Demixed PCA

Jorge Aurelio Menéndez August 18, 2020

These are personal notes on the demixed PCA method presented in [Kobak et al., 2016].

Contents

1. The data

We'll consider a data set of the following form:

$$\left\{\mathbf{x}_{tsdk}\right\}_{t,s,d,k}$$

where $\mathbf{x}_{tsdk} \in \mathbb{R}^N$ is a vector of recorded neural activity

- ▷ at time bin $t \in \{1, \ldots, T\}$
- ▷ in response to stimulus $s \in \{1, \ldots, S\}$
- ▷ in condition $d \in \{1, \ldots, D\}$ (this could also be, for example, the decision made by the subject on a given trial)
- ▷ at trial $k \in \{1, \ldots, K\}$

Our goal is to identify the primary dimensions along which these activity patterns differ with changes in one (or two) of these four independent variables.

2. Decomposing the covariance

We begin by decomposing these patterns of activity into variable-dependent fluctuations around variable-specific means. We define the ϕ-specific mean as

$$\bar{\mathbf{x}}_\phi := \langle \mathbf{x}_{tsdk} \rangle_{\neg\phi}$$

where the angular brackets $\langle\cdot\rangle_{\neg\phi}$ denote an average over all other variables. For example, the trial-averaged activity at time t in response to stimulus s in condition d is given by

$$\bar{\mathbf{x}}_{tsd} := \langle \mathbf{x}_{tsdk} \rangle_k = \frac{1}{K} \sum_k \mathbf{x}_{tsdk}$$

The mean activity at time t averaged over all stimuli, conditions, and trials, is given by

$$\bar{\mathbf{x}}_t := \langle \mathbf{x}_{tsdk} \rangle_{sdk} = \frac{1}{SDK} \sum_{s,d,k} \mathbf{x}_{tsdk}$$

And so forth.

With this definition in hand, we can begin by decomposing each data point into trial-by-trial fluctuations

$$\mathbf{x}_{tsdk} = \bar{\mathbf{x}}_{tsd} + \underbrace{(\mathbf{x}_{tsdk} - \bar{\mathbf{x}}_{tsd})}_{d\mathbf{x}_{tsdk}}$$

where $d\mathbf{x}_{tsdk}$ denotes trial-by-trial fluctuations around the trial average $\bar{\mathbf{x}}_{tsd}$. We next decompose the trial-average activity into fluctuations around the temporal mean $\bar{\mathbf{x}}_t$,

$$= \bar{\mathbf{x}}_t + \underbrace{(\bar{\mathbf{x}}_{tsd} - \bar{\mathbf{x}}_t)}_{d\mathbf{x}_t} + d\mathbf{x}_{tsdk}$$

The term $d\mathbf{x}_t$ denotes time-specific fluctuations across different stimuli and conditions, averaged over trials. To isolate the effects of changing the stimulus or condition *only*, we next decompose $d\mathbf{x}_t$ into

- ▷ stimulus-specific fluctuations $d\mathbf{x}_{ts}$ (fluctuations across different stimuli, averaged over trials and conditions)
- ▷ condition-specific fluctuations $d\mathbf{x}_{td}$ (fluctuations across different conditions, averaged over trials and stimuli), and
- ▷ additional interaction effects $d\mathbf{x}_{tsd}$

$$\begin{aligned}
&= \bar{\mathbf{x}}_t + (\underbrace{\bar{\mathbf{x}}_{ts} - \bar{\mathbf{x}}_t}_{d\mathbf{x}_{ts}}) + (\bar{\mathbf{x}}_{tsd} - \bar{\mathbf{x}}_{ts}) + d\mathbf{x}_{tsdk} \\
&= \bar{\mathbf{x}}_t + d\mathbf{x}_{ts} + (\underbrace{\bar{\mathbf{x}}_{td} - \bar{\mathbf{x}}_t}_{d\mathbf{x}_{td}}) + (\underbrace{\bar{\mathbf{x}}_{tsd} - \bar{\mathbf{x}}_{td} - \bar{\mathbf{x}}_{ts} + \bar{\mathbf{x}}_t}_{d\mathbf{x}_{tsd}}) + d\mathbf{x}_{tsdk}
\end{aligned}$$

