T)(H_T^\top \mathbf{x}) \\ \bar{\mathbf{x}}^{(2)} &= \bar{\mathbf{x}}^{(1)} - (TCH_N)(H_N^\top \bar{\mathbf{x}}^{(1)}) \\ \bar{\mathbf{x}} &= \bar{\mathbf{x}}^{(2)} - (TNH_C)(H_C^\top \bar{\mathbf{x}}^{(2)}).\end{aligned}$$

Note that (NCH_T) , (TCH_N) , and (TNH_C) copy the measured marginal means into the appropriate locations in the vectorized tensor. The resulting tensor \bar{X} has zero marginal means with implied marginal mean tensor $\hat{M} = X - \bar{X}$. Note also by construction that

$\hat{M} \in \mathcal{M}$; that is, \hat{M} is a valid marginal mean tensor.

Proposition 1.1: *The marginal mean tensor \hat{M} that results from sequential mean subtraction is invariant to the order in which the marginal means are subtracted.*

Proof: Expand $\bar{\mathbf{x}} = \text{vec}(\bar{X})$ as:

$$\begin{aligned} \bar{\mathbf{x}} &= (I_{TNC} - TNH_C H_C^\top) (I_{TNC} - TCH_N H_N^\top) (I_{TNC} - NCH_T H_T^\top) \mathbf{x} \\ &= \left((I_C \otimes I_N \otimes I_T) - \left(I_C \otimes \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \otimes \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^\top \right) \right) \\ &\quad \left((I_C \otimes I_N \otimes I_T) - \left(\frac{1}{C} \mathbf{1}_C \mathbf{1}_C^\top \otimes I_N \otimes \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^\top \right) \right) \\ &\quad \left((I_C \otimes I_N \otimes I_T) - \left(\frac{1}{C} \mathbf{1}_C \mathbf{1}_C^\top \otimes \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \otimes I_T \right) \right) \mathbf{x} \\ &\stackrel{\text{def}}{=} P_C P_N P_T \mathbf{x}. \end{aligned}$$

The order of mean subtraction is equivalent to the order of matrix multiplication of the mean-centering matrices P_C , P_N , and P_T . To show that that this mean subtraction is order invariant, it is sufficient to show that P_C , P_N , and P_T commute. From the mixed-product property of the Kronecker product, $AC \otimes BD = (A \otimes B)(C \otimes D)$, this commutation can be readily seen by noting that the multiplication of any pair of the matrices P_C , P_N , and P_T only involves the multiplication of the submatrices $(\frac{1}{T} \mathbf{1}_T \mathbf{1}_T^\top)$, $(\frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top)$, and $(\frac{1}{C} \mathbf{1}_C \mathbf{1}_C^\top)$ by themselves or by the (appropriately sized) identity matrix. Every matrix commutes with itself and the identity; thus P_C , P_N , and P_T commute, which completes the proof.

Proposition 1.2: *Sequential mean subtraction produces the least norm marginal mean tensor; that is, $\hat{\mathbf{m}} = \text{vec}(\hat{M}) = \text{argmin}_{\mathbf{m} \in \mathcal{M}} \|\mathbf{m}\|_2^2$.*

Proof: We already have from above that $\hat{\mathbf{m}} \in \mathcal{M}$. The least norm solution is the orthogonal projection of the origin onto the feasible set \mathcal{M} , and thus it is sufficient to show that the difference between $\hat{\mathbf{m}}$ and any other feasible point $\mathbf{m} \in \mathcal{M}$ is orthogonal to $\hat{\mathbf{m}}$, in which case:

$$\begin{aligned}\|\mathbf{m}\|_2^2 &= \|\mathbf{m} - \hat{\mathbf{m}} + \hat{\mathbf{m}}\|_2^2 \\ &= \|\mathbf{m} - \hat{\mathbf{m}}\|_2^2 + \|\hat{\mathbf{m}}\|_2^2 \quad (\text{by orthogonality}) \\ &\geq \|\hat{\mathbf{m}}\|_2^2.\end{aligned}$$

