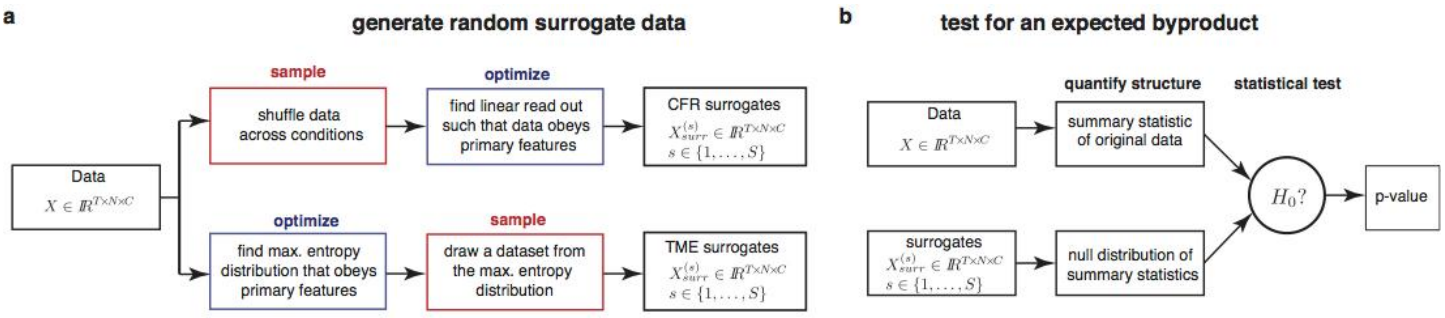
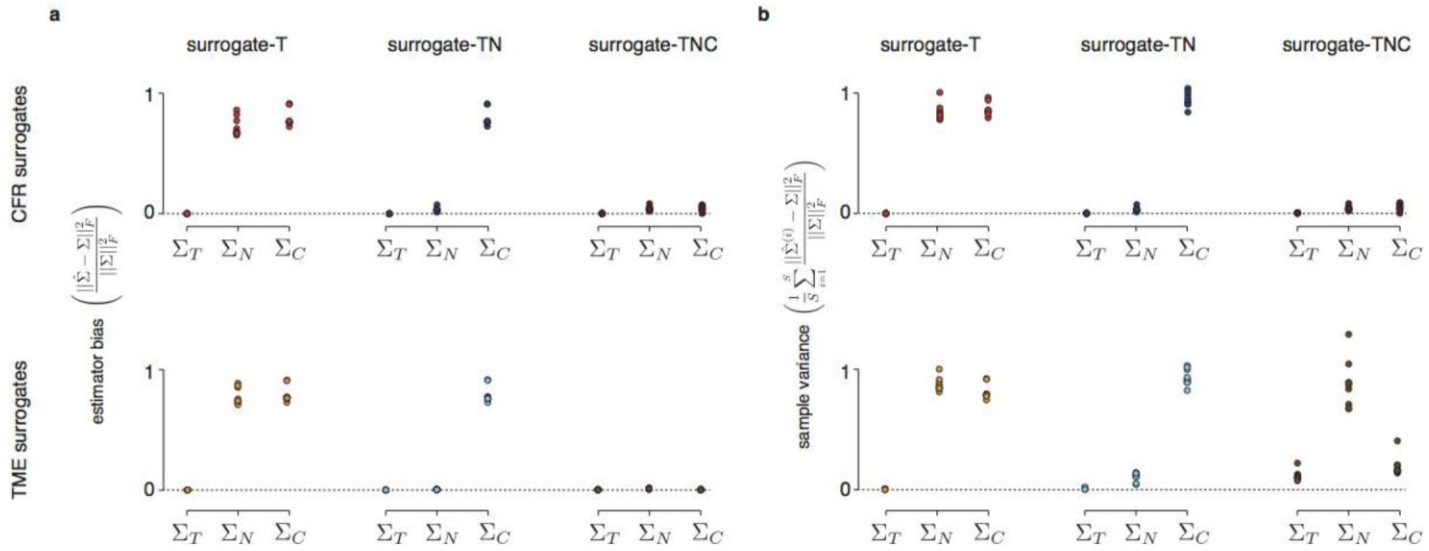


Supplementary Figure 1

Corrected Fisher randomization (CFR) and tensor maximum entropy (TME) methods.

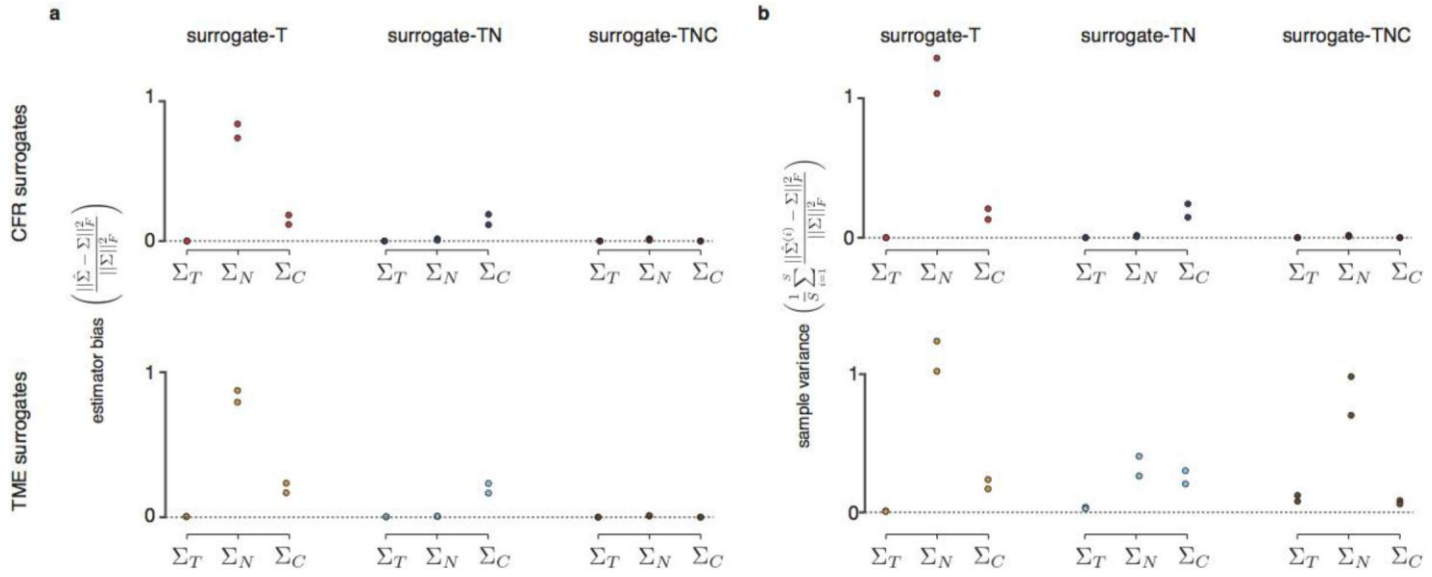
(**a**) Creating *S* surrogate datasets using CFR (top) and TME methods (bottom). Sample CFR: original neural responses are shuffled across conditions, yielding a randomized dataset with distorted primary features. Optimize CFR: find the best neural readout that operates on the shuffled data to retain the primary features of the original dataset. Optimize TME: find the maximum entropy distribution of tensor-valued datasets that has the same primary features (first and second marginal moments) as the original dataset. Sample TME: use efficient Kronecker methods to sample from the maximum entropy distribution. (**b**) Using surrogate datasets from CFR or TME to test the null hypothesis that population structure is an expected byproduct of given primary features. Quantify structure: evaluate a summary statistic such as R^2 or variance explained that quantifies the degree that a hypothesized population structure exists in population data. Evaluate the summary statistic from the original neural data and from *S* surrogate datasets. Use the *S* values of the summary statistics from the surrogate datasets as the null distribution of population structure arising from the primary features alone. The summary statistic from the original neural data is then compared to that null distribution to obtain a p-value for the null hypothesis.