Lastly, we extract average temporal fluctuations around the grand mean $\bar{\mathbf{x}}$:

$$\begin{aligned}
&= \bar{\mathbf{x}} + (\underbrace{\bar{\mathbf{x}}_t - \bar{\mathbf{x}}}_{d\mathbf{x}_t}) + d\mathbf{x}_{ts} + d\mathbf{x}_{td} + d\mathbf{x}_{tsd} + d\mathbf{x}_{tsdk} \\
&= \bar{\mathbf{x}} + d\mathbf{x}_t + d\mathbf{x}_{ts} + d\mathbf{x}_{td} + d\mathbf{x}_{tsd} + d\mathbf{x}_{tsdk}
\end{aligned}$$

The interaction term $d\mathbf{x}_{tsd}$ quantifies joint stimulus/condition pair-specific fluctuations beyond those captured by $d\mathbf{x}_{ts}$ and $d\mathbf{x}_{td}$:

$$d\mathbf{x}_{tsd} = (\bar{\mathbf{x}}_{tsd} - \bar{\mathbf{x}}_{td}) - d\mathbf{x}_{ts} = (\bar{\mathbf{x}}_{tsd} - \bar{\mathbf{x}}_{ts}) - d\mathbf{x}_{td}$$

We can now use this decomposition to similarly decompose the data covariance $\mathbf{\Sigma}$. We first define the *centered* data, or fluctuations around the grand mean,

$$\begin{aligned}
d\bar{\mathbf{x}}_{tsdk} &:= \mathbf{x}_{tsdk} - \bar{\mathbf{x}} \\
&= d\mathbf{x}_t + d\mathbf{x}_{ts} + d\mathbf{x}_{td} + d\mathbf{x}_{tsd} + d\mathbf{x}_{tsdk}
\end{aligned}$$

This allows us to succinctly express the covariance as

$$\begin{aligned}
\mathbf{\Sigma} &:= \left\langle d\bar{\mathbf{x}}_{tsdk} d\bar{\mathbf{x}}_{tsdk}^T \right\rangle_{tsdk} = \frac{1}{TSDK} \sum_{t,s,d,k} d\bar{\mathbf{x}}_{tsdk} d\bar{\mathbf{x}}_{tsdk}^T \\
&= \sum_{\phi,\phi' \in \{t,ts,td,tsd,tsdk\}} \left\langle d\mathbf{x}_\phi d\mathbf{x}_{\phi'}^T \right\rangle_{tsdk}
\end{aligned}$$

We next note that each of the five fluctuation terms in $d\bar{\mathbf{x}}_{tsdk}$ are in fact uncorrelated with each other: for any $\phi \neq \phi'$,

$$\left\langle d\mathbf{x}_\phi d\mathbf{x}_{\phi'}^T \right\rangle_{tsdk} = 0$$

The reason for this is that each fluctuation term $d\mathbf{x}_\phi$ is defined as fluctuations relative to some ϕ'-specific mean, where $\phi' \neq \phi$. Thus, by definition, these fluctuations average out to 0 when averaged over ϕ'. A few examples:

$$\begin{aligned}
\left\langle d\mathbf{x}_t d\mathbf{x}_{ts}^T \right\rangle_{tsdk} &= \left\langle d\mathbf{x}_t \left\langle d\mathbf{x}_{ts} \right\rangle_s^T \right\rangle_t = \left\langle d\mathbf{x}_t (\left\langle \bar{\mathbf{x}}_{ts} \right\rangle_s - \bar{\mathbf{x}}_t)^T \right\rangle_t \\
&= \left\langle d\mathbf{x}_t (\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_t)^T \right\rangle_t = 0 \\
\left\langle d\mathbf{x}_{ts} d\mathbf{x}_{tsd}^T \right\rangle_{tsdk} &= \left\langle d\mathbf{x}_{ts} \left\langle d\mathbf{x}_{tsd} \right\rangle_d^T \right\rangle_{ts} = \left\langle d\mathbf{x}_{ts} (\left\langle \bar{\mathbf{x}}_{tsd} \right\rangle_d - \bar{\mathbf{x}}_{ts} - \left\langle \bar{\mathbf{x}}_{td} \right\rangle_d + \bar{\mathbf{x}}_t)^T \right\rangle_{ts} \\
&= \left\langle d\mathbf{x}_{ts} (\bar{\mathbf{x}}_{ts} - \bar{\mathbf{x}}_{ts} - \bar{\mathbf{x}}_t + \bar{\mathbf{x}}_t)^T \right\rangle_{ts} = 0
\end{aligned}$$

All the cross-terms in the sum are therefore 0, leaving us with the following expression for the covariance:

$$\begin{aligned}
\mathbf{\Sigma} &= \mathbf{\Sigma}_t + \mathbf{\Sigma}_{ts} + \mathbf{\Sigma}_{td} + \mathbf{\Sigma}_{tsd} + \mathbf{\Sigma}_{tsdk} \\
\mathbf{\Sigma}_\phi &:= \left\langle d\mathbf{x}_\phi d\mathbf{x}_\phi^T \right\rangle_\phi, \quad \phi \in \{t, ts, td, tsd, tsdk\}
\end{aligned}$$

Each covariance term corresponds to the covariance of the variable-specific fluctuations we used to decompose the data.