Using the fact $\hat{\mathbf{m}} = \mathbf{x} - \bar{\mathbf{x}} = (I - P_C P_N P_T)\mathbf{x}$ (see proof of Proposition 1.1), we have:

$$\begin{aligned}\hat{\mathbf{m}}^\top (\hat{\mathbf{m}} - \mathbf{m}) &= \hat{\mathbf{m}}^\top \hat{\mathbf{m}} - \hat{\mathbf{m}}^\top \mathbf{m} \\ &= \mathbf{x}^\top (I - P_C P_N P_T)(I - P_C P_N P_T)\mathbf{x} - \mathbf{x}^\top (I - P_C P_N P_T)\mathbf{m} \\ &= \mathbf{x}^\top (I - P_C P_N P_T)\mathbf{x} - \mathbf{x}^\top (I - P_C P_N P_T)\mathbf{m} \\ &= \mathbf{x}^\top (I - P_C P_N P_T)(\mathbf{x} - \mathbf{m}) \\ &= \mathbf{x}^\top (I - P_C P_N P_T)\bar{\mathbf{x}} \\ &= \mathbf{x}^\top (\bar{\mathbf{x}} - \bar{\mathbf{x}}) \\ &= \mathbf{0}.\end{aligned}$$

The third equality is because $(I - P_C P_N P_T)$ is idempotent. Further note that $\bar{\mathbf{x}} = \mathbf{x} - \mathbf{m}$ has zero marginal means because both \mathbf{x} and \mathbf{m} satisfy the mean constraint ($\mathbf{x}, \mathbf{m} \in \mathcal{M}$), and the operation of $P_C P_N P_T$ is equivalent to sequentially subtracting the marginal means of a tensor (see proof of proposition 1.1). As a result, $P_C P_N P_T \bar{\mathbf{x}} = \bar{\mathbf{x}}$ as the subtracted marginal means are equal to zero in this case. Thus, $\hat{\mathbf{m}} \perp (\hat{\mathbf{m}} - \mathbf{m})$ and consequently $\|\hat{\mathbf{m}}\|_2^2 \leq \|\mathbf{m}\|_2^2 \forall \mathbf{m} \in \mathcal{M}$, which completes the proof.

Supplementary Note 2

This note proves a proposition regarding the constraint placed on the readout matrix K in the CFR method, such that the application of K does not distort the marginal means. It also discusses optimization practicalities.

Preliminaries: We have a surrogate dataset $\bar{S}_0 \in \mathbb{R}^{T \times N \times C}$ with zero marginal means (see definitions and preliminaries in Supplementary Note 1), to which we apply the neural readout matrix $K \in \mathbb{R}^{N \times N}$ so that the resulting surrogate $\bar{S} \in \mathbb{R}^{T \times N \times C}$, where $\bar{S}(t, :, c) = K^\top \bar{S}_0(t, :, c)$ for condition $c \in \{1, \dots, C\}$ and time $t \in \{1, \dots, T\}$, will have the correct marginal covariances. Note that, for any K , the resulting surrogate \bar{S} will have zero mean along the neural mode ($\mu_N = \mathbf{0}$), if \bar{S}_0 has zero mean along the neural mode:

$$\begin{aligned}
 \mu_N &= \frac{1}{NC} \sum_{t=1}^T \sum_{c=1}^C \bar{S}(t, :, c) \\
 &= \frac{1}{NC} \sum_{t=1}^T \sum_{c=1}^C K^\top \bar{S}_0(t, :, c) \\
 &= K^\top \frac{1}{NC} \left(\sum_{t=1}^T \sum_{c=1}^C \bar{S}_0(t, :, c) \right) \\
 &= K^\top \mathbf{0}_N \\
 &= \mathbf{0}_N.
 \end{aligned}$$

However, the mean for the other tensor modes can be non-zero even when the mean of \bar{S}_0 is zero along these other modes.