Supplementary Figure 2

Quantification of primary features in the surrogate datasets based on motor cortex data.

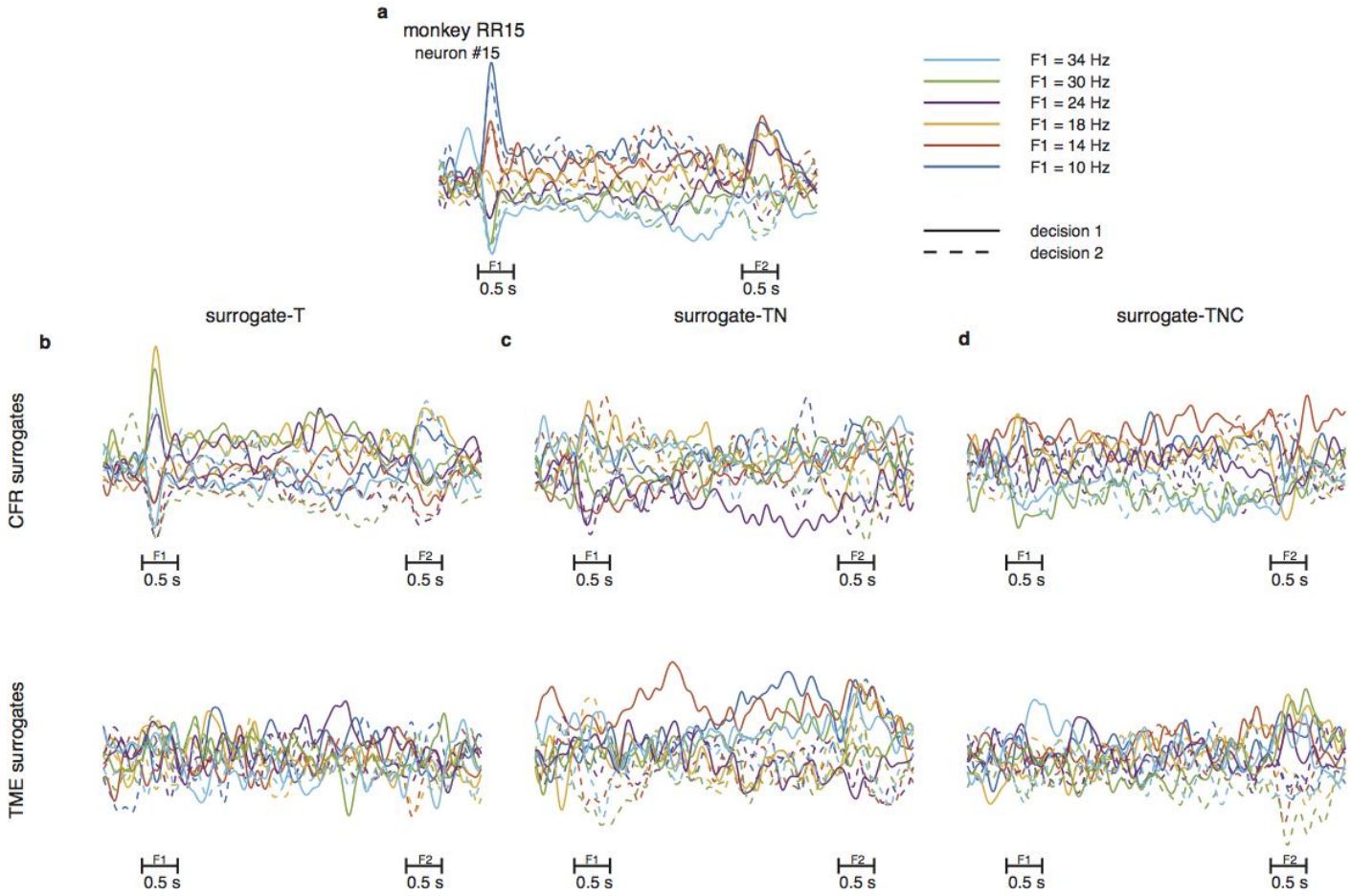
Each surrogate dataset $X_{surr}^{(i)}$ (for $i \in \{1, \ldots, S\}$, S=100) has marginal covariance $\Sigma_T^{(i)}$, $\Sigma_N^{(i)}$, and $\Sigma_C^{(i)}$, which, in the surrogate-TNC control (right column of each panel), should match the specified primary features Σ_T, Σ_N, and Σ_C of the original neural data. The right column demonstrates that fit: the top two rows (panel **a**) of the right column shows that CFR and TME match the true covariances in expectation (see equation on vertical axis); the bottom two rows (panel **b**) show that CFR has very minor variance around that mean, whereas TME has meaningful variance (see equation on vertical axis), as expected. Each dot corresponds to one of the nine separate datasets collected in the motor cortex (A, B, J1, J2, J3, J4, J-Array, N, and N-Array; see Methods for data details). The left and middle columns of each panel of the figure correspond to the surrogate-T (left) and surrogate-TN (middle) controls. Here we see as expected that all covariances that are specified in the control Σ_T in the left column; both Σ_T and Σ_N in the middle column) are very well matched. Correspondingly, we also see that covariances that are not specified in the control do not match the moments of the data. Taken together, these data demonstrate quantitatively that the CFR and TME methods behave as desired, both in terms of preserving the specified structure, and in terms of destroying structure not specified.