Note that this decomposition is not at all unique. Many such decompositions can be written down in terms of fluctuations around other variable-specific means or other interaction effects. The decomposition chosen here is just a natural one for neural data.

3. Computing the dPC's

We are now ready to extract the *demixed principal components*, or dPC's. We can think of these as the dimensions along which the neural activity changes most with isolated changes in one or two independent variables. For example, the dimensions along which neural activity changes most with a change in the stimulus, on average over all time, conditions, and trials. In some sense, these are the dimensions of activity most informative about the stimulus.

To build up intuition for the mathematical formulation of dPC's, we consider first trying to capture the dimensions of highest variance across all variables, i.e. standard principal components analysis. We'll do this using an autoencoder, whereby we reduce the dimensionality of the data to a few "latent variables" $\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \dots & z_Q \end{bmatrix}^T$ that maximally preserve the variability present in the data. We use a linear autoencoder of the form

$$\mathbf{z}_{tsdk} := \mathbf{F} d\bar{\mathbf{x}}_{tsdk}$$

where $\mathbf{F}$ is the $Q \times N$ encoding matrix that determines how each data point is transformed into its latent representation. Since we're focused on capturing dimensions of *variability*, we'll subtract the grand mean from the data and only focus on reducing the dimensionality of the fluctuations around this grand mean, given by the centered data $d\bar{\mathbf{x}}_{tsdk}$. To enforce that these latents actually preserve information about this variability, we'll enforce that there exist an $N \times Q$ decoding matrix $\mathbf{D}$ that can linearly reconstruct these centered data points from their latent representation:

$$d\hat{\mathbf{x}}_{tsdk} := \mathbf{D}\mathbf{z}_{tsdk}$$

where $d\hat{\mathbf{x}}_{tsdk}$ denotes the reconstruction. We then optimize the encoding and decoding matrices $\mathbf{F}, \mathbf{D}$ to minimize the squared reconstruction error

$$\hat{\mathbf{F}}, \hat{\mathbf{D}} := \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \|d\bar{\mathbf{x}}_{tsdk} - d\hat{\mathbf{x}}_{tsdk}\|^2 \right\rangle_{tsdk} = \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \|d\bar{\mathbf{x}}_{tsdk} - \mathbf{D}\mathbf{F} d\bar{\mathbf{x}}_{tsdk}\|^2 \right\rangle_{tsdk}$$

The "principal components" are then given by the rows of $\hat{\mathbf{F}}$: the dimensions determined to be most important to preserve information about the data in the latent representation.

The solution to this equation can be obtained by first expressing the reconstruction error in terms of the data covariance $\mathbf{\Sigma}$:

$$\begin{aligned}
\hat{\mathbf{F}}, \hat{\mathbf{D}} &= \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \|(\mathbf{I} - \mathbf{D}\mathbf{F}) d\bar{\mathbf{x}}_{tsdk}\|^2 \right\rangle_{tsdk} \\
&= \arg\min_{\mathbf{F},\mathbf{D}} \mathrm{Tr}\Big[(\mathbf{I} - \mathbf{D}\mathbf{F})^T (\mathbf{I} - \mathbf{D}\mathbf{F}) \underbrace{\left\langle d\bar{\mathbf{x}}_{tsdk} d\bar{\mathbf{x}}_{tsdk}^T \right\rangle_{tsdk}}_{\mathbf{\Sigma}} \Big] \\
&= \arg\min_{\mathbf{F},\mathbf{D}} \left\| \mathbf{\Sigma}^{\frac{1}{2}} - \mathbf{D}\mathbf{F}\mathbf{\Sigma}^{\frac{1}{2}} \right\|_{\mathcal{F}}^2
\end{aligned}$$

where in the third line we recognized the covariance matrix $\mathbf{\Sigma}$ and in the last line we took its matrix square root $\mathbf{\Sigma}^{\frac{1}{2}}$. The notation $\|\cdot\|_{\mathcal{F}}$ denotes the matrix Frobenius norm[1]. We thus note that the reconstruction error can be expressed as the Frobenius norm of the difference between a full-rank matrix $\mathbf{\Sigma}^{\frac{1}{2}}$ and a *low-rank* matrix $\mathbf{D}\mathbf{F}\mathbf{\Sigma}^{\frac{1}{2}}$, which in this case is rank Q (i.e. the number of rows/columns of $\mathbf{D}/\mathbf{F}$). By the Eckart-Young-Mirsky theorem[2] this norm is minimized when the rank-Q matrix is equal to the top Q components of the singular value decomposition of the full-rank matrix. Because the covariance matrix $\mathbf{\Sigma}$ is symmetric, we have that its singular value decomposition is equal to its eigendecomposition:

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{S}\mathbf{U}^T$$

[1] $\|\mathbf{A}\|_{\mathcal{F}} = \sqrt{\mathrm{Tr}\left[\mathbf{A}^T\mathbf{A}\right]}$

[2] https://en.wikipedia.org/wiki/Low-rank_approximation#Basic_low-rank_approximation_problem

where $\mathbf{U}$ contains the (orthogonal) eigenvectors of $\mathbf{\Sigma}$ in its columns, and $\mathbf{S}$ is a diagonal matrix containing their associated (real and positive) eigenvalues on the diagonal, orderd by size. Its matrix square root can therefore be similarly decomposed as $\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T$. The reconstruction error is therefore minimized when

$$\mathbf{DF}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{U}_Q\mathbf{S}_Q^{\frac{1}{2}}\mathbf{U}_Q^T$$

where $\mathbf{U}_Q$ denotes the matrix $\mathbf{U}$ truncated to its first Q columns (i.e. to the Q eigenvectors with largest eigenvalues), and similarly $\mathbf{S}_Q$ is the $Q \times Q$ diagonal matrix with the associated eigenvalues. Multiplying both sides by the inverse of $\mathbf{\Sigma}^{\frac{1}{2}}$, we arrive at our solution:

$$\hat{\mathbf{D}}\hat{\mathbf{F}} = \mathbf{U}_Q\mathbf{U}_Q^T$$

Any pair of decoding matrices $\mathbf{D}, \mathbf{F}$ satisfying this equation will minimize the reconstruction error. A natural choice is

$$\hat{\mathbf{D}} = \mathbf{U}_Q$$
$$\hat{\mathbf{F}} = \mathbf{U}_Q^T$$

whereby the principal components are given by the rows of $\mathbf{U}_Q^T$ – that is, the top Q eigenvectors of the covariance matrix, which is exactly the solution of PCA!

We'll now use exactly this same formalism to compute the *demixed* principal components. The only difference is that now we will build an autoencoder optimized to capture variable-specific variability; that is, variability across different settings of a single independent variable (or pair thereof), rather than over the whole data set. This requires simply changing the target of our reconstruction:

$$\hat{\mathbf{F}}_\phi, \hat{\mathbf{D}}_\phi := \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \|d\mathbf{x}_\phi - \mathbf{DF}d\bar{\mathbf{x}}_{tsdk}\|^2 \right\rangle_{tsdk}$$

where $d\mathbf{x}_\phi$ denotes the particular fluctuations we want to capture, over changes in a specific variable ϕ. The rows of $\hat{\mathbf{F}}_\phi$ then denote the dPC's of ϕ-specific variability. For example, the dPC's of stimulus-specific variability are given by the rows of the matrix $\hat{\mathbf{F}}_{ts}$, defined by

$$\hat{\mathbf{F}}_{ts}, \hat{\mathbf{D}}_{ts} := \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \|d\mathbf{x}_{ts} - \mathbf{DF}d\bar{\mathbf{x}}_{tsdk}\|^2 \right\rangle_{tsdk}$$

where $d\mathbf{x}_{ts}$ are the stimulus-specific fluctuations defined in the previous section.

Note that in this case the target reconstruction isn't the same as the input to the autoencoder, so we can't follow the same steps we did above to obtain a solution. This problem instead resembles a classic ordinary least-squares regression problem of the form

$$\hat{\mathbf{W}}_\phi^{\text{OLS}} := \arg\min_{\mathbf{W}} \left\langle \|d\mathbf{x}_\phi - \mathbf{W}d\bar{\mathbf{x}}_{tsdk}\|^2 \right\rangle_{tsdk} = \left\langle d\mathbf{x}_\phi d\bar{\mathbf{x}}_{tsdk}^T \right\rangle_{tsdk} \left\langle d\bar{\mathbf{x}}_{tsdk} d\bar{\mathbf{x}}_{tsdk}^T \right\rangle_{tsdk}^{-1} = \mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-1}$$