Proposition 2.1: *If the readout K has eigenvector $\mathbf{1}_N$ with a corresponding eigenvalue of zero, the resulting surrogate \bar{S} will maintain the zero marginal means of \bar{S}_0 .*

Proof: Write the tensor \bar{S} in a vector form as $\bar{\mathbf{s}} = (I_C \otimes K^\top \otimes I_T)\bar{\mathbf{s}}_0$, where $\bar{\mathbf{s}}_0 = \text{vec}(\bar{S}_0)$ and $\bar{\mathbf{s}} = \text{vec}(\bar{S})$. Then:

$$\begin{aligned}
\mu_T &= H_T^\top \bar{\mathbf{s}} \\
&= \frac{1}{NC} (\mathbf{1}_C^\top \otimes \mathbf{1}_N^\top \otimes I_T) (I_C \otimes K^\top \otimes I_T) \bar{\mathbf{s}}_0 \\
&= \frac{1}{NC} (\mathbf{1}_C^\top I_C \otimes \mathbf{1}_N^\top K^\top \otimes I_T I_T) \bar{\mathbf{s}}_0 \\
&= \frac{1}{NC} (\mathbf{1}_C^\top \otimes (K \mathbf{1}_N)^\top \otimes I_T) \bar{\mathbf{s}}_0 \\
&= \frac{1}{NC} (\mathbf{1}_C^\top \otimes \mathbf{0} \mathbf{1}_N^\top \otimes I_T) \bar{\mathbf{s}}_0 \\
&= 0 \frac{1}{NC} (\mathbf{1}_C^\top \otimes \mathbf{1}_N^\top \otimes I_T) \bar{\mathbf{s}}_0 \\
&= \mathbf{0}_T.
\end{aligned}$$

Exchanging H_C for H_T , the same result for μ_C is immediate, by the same steps, which completes the proof.

Implementation Note: This eigenvector condition can be imposed on K by right multiplying $(I_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top)$, a special case of the general fact that a zero eigenvalue can be imposed via subtraction of a normalized rank-one outer product:

$$K \left(I - \frac{1}{\|\mathbf{v}\|_2^2} \mathbf{v} \mathbf{v}^\top \right) \mathbf{v} = K \mathbf{v} - \frac{1}{\|\mathbf{v}\|_2^2} K \mathbf{v} \|\mathbf{v}\|_2^2 = \mathbf{0}.$$

This linear projection integrates easily into the optimization by using a projected gradient: instead of optimizing $f(K)$ with gradient steps $\eta \nabla_K f$ (η : step size), we take projected

gradient steps $\eta G_K f = \eta \nabla_{K} f \left(I - \frac{1}{\|\mathbf{v}\|_2^2} \mathbf{v} \mathbf{v}^\top \right)$ to remain in the feasible set:

$$\begin{aligned}
 K_{i+1} &= (K_i - \eta_i \nabla_{K_i} f) \left(I - \frac{1}{\|\mathbf{v}\|_2^2} \mathbf{v} \mathbf{v}^\top \right) \\
 &= K_i \left(I - \frac{1}{\|\mathbf{v}\|_2^2} \mathbf{v} \mathbf{v}^\top \right) - \eta_i \nabla_{K_i} f \left(I - \frac{1}{\|\mathbf{v}\|_2^2} \mathbf{v} \mathbf{v}^\top \right) \\
 &= K_i - \eta_i \nabla_{K_i} f \left(I - \frac{1}{\|\mathbf{v}\|_2^2} \mathbf{v} \mathbf{v}^\top \right) \\
 &= K_i - \eta_i G_{K_i} f.
 \end{aligned}$$

The advantage here is that we do not need to impose the constraint once we have this form of the projected gradient: the projected gradient can be used directly with any unconstrained optimization package, and it will yield solutions that satisfy the constraints if K is initialized properly (i.e., K_0 satisfy the constraint).

Supplementary Note 3

This note provides the main proof of the form of the maximum entropy distribution used in TME. To proceed in full generality with respect to the number of tensor modes, we require additional notation:

\oplus : Kronecker sum (i.e., $A \oplus B = A \otimes I + I \otimes B$).

n : index over tensor modes ($n \in \{1, \dots, N\}$).

\bar{n} : all tensor modes except the n^{th} mode.

D_n : dimensionality of a tensor along the n^{th} mode (number of elements).

$D_{\bar{n}}$: the product of the dimensionalities of all tensor modes except the n^{th} mode.

Z_{D_n} : matrix unfolding of $Z \in \mathbb{R}^{D_1 \times \dots \times D_N}$ along the n^{th} mode, namely $X_{D_n} \in \mathbb{R}^{D_n \times D_{\bar{n}}}$.