Supplementary Figure 3

Quantification of primary features in the surrogate datasets based on PFC data.

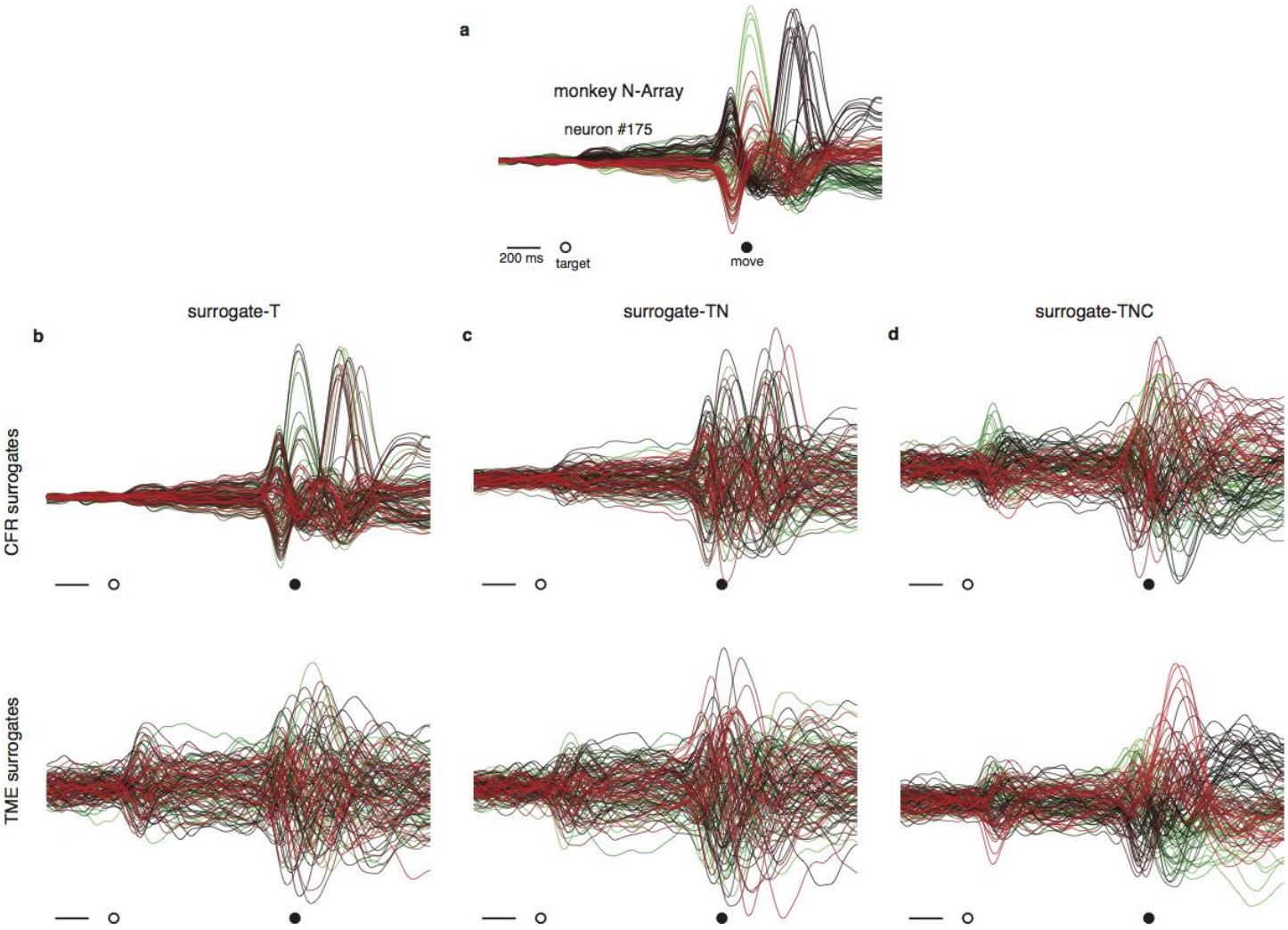
Each surrogate dataset $X_{surr}^{(i)}$ (for $i \in \{1, \ldots, S\}$, S=100) has marginal covariances $\Sigma_T^{(i)}$, $\Sigma_N^{(i)}$, and $\Sigma_C^{(i)}$, which, in the surrogate-TNC control (right column of each panel), should match the specified primary features Σ_T, Σ_N, and Σ_C of the original neural data. The right column demonstrates that fit: the top two rows (panel **a**) of the right column shows that CFR and TME match the true covariances in expectation (see equation on vertical axis); the bottom two rows (panel **b**) show that CFR has very minor variance around that mean, whereas TME has meaningful variance (again see equation on vertical axis), as expected. Each dot corresponds to one of the two datasets collected in the prefrontal cortex (RR15 and RR14; see Methods for data details). The left and middle columns of each panel of the figure correspond to the surrogate-T (left) and surrogate-TN (middle) controls. Here we see as expected that all covariances that are specified in the control (Σ_T in the left column; both Σ_T and Σ_N in the middle column) are very well matched. Correspondingly, we also see that covariances that are not specified in the control do not match the moments of the data. Taken together, these data demonstrate quantitatively that the CFR and TME methods behave as desired, both in terms of preserving the specified structure, and in terms of destroying structure not specified.