$\Sigma^{(n)}$: the marginal $D_n \times D_n$ covariance of X along the n^{th} mode (note by necessity we have switched notation from the preceding, where this matrix was denoted Σ_n ; e.g., Σ_T).

Throughout we will assume without loss of generality that all modal means are 0 (with mean treatment as discussed in the main text).

Theorem 3.1 *Given tensor data $X \in \mathbb{R}^{D_1 \times \dots \times D_N}$ with marginal covariances $\Sigma^{(1)}, \dots, \Sigma^{(N)}$ that factorize as $\Sigma^{(n)} = Q^{(n)} S^{(n)} Q^{(n)\top}$ (svd), the maximum entropy problem*

$$\begin{aligned}
 & \underset{p}{\text{maximize}} && - \int p(Z) \log p(Z) dZ \\
 & \text{subject to} && \int p(Z) dZ = 1 \\
 & && p(Z) \geq 0 \\
 & && E_p(Z_{D_n} Z_{D_n}^\top) = \Sigma^{(n)} \quad \forall n \in \{1, \dots, N\}
 \end{aligned}$$

has solution $\hat{p}(Z)$, a tensor variate probability distribution, with density on $\mathbf{z} = \text{vec}(Z)$:

$$\hat{p}(\mathbf{z}) = \mathcal{N} \left(0, \frac{1}{2} \left(\otimes_{n=1}^N Q^{(n)} \right) \left(\oplus_{n=1}^N \Lambda^{(n)} \right)^{-1} \left(\otimes_{n=1}^N Q^{(n)} \right)^\top \right),$$

where the $\Lambda^{(n)} = \text{diag}\{\lambda_1^{(n)}, \dots, \lambda_{d_n}^{(n)}\}$ are a function of $S^{(n)} = \text{diag}\{s_1^{(n)}, \dots, s_{d_n}^{(n)}\}$, solving the system of equations:

$$\begin{aligned} s_{d_1}^{(1)} &= \frac{1}{2} \sum_{d_2=1}^{D_2} \dots \sum_{d_N=1}^{D_N} \frac{1}{\lambda_{d_1}^{(1)} + \lambda_{d_2}^{(2)} + \dots + \lambda_{d_N}^{(N)}}, \quad \forall d_1 \in \{1, \dots, D_1\} \\ &\vdots \\ s_{d_N}^{(N)} &= \frac{1}{2} \sum_{d_1=1}^{D_1} \dots \sum_{d_{N-1}=1}^{D_{N-1}} \frac{1}{\lambda_{d_1}^{(1)} + \lambda_{d_2}^{(2)} + \dots + \lambda_{d_N}^{(N)}}, \quad \forall d_N \in \{1, \dots, D_N\}. \end{aligned}$$

Proof: Optimization of the Lagrangian in the standard fashion, with Lagrange multiplier matrices $L^{(n)}$, yields the expected exponential family form:

$$\hat{p}(\mathbf{z}) \propto \exp \left\{ - \sum_{n=1}^N \text{tr} \left(L^{(n)\top} Z_{D_n} Z_{D_n}^\top \right) \right\}.$$

Rearranging the exponent and factorizing $L^{(n)} = U^{(n)} \Lambda^{(n)} U^{(n)\top}$ (which must be symmetric to produce a positive definite quadratic form, required for \hat{p} to integrate to 1) produces:

$$\begin{aligned} \sum_{n=1}^N \text{tr} \left(L^{(n)\top} Z_{D_n} Z_{D_n}^\top \right) &= \text{tr} \left(Z_{D_1}^\top L^{(1)} Z_{D_1} \right) + \dots + \text{tr} \left(Z_{D_N}^\top L^{(N)} Z_{D_N} \right) \\ &= \mathbf{z}^\top \left(I_{D_N} \otimes \dots \otimes L^{(1)} \right) \mathbf{z} + \dots + \mathbf{z}^\top \left(L^{(N)} \otimes \dots \otimes I_{D_1} \right) \mathbf{z} \\ &= \mathbf{z}^\top \left(L^{(N)} \oplus \dots \oplus L^{(1)} \right) \mathbf{z} \\ &= \mathbf{z}^\top \left(U^{(N)} \Lambda^{(N)} U^{(N)\top} \oplus \dots \oplus U^{(1)} \Lambda^{(1)} U^{(1)\top} \right) \mathbf{z} \\ &= \mathbf{z}^\top \left(\otimes_{n=1}^N U^{(n)} \right) \left(\oplus_{n=1}^N \Lambda^{(n)} \right) \left(\otimes_{n=1}^N U^{(n)} \right)^\top \mathbf{z}, \end{aligned}$$