Supplementary Figure 4

Neural responses in prefrontal cortex vs. surrogates.

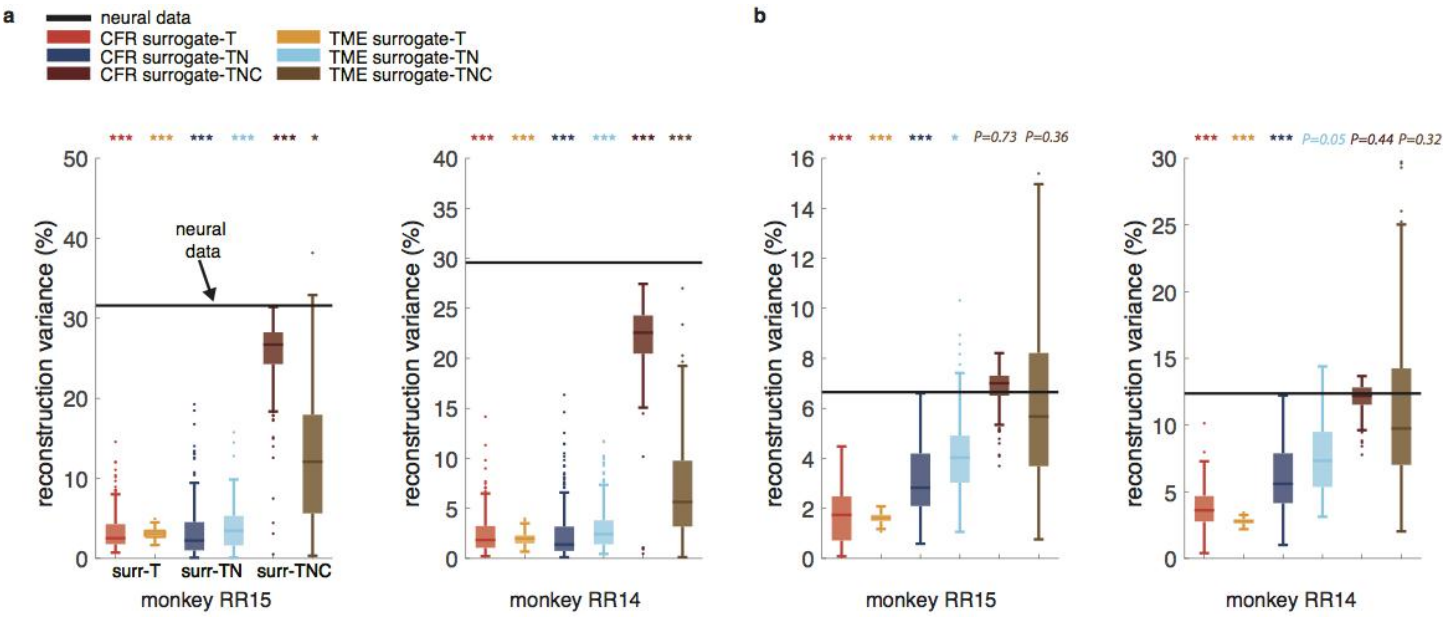
This figure is similar to **Fig. 3** from the main text but showing preprocessed data with soft-normalization and mean-condition subtraction. (**a**) Example neuron (neuron number 15 of 571 total) from prefrontal cortex. Each trace is the trial-averaged firing of the twelve task conditions (six stimuli and two decisions). The traces color and style reflect the twelve possible experimental conditions. Horizontal bars denote the times of first (F1) and second (F2) vibrotactile stimuli. (**b-d**) Example neurons from one surrogate-T, surrogate-TN, and surrogate-TNC dataset, respectively. **b-d** panels follow the same convention as panel **a**. Top panels in **b-d** are surrogate datasets generated using Corrected Fisher Randomization (CFR) and bottom panels in **b-d** are surrogate datasets generated using Tensor Maximum Entropy (TME).



Supplementary Figure 5

Neural responses in motor cortex vs. surrogates.

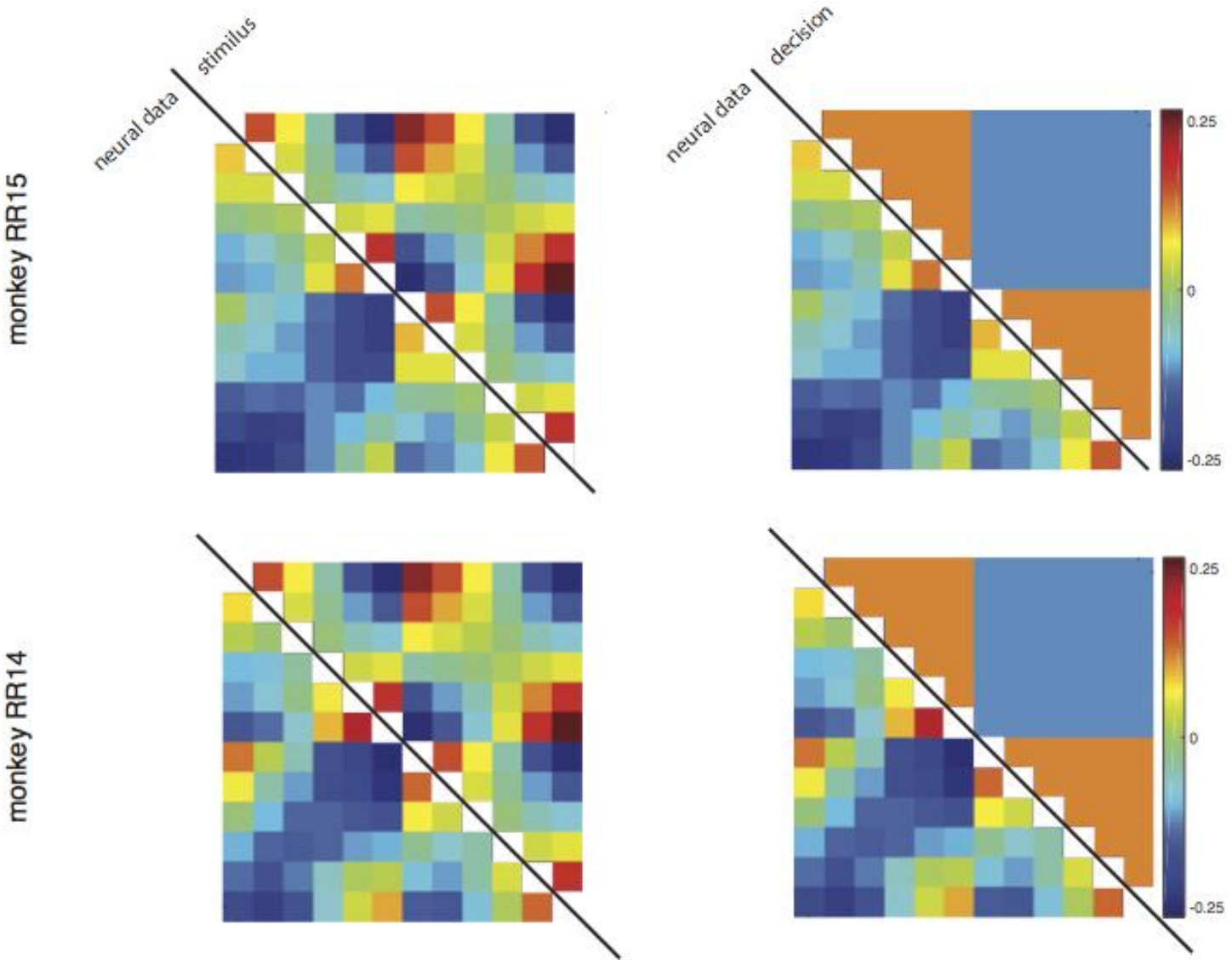
This figure is showing preprocessed data with soft-normalization and mean-condition subtraction. (**a**) Example neuron (neuron number 175 of 218 total) recorded from the motor cortex of one monkey during the delayed-reach task. Each trace is the trial-averaged normalized rate during one of 108 reaching conditions (neuron 175 from monkey N-Array; see Methods). The trace color indicates the reach condition from **Fig. 5a** in the main text. (**b-d**) Example neurons from one surrogate-T, surrogate-TN, and surrogate-TNC dataset, respectively. **b-d** panels follow the same convention as panel **a**. Top panels in **b-d** are surrogate datasets generated using Corrected Fisher Randomization (CFR) and bottom panels in **b-d** are surrogate datasets generated using Tensor Maximum Entropy (TME). Scale bars and time markers in **b-d** match the scale bar and time markers in **a**.



Supplementary Figure 6

Decision (or stimulus) readouts in PFC are not (or are) an expected byproduct.

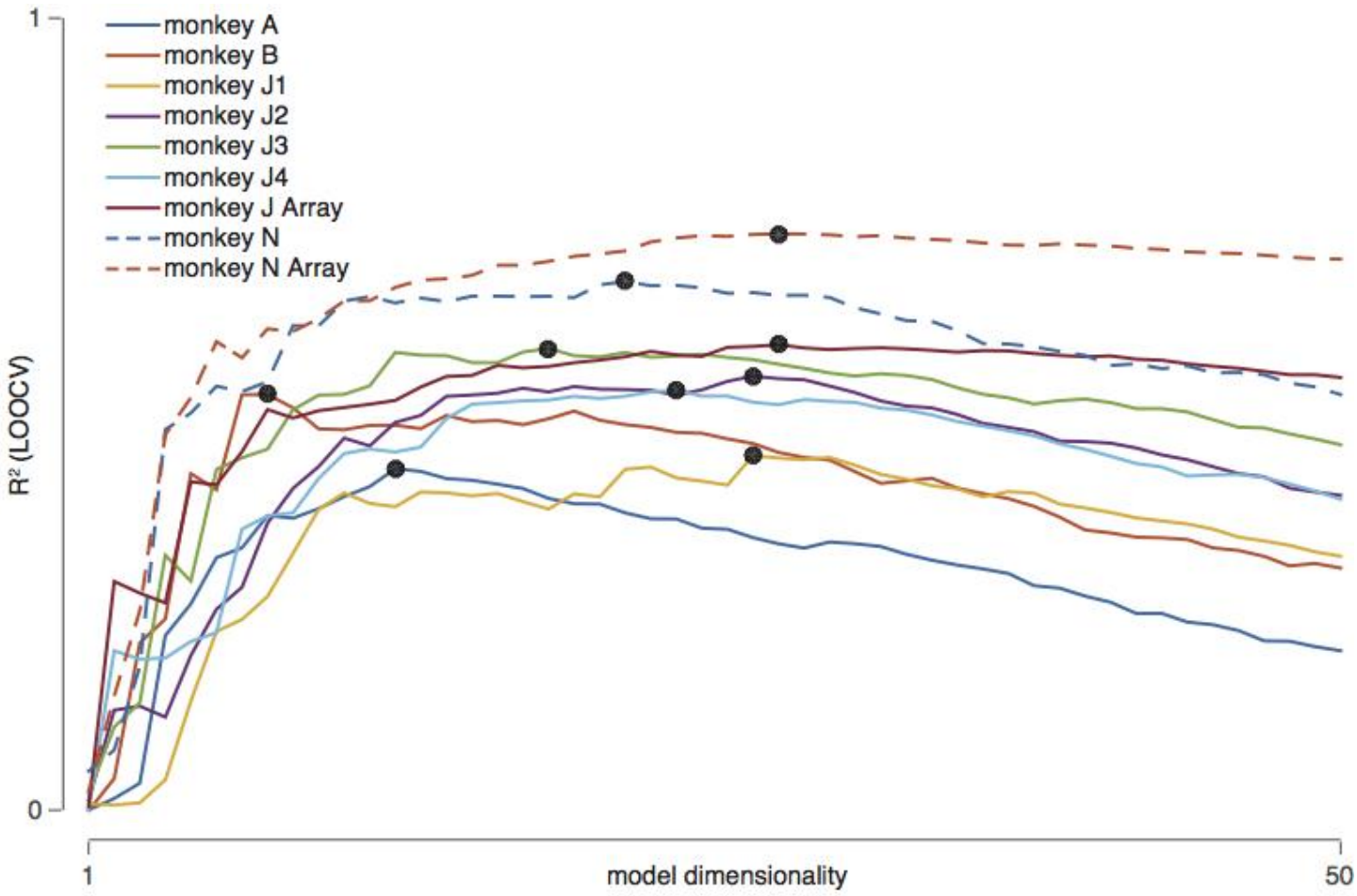
This figure is similar to **Fig. 4e,f** but with reconstruction variance (termed explained variance in Kobak *et al.*[37]) instead of the conventional percentage variance explained. (**a**) Percent reconstruction variance of the population projection onto the top decision readout. Black lines show the percent variance explained from the original neural data, colored box-whisker plots show the variance explained distribution from 200 surrogate samples (same convention as **Fig. 4 e,f**; stars denote significantly high variance; upper-tail test). The variance of the decision projection is calculated during the decision epoch (100 ms after the second stimulus onset, until the second stimulus offset). (**b**) Same as **a** but for reconstruction variance of the population projection onto the top stimulus readout. The variance of the stimulus projection is calculated during the stimulus epoch (100 ms after first stimulus onset, until the second stimulus onset).