where the last line is technically involved; we prove it in Proposition 4.1. This quadratic form makes apparent the anticipated Gaussian $\hat{p}(\mathbf{z}) = \mathcal{N}(0, \Psi)$, with covariance:

$$\Psi = \frac{1}{2} \left(\otimes_{n=1}^N U^{(n)} \right) \left(\oplus_{n=1}^N \Lambda^{(n)} \right)^{-1} \left(\otimes_{n=1}^N U^{(n)} \right)^\top, \quad (1)$$

proving the form of the distribution as stated in the theorem. What remains then is to choose the $U^{(n)}$ and $\Lambda^{(n)}$ such that this distribution has marginal covariances $\Sigma^{(n)} = Q^{(n)} S^{(n)} Q^{(n)\top}$. Without loss of generality (since we can always consider a different tensor mode and vectorize the tensor starting with that mode), we consider the first marginal covariance $\Psi^{(1)}$ of this distribution, which is the sum of the $D_1 \times D_1$ main-diagonal blocks of Ψ (a technical detail proven in Proposition 4.2). To consider the form of these blocks, we write

$$\Psi = \frac{1}{2} (U^{(\bar{1})} \otimes U^{(1)}) \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_{D_{\bar{1}}} \end{bmatrix}^{-1} (U^{(\bar{1})} \otimes U^{(1)})^\top,$$

where $\Lambda_i \in \mathbb{R}^{D_1 \times D_1}$ is the i^{th} diagonal block of the inverse of the (diagonal) eigenvalue matrix $(\oplus_{n=1}^N \Lambda^{(n)})$ (note the critical notational distinction between Λ_i , the blocks of this Kronecker sum matrix, and $\Lambda^{(n)}$, the constituents of the Kronecker sum that arise from the Lagrange multipliers). Proposition 4.3 shows the form of the i th main-diagonal block of Ψ to be $\sum_{d=1}^{D_{\bar{1}}} \left(U_{i,d}^{(\bar{1})} \right)^2 U^{(1)} \Lambda_d^{-1} U^{(1)\top}$, and thus $\Psi^{(1)}$, the sum of these blocks, is:

$$\begin{aligned} \Psi^{(1)} &= \frac{1}{2} \sum_{i=1}^{D_{\bar{1}}} \sum_{d=1}^{D_{\bar{1}}} \left(U_{i,d}^{(\bar{1})} \right)^2 U^{(1)} \Lambda_d^{-1} U^{(1)\top} \\ &= U^{(1)} \left(\frac{1}{2} \sum_{d=1}^{D_{\bar{1}}} \Lambda_d^{-1} \sum_{i=1}^{D_{\bar{1}}} \left(U_{i,d}^{(\bar{1})} \right)^2 \right) U^{(1)\top} \\ &= U^{(1)} \left(\frac{1}{2} \sum_{d=1}^{D_{\bar{1}}} \Lambda_d^{-1} \right) U^{(1)\top}, \end{aligned}$$

where the last line results because $U^{(1)}$ is an orthogonal matrix. This is a key result, in so much as we aimed to set $\Psi^{(n)}$ to $\Sigma^{(n)} = Q^{(n)}S^{(n)}Q^{(n)\top}$ (the constraint); it is now proven that the Lagrange multiplier eigenvectors $U^{(n)}$ are each equal to $Q^{(n)}$, the given eigenvectors of the marginal covariance constraints, which is then substituted into Equation 1 to give the eigenvector form in the theorem statement, completing that piece of the proof.

Further, we also see that the Lagrange multiplier eigenvalues $\Lambda^{(n)}$ are the solutions to $S^{(n)} = \frac{1}{2} \sum_{d=1}^{D_n} \Lambda_d^{-1}$, where $S^{(n)}$ are the given eigenvalues of the marginal covariance constraints. Explicitly, the expression for the d_n th element of the n th constraint eigenvalue (a form which is detailed in Proposition 4.4) is, for all $d_n \in \{1, \dots, D_n\}$:

$$s_{d_n}^{(n)} = \frac{1}{2} \sum_{d_1=1}^{D_1} \cdots \sum_{d_{n-1}=1}^{D_{n-1}} \sum_{d_{n+1}=1}^{D_{n+1}} \cdots \sum_{d_N=1}^{D_N} \frac{1}{\lambda_{d_1}^{(1)} + \cdots + \lambda_{d_{n-1}}^{(n-1)} + \lambda_{d_n}^{(n)} + \lambda_{d_{n+1}}^{(n+1)} + \cdots + \lambda_{d_N}^{(N)}},$$

which is the eigenvalue form in the theorem statement, thus completing the proof.

Optimization Note: The above system of equations has no closed form solution, but there is a bijection between the set of $s_{d_n}^{(n)}$ eigenvalues (given) and the set of $\lambda_{d_n}^{(n)}$ eigenvalues (unknown). Accordingly we numerically optimize the squared error objective over these $D_1 + \dots + D_N$ variables:

$$\begin{aligned} \min_{\lambda_1^{(1)} \dots \lambda_{D_N}^{(N)}} & \sum_{d_1=1}^{D_1} \left(s_{d_1}^{(1)} - \frac{1}{2} \sum_{d_2=1}^{D_2} \cdots \sum_{d_N=1}^{D_N} \frac{1}{\lambda_{d_1}^{(1)} + \lambda_{d_2}^{(2)} + \cdots + \lambda_{d_N}^{(N)}} \right)^2 + \\ & \cdots + \sum_{d_N=1}^{D_N} \left(s_{d_N}^{(N)} - \frac{1}{2} \sum_{d_1=1}^{D_1} \cdots \sum_{d_{N-1}=1}^{D_{N-1}} \frac{1}{\lambda_{d_1}^{(1)} + \lambda_{d_2}^{(2)} + \cdots + \lambda_{d_N}^{(N)}} \right)^2. \end{aligned}$$

This objective is fast to compute and can be readily differentiated. When the $S^{(n)}$ matrices are well conditioned, optimizing this objective converges quickly to the global optimum

(namely 0, up to machine precision). When the given $S^{(n)}$ are poorly conditioned (i.e., close to low rank), we found that optimization can be sped up significantly by instead defining the objective as the squared loss of the log eigenvalues, namely:

$$\min_{\nu_1^{(1)}, \dots, \nu_{D_N}^{(N)}} \sum_{d_1=1}^{D_1} \left(\log s_{d_1}^{(1)} - \log \sum_{d_2=1}^{D_2} \dots \sum_{d_N=1}^{D_N} \frac{0.5}{e^{\nu_{d_1}^{(1)}} + e^{\nu_{d_2}^{(2)}} + \dots + e^{\nu_{d_N}^{(N)}}} \right)^2 +$$

$$\dots + \sum_{d_N=1}^{D_N} \left(\log s_{d_N}^{(N)} - \log \sum_{d_1=1}^{D_1} \dots \sum_{d_{N-1}=1}^{D_{N-1}} \frac{0.5}{e^{\nu_{d_1}^{(1)}} + e^{\nu_{d_2}^{(2)}} + \dots + e^{\nu_{d_N}^{(N)}}} \right)^2,$$

which optimizes quickly to machine precision in all situations we have tested.

Supplementary Note 4

This note provides supporting technical proofs that are necessary for the main proof that precedes (Supplementary Note 3).

Proposition 4.1: $L^{(N)} \oplus \dots \oplus L^{(1)} = (U^{(N)} \otimes \dots \otimes U^{(1)})(\Lambda^{(N)} \oplus \dots \oplus \Lambda^{(1)})(V^{(N)} \otimes \dots \otimes V^{(1)})^\top$ where $L^{(n)} = U^{(n)}\Lambda^{(n)}V^{(n)\top}$ (svd) for $n \in \{1, \dots, N\}$.