Supplementary Figure 7

Single-neuron tuning to stimulus is prominent in PFC.

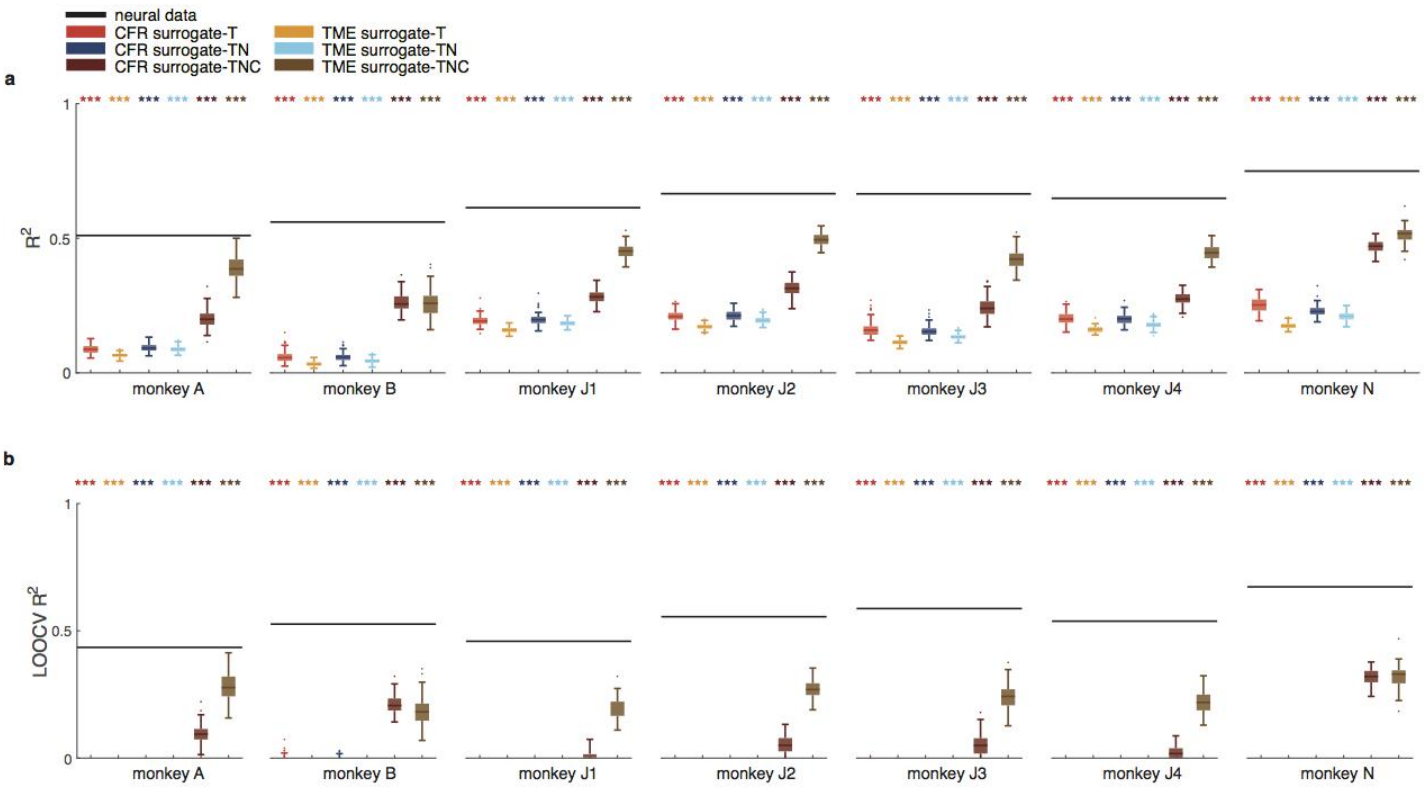
Left column: the lower triangle is the unique values of the condition covariance matrix (lower triangle of Σ_C) of the original neural data (normalized to have unit norm); the upper triangle is the unique values of the outer product of the stimulus vector with itself (again normalized to have unit norm, shown above the diagonal line). Right column: lower triangle is the same as the left panel but the upper triangle is now the unique values of the outer product of the decision vector (right). Top row is for monkey RR15 and bottom row is for monkey RR14 (see Methods for data details). The left column has a closer qualitative match between upper and lower triangles, indicating stimulus tuning is similar to the condition covariance matrix. Compare to the right column, which looks highly non-symmetric, meaning that Σ_C does not capture decision tuning well.



Supplementary Figure 8

Dimensionality of the linear dynamical system structure in motor cortex.