Proof: We will leverage the following known linear algebraic properties:

- property 1: $\exp(A \oplus B) = \exp(A) \otimes \exp(B)$
- property 2: $\exp(Z) = U_Z \exp(S_Z) V_Z^\top$ where $Z = U_Z S_Z V_Z^\top$
- property 3: $AC \otimes BD = (A \otimes B)(C \otimes D)$

Then consider the matrix exponential:

$$\begin{aligned}
 & \exp(L^{(N)} \oplus \dots \oplus L^{(1)}) \\
 &= \exp(L^{(N)}) \otimes \dots \otimes \exp(L^{(1)}) \quad (\text{property 1}) \\
 &= U^{(N)} \exp(\Lambda^{(N)}) V^{(N)\top} \otimes \dots \otimes U^{(1)} \exp(\Lambda^{(1)}) V^{(1)\top} \quad (\text{property 2}) \\
 &= (U^{(N)} \otimes \dots \otimes U^{(1)}) (\exp(\Lambda^{(N)}) \otimes \dots \otimes \exp(\Lambda^{(1)})) (V^{(N)} \otimes \dots \otimes V^{(1)})^\top \quad (\text{property 3}) \\
 &= (U^{(N)} \otimes \dots \otimes U^{(1)}) \exp(\Lambda^{(N)} \oplus \dots \oplus \Lambda^{(1)}) (V^{(N)} \otimes \dots \otimes V^{(1)})^\top. \quad (\text{property 1})
 \end{aligned}$$

From property 2, the left singular vectors, right singular vectors, and singular values of $(L^{(N)} \oplus \dots \oplus L^{(1)})$ are equal to $(U^{(N)} \otimes \dots \otimes U^{(1)})$, $(V^{(N)} \otimes \dots \otimes V^{(1)})$, and $(\Lambda^{(N)} \oplus \dots \oplus \Lambda^{(1)})$, respectively, which completes the proof.

Proposition 4.2: Let $Z \in \mathbb{R}^{D_1 \times \dots \times D_N}$ and $\Psi = E(\mathbf{z}\mathbf{z}^\top) \in \mathbb{R}^{D_1 \dots D_N \times D_1 \dots D_N}$ for $\mathbf{z} = \text{vec}(Z)$. Then $\Psi^{(1)} = E(Z_{D_1} Z_{D_1}^\top) \in \mathbb{R}^{D_1 \times D_1}$ is the sum of the $D_1 \times D_1$ diagonal blocks of Ψ .

Proof: Denote the d^{th} column of Z_{D_1} by $\mathbf{v}_d \in \mathbb{R}^{D_1}$ for $d \in \{1, \dots, D_1\}$, then $\mathbf{z} = [\mathbf{v}_1^\top, \dots, \mathbf{v}_{D_1}^\top]^\top$. We then write:

$$\begin{aligned} \Psi^{(1)} &= E(Z_{D_1} Z_{D_1}^\top) \\ &= E\left(\begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_{D_1} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_{D_1}^\top \end{bmatrix}\right) \\ &= \sum_{d=1}^{D_1} E(\mathbf{v}_d \mathbf{v}_d^\top), \end{aligned}$$

the summands of which each correspond to a $D_1 \times D_1$ diagonal block of Ψ , which thus completes the proof.

Proposition 4.3: Given a matrix Ψ with singular value decomposition

$$\Psi = (U^B \otimes U^A) \begin{bmatrix} \Phi_1 & & & \\ & \Phi_2 & & \\ & & \dots & \\ & & & \Phi_{D_B} \end{bmatrix} (U^B \otimes U^A)^\top,$$

where the Φ_i are the D_B main-diagonal blocks of size $D_A \times D_A$, the i th diagonal block of Ψ has the form $U^A \left(\sum_{d=1}^{D_B} (U_{i,d}^B)^2 \Phi_d \right) U^A^\top$.

Proof: Considering first the left multiplication of $(U^B \otimes U^A)$,

$$\begin{aligned} (U^B \otimes U^A) \begin{bmatrix} \Phi_1 & & & \\ & \Phi_2 & & \\ & & \ddots & \\ & & & \Phi_{D_B} \end{bmatrix} &= \begin{bmatrix} U_{1,1}^B U^A & \dots & U_{1,D_B}^B U^A \\ \vdots & \ddots & \vdots \\ U_{D_B,1}^B U^A & \dots & U_{D_B,D_B}^B U^A \end{bmatrix} \begin{bmatrix} \Phi_1 & & & \\ & \Phi_2 & & \\ & & \ddots & \\ & & & \Phi_{D_B} \end{bmatrix} \\ &= \begin{bmatrix} U_{1,1}^B U^A \Phi_1 & \dots & U_{1,D_B}^B U^A \Phi_{D_B} \\ \vdots & \ddots & \vdots \\ U_{D_B,1}^B U^A \Phi_1 & \dots & U_{D_B,D_B}^B U^A \Phi_{D_B} \end{bmatrix}. \end{aligned}$$