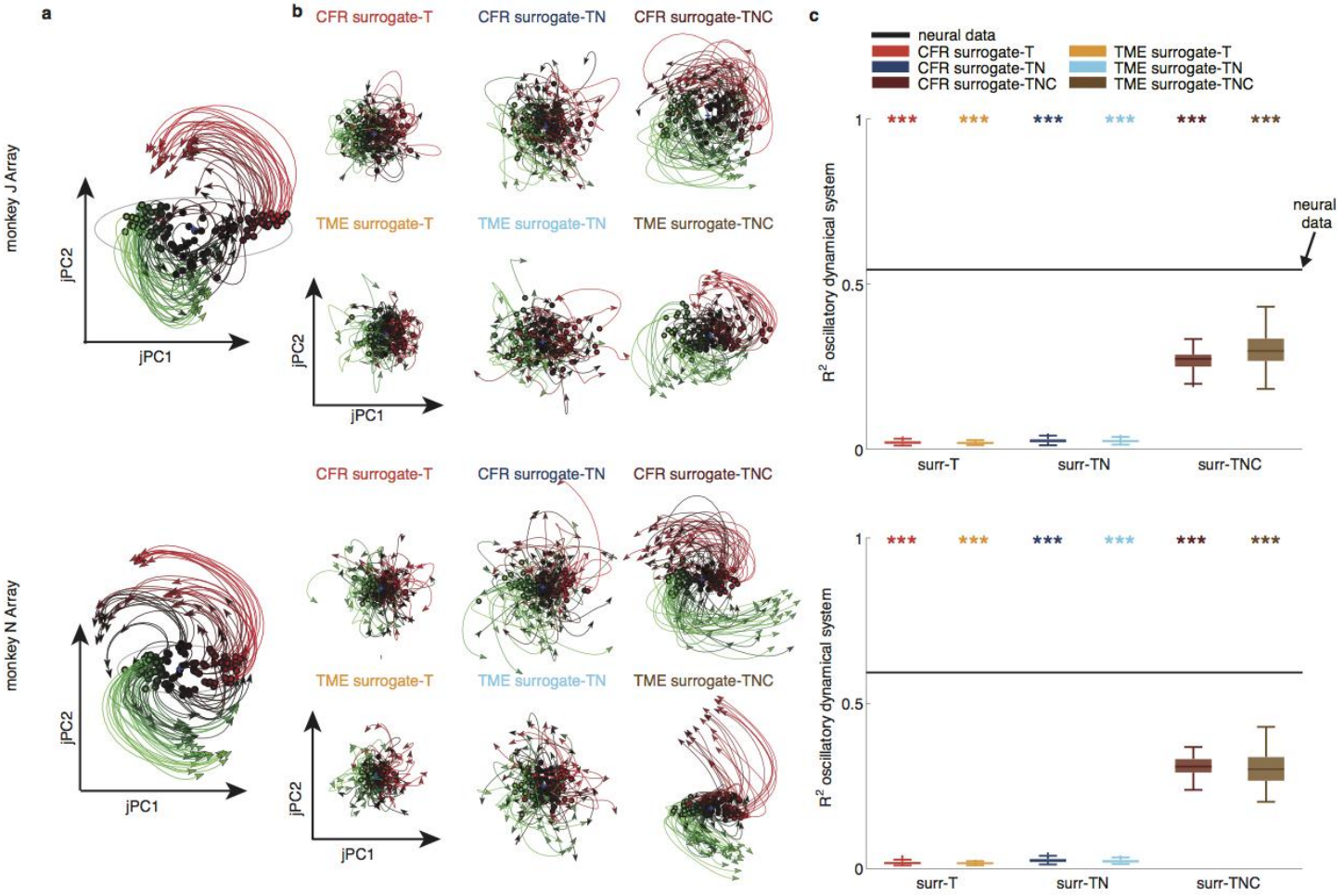
We measured the fit quality (R^2) of the motor cortex data to a linear dynamical system with different dimensionalities using leave-one-condition-out cross-validation. Model dimensionalities in the horizontal axis reflect the number of principal components retained in the population response before fitting the dynamical system. Colored traces show the R^2 from different datasets from the motor cortex (nine datasets). Black dot is the highest cross-validated R^2 value, which is used in **Fig. 6a** and **Supplementary Fig. 9.** Note that all test results are robust to this choice of dimensionality over a wide range (all but the smallest dimensionalities; see **Fig. 6b**).



Supplementary Figure 9

Population dynamics in motor cortex from many datasets are consistently significant.

Neural responses in the motor cortex from the 400 ms duration reflecting movement response were projected onto the top PCs (see **Supplementary Fig. 8**), and then fit to a linear dynamical system. (**a**) Quality of fit (R^2) of the original neural responses to the dynamical system model. Black bar denotes the R^2 from the original neural data. Colored bars denote the median R^2 from 100 different surrogates from each control type (error bars denote the 95th percentile of the distribution). Stars denote significantly higher R^2 value in the original data than in the surrogates (P<0.001; upper-tail test). (**b**) Same as **a**, but using leave-one-condition-out cross-validated R^2.



Supplementary Figure 10

Quasirhythmic population response in motor cortex is not an expected byproduct.

Churchland *et al.*[9] identified a quasi-rhythmic population response during reaching from motor cortex data. They designed a dimensionality reduction method (jPCA; Churchland *et al.*[9]) to identify a population projection, which revealed the presence or absence of rotational dynamics in the neural trajectories. Some have expressed concern that the observed population oscillations may be an expected byproduct of the primary features of neural data. (**a**) Projections of population responses onto the top jPCA plane. Each trace shows neural trajectories reflecting the first 200 ms of movement-related activity for one of 108 reaching conditions. The colors of the traces are based on the variance of the preparatory state along jPC1. (**b**) jPCA projections from one sample of each surrogate type. (**c**) Quality of fit (R^2) of the jPCA oscillatory dynamical system model of the original neural data and the surrogate datasets (same convention as **Fig. 6a**). This figure validates the claim that quasi-rhythmic dynamics are not an expected byproduct.