By repeating the steps for the right multiplication of $(U^B \otimes U^A)^\top$, the i th block along the main diagonal is seen to be $\sum_{d=1}^{D_B} (U_{i,d}^B)^2 U^A \Phi_d U^A^\top$, which completes the proof.

Proposition 4.4: Let $\Psi^{(1)} = \frac{1}{2} \sum_{d=1}^{D_1} \Lambda_d^{-1} \in \mathbb{R}^{D_1 \times D_1}$, where Λ_d is the d^{th} main-diagonal block of $\Lambda = \Lambda^{(N)} \oplus \dots \oplus \Lambda^{(1)}$, and $\Lambda^{(n)}$ is a diagonal matrix with D_n elements. Then $\psi_{d_1}^{(1)}$, the d_1 th diagonal element of $\Psi^{(1)}$, is equal to $\frac{1}{2} \sum_{d_N=1}^{D_N} \dots \sum_{d_2=1}^{D_2} \frac{1}{\lambda_{d_1}^{(1)} + \lambda_{d_2}^{(2)} + \dots + \lambda_{d_N}^{(N)}}$.

Proof: Note that the Kronecker sum of diagonal matrices $\Lambda^{(N)}, \dots, \Lambda^{(1)}$ (i.e., Λ) can be seen as a counting system where the most to least significant digit goes from elements of $\Lambda^{(N)}$ to the elements of $\Lambda^{(1)}$, and those elements are then added to form the entry in the matrix. Λ^{-1} then simply inverts each entry. For example, if $N = 2$, then:

$$(\Lambda^{(2)} \oplus \Lambda^{(1)})^{-1} = \begin{bmatrix} \left[\begin{array}{ccc} \frac{1}{\lambda_1^{(1)} + \lambda_1^{(2)}} & & \\ & \ddots & \\ & & \frac{1}{\lambda_{D_1}^{(1)} + \lambda_1^{(2)}} \end{array} \right] & & \\ & \ddots & \\ & & \left[\begin{array}{ccc} \frac{1}{\lambda_1^{(1)} + \lambda_{D_2}^{(2)}} & & \\ & \ddots & \\ & & \frac{1}{\lambda_{D_1}^{(1)} + \lambda_{D_2}^{(2)}} \end{array} \right] \end{bmatrix}.$$

We can see from the above example that the entry in the d_i th position of each of the $D_1 \times D_1$ main-diagonal blocks share the $\lambda_{d_i}^{(1)}$ element in the entry's denominator, and the other elements of that denominator change systematically (in the counting fashion) as we move down along the main-diagonal blocks. Accordingly,

$$\psi_{d_1}^{(1)} = \frac{1}{2} \sum_{d=1}^{D_1} \Lambda_d^{-1} = \frac{1}{2} \sum_{d_N=1}^{D_N} \cdots \sum_{d_2=1}^{D_2} \frac{1}{\lambda_{d_1}^{(1)} + \lambda_{d_2}^{(2)} + \dots + \lambda_{d_N}^{(N)}},$$

which completes the proof. In full generality for the d_n th eigenvalue of $\Psi^{(n)}$, for all $d_n \in \{1, \dots, D_n\}$, the same steps show that

$$\psi_{d_n}^{(n)} = \frac{1}{2} \sum_{d_1=1}^{D_1} \cdots \sum_{d_{n-1}=1}^{D_{n-1}} \sum_{d_{n+1}=1}^{D_{n+1}} \cdots \sum_{d_N=1}^{D_N} \frac{1}{\lambda_{d_1}^{(1)} + \dots + \lambda_{d_{n-1}}^{(n-1)} + \lambda_{d_n}^{(n)} + \lambda_{d_{n+1}}^{(n+1)} + \dots + \lambda_{d_N}^{(N)}}.$$