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 $\mathbf{x} = \mathbf{vec}(X) \in \mathbb{R}^{TNC}$ be the vectorized tensor ($\mathbf{vec}(.)$ 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 $H_N = \frac{1}{TC}(1_C \otimes I_N \otimes 1_T) \in \mathbb{R}^{TNC\times N}$ and $H_C = \frac{1}{TN}(I_C \otimes 1_N \otimes 1_T) \in \mathbb{R}^{TNC\times C}$, 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 $\mathbf{m} = \mathbf{vec}(M)$ 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:

$$\bar{\mathbf{x}}^{(1)} = \mathbf{x} - (NCH_T)\left({H_T}^\top \mathbf{x}\right)$$
$$\bar{\mathbf{x}}^{(2)} = \bar{\mathbf{x}}^{(1)} - (TCH_N)\left({H_N}^\top \bar{\mathbf{x}}^{(1)}\right)$$
$$\bar{\mathbf{x}} = \bar{\mathbf{x}}^{(2)} - (TNH_C)\left({H_C}^\top \bar{\mathbf{x}}^{(2)}\right).$$

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

Nature Neuroscience: doi:10.1038/nn.4617

$\hat{M} \in \mathcal{M}$; 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 $\bar{\mathbf{x}} = \mathbf{vec}(\bar{X})$ as:

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

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,* $\hat{\mathbf{m}} = \mathbf{vec}(\hat{M}) = \mathrm{argmin}_{\mathbf{m} \in \mathcal{M}} \|\mathbf{m}\|_2^2$.

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 $\bar{S}_0 \in I\!R^{T\times N\times C}$ with zero marginal means (see definitions and preliminaries in Supplementary Note 1), to which we apply the neural readout matrix $K \in I\!R^{N\times N}$ so that the resulting surrogate $\bar{S} \in I\!R^{T\times N\times C}$, where $\bar{S}(t,:,c) = K^\top \bar{S}_0(t,:,c)$ for condition $c \in \{1,\ldots,C\}$ and time $t \in \{1,\ldots,T\}$, 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 $\bar{S}_0$.*

Nature Neuroscience: doi:10.1038/nn.4617

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 $\bar{\mathbf{s}}_0 = \mathbf{vec}(\bar{S}_0)$ and $\bar{\mathbf{s}} = \mathbf{vec}(\bar{S})$. Then:

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

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 $\left(I_N - \frac{1}{N}\mathbf{1}_N{\mathbf{1}_N}^\top\right)$, a special case of the general fact that a zero eigenvalue can be imposed via subtraction of a normalized rank-one outer product:

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

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

Nature Neuroscience: doi:10.1038/nn.4617

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).

Nature Neuroscience: doi:10.1038/nn.4617

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 ($n \in \{1, \ldots, N\}$).

$\bar{n}$: 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 $Z \in I\!R^{D_1 \times \ldots \times D_N}$ along the nth mode, namely $X_{D_n} \in I\!R^{D_n \times D_{\bar{n}}}$.

$\Sigma^{(n)}$: the marginal $D_n \times D_n$ covariance of X along the nth 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 $X \in I\!R^{D_1 \times \ldots \times D_N}$ with marginal covariances $\Sigma^{(1)}, \ldots, \Sigma^{(N)}$ that factorize as $\Sigma^{(n)} = Q^{(n)} S^{(n)} Q^{(n)\top}$ (svd), the maximum entropy problem*

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

Nature Neuroscience: doi:10.1038/nn.4617

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

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

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

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

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

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

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:

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

Nature Neuroscience: doi:10.1038/nn.4617

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:

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

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 $\Lambda_i \in I\!R^{D_1 \times D_1}$ is the i^{th} diagonal block of the inverse of the (diagonal) eigenvalue matrix $\left(\oplus_{n=1}^{N} \Lambda^{(n)}\right)$ (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 ith 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 $U^{(\bar{1})}$ 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 $S^{(n)} = \frac{1}{2} \sum_{d=1}^{D_{\bar{n}}} \Lambda_d^{-1}$, where $S^{(n)}$ are the given eigenvalues of the marginal covariance constraints. Explicitly, the expression for the d_nth element of the nth constraint eigenvalue (a form which is detailed in Proposition 4.4) is, for all $d_n \in \{1, ..., D_n\}$:

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

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 $D_1 + ... + D_N$ variables:

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

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:

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

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

Nature Neuroscience: doi:10.1038/nn.4617

Supplementary Note 4

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

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

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:

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

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

Nature Neuroscience: doi:10.1038/nn.4617

Proposition 4.2: *Let* $Z \in \mathbb{R}^{D_1 \times \ldots \times D_N}$ *and* $\Psi = E\left(\mathbf{z}\mathbf{z}^\top\right) \in \mathbb{R}^{D_1 \ldots D_N \times D_1 \ldots D_N}$ *for* $\mathbf{z} = \mathbf{vec}(Z)$. *Then* $\Psi^{(1)} = E\left(Z_{D_1} Z_{D_1}^\top\right) \in \mathbb{R}^{D_1 \times D_1}$ *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 $d \in \{1, \ldots, D_{\bar{1}}\}$, then $\mathbf{z} = [\mathbf{v}_1{}^\top, \ldots, \mathbf{v}_{D_{\bar{1}}}{}^\top]^\top$. We then write:

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

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*

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

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* $U^A \left(\sum_{d=1}^{D_B} \left(U_{i,d}^B\right)^2 \Phi_d\right) U^{A\top}$.

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

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

$$= \begin{bmatrix} U^B_{1,1}U^A\Phi_1 & \dots & U^B_{1,D_B}U^A\Phi_{D_B} \\ \vdots & \ddots & \vdots \\ U^B_{D_B,1}U^A\Phi_1 & \dots & U^B_{D_B,D_B}U^A\Phi_{D_B}\end{bmatrix}.$$

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

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

Proof: Note that the Kronecker sum of diagonal matrices $\Lambda^{(N)}, \ldots, \Lambda^{(1)}$ (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:

$$(\Lambda^{(2)} \oplus \Lambda^{(1)})^{-1} = \begin{bmatrix} \begin{bmatrix}\frac{1}{\lambda^{(1)}_1+\lambda^{(2)}_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda^{(1)}_{D_1}+\lambda^{(2)}_1}\end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix}\frac{1}{\lambda^{(1)}_1+\lambda^{(2)}_{D_2}} & & \\ & \ddots & \\ & & \frac{1}{\lambda^{(1)}_{D_1}+\lambda^{(2)}_{D_2}}\end{bmatrix}\end{bmatrix}.$$

Nature Neuroscience: doi:10.1038/nn.4617

We can see from the above example that the entry in the d_ith 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,

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

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

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

Nature Neuroscience: doi:10.1038/nn.4617