where OLS stands for "ordinary least-squares". The final equality follows from our above observation that $d\mathbf{x}_\phi$ is uncorrelated with all the other the components of $d\bar{\mathbf{x}}_{tsdk}$, so that $\left\langle d\mathbf{x}_\phi d\bar{\mathbf{x}}_{tsdk}^T \right\rangle_{tsdk} = \left\langle d\mathbf{x}_\phi d\mathbf{x}_\phi^T \right\rangle_\phi = \mathbf{\Sigma}_\phi$. We can use this ordinary least-squares solution to decompose the reconstruction error as follows:

$$\left\langle \|d\mathbf{x}_\phi - \mathbf{DF}d\bar{\mathbf{x}}_{tsdk}\|^2 \right\rangle_{tsdk} = \left\langle \left\| \underbrace{(d\mathbf{x}_\phi - \hat{\mathbf{W}}_\phi^{\text{OLS}} d\bar{\mathbf{x}}_{tsdk})}_{\text{regression residuals}} + \underbrace{(\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}})d\bar{\mathbf{x}}_{tsdk}}_{\text{low-rank approximation error}} \right\|^2 \right\rangle_{tsdk}$$

The *regression residuals* denote errors that cannot be accounted even by the optimal least-squares solution. The *low-rank approximation error*, on the other hand, is due to the fact that the rank-Q matrix $\mathbf{DF}$ can't reproduce this optimal solution unless $Q = N$. This is effectively quantifying the amount of information we are throwing away by reducing dimensionality in our latent representation of the data, which is only sensitive to Q dimensions of the data (the Q dPC's in the rows of $\mathbf{F}$). The dPC's we are after are the ones the minimize this approximation error, i.e. the ones that throw away the least amount of information. Indeed, these two sources of error are in fact uncorrelated, meaning that minimizing the

full reconstruction error is exactly equivalent to minimizing the low-rank approximation error:

$$\begin{aligned}
\hat{\mathbf{F}}_\phi, \hat{\mathbf{D}}_\phi &= \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \left\| (d\mathbf{x}_\phi - \hat{\mathbf{W}}_\phi^{\text{OLS}} d\bar{\mathbf{x}}_{tsdk}) + (\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} \\
&= \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \left\| d\mathbf{x}_\phi - \hat{\mathbf{W}}_\phi^{\text{OLS}} d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} + \left\langle \left\| (\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} \\
&\qquad + 2 \left\langle \left(d\mathbf{x}_\phi - \hat{\mathbf{W}}_\phi^{\text{OLS}} d\bar{\mathbf{x}}_{tsdk} \right)^T (\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) d\bar{\mathbf{x}}_{tsdk} \right\rangle_{tsdk} \\
&= \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \left\| d\mathbf{x}_\phi - \hat{\mathbf{W}}_\phi^{\text{OLS}} d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} + \left\langle \left\| (\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} \\
&\qquad + 2\text{Tr} \left[(\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) \left(\left\langle d\bar{\mathbf{x}}_{tsdk} d\mathbf{x}_\phi^T \right\rangle_{tsdk} - \left\langle d\bar{\mathbf{x}}_{tsdk} d\bar{\mathbf{x}}_{tsdk}^T \right\rangle_{tsdk} \left(\hat{\mathbf{W}}_\phi^{\text{OLS}} \right)^T \right) \right] \\
&= \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \left\| d\mathbf{x}_\phi - \hat{\mathbf{W}}_\phi^{\text{OLS}} d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} + \left\langle \left\| (\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} \\
&\qquad + 2\text{Tr} \left[(\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) \left(\mathbf{\Sigma}_\phi - \mathbf{\Sigma} \left(\mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\phi \right) \right) \right] \\
&= \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \left\| d\mathbf{x}_\phi - \hat{\mathbf{W}}_\phi^{\text{OLS}} d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} + \left\langle \left\| (\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk} \\
&= \arg\min_{\mathbf{F},\mathbf{D}} \left\langle \left\| (\mathbf{DF} - \hat{\mathbf{W}}_\phi^{\text{OLS}}) d\bar{\mathbf{x}}_{tsdk} \right\|^2 \right\rangle_{tsdk}
\end{aligned}$$

where in the fourth line we plugged in the OLS solution $\hat{\mathbf{W}}_\phi^{\text{OLS}} = \mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-1}$, and in the last line we dropped all terms that did not depend on $\mathbf{F}, \mathbf{D}$. We can now proceed as we did above, by (1) using the trace trick to express the reconstruction error in terms of the covariance, (2) using the matrix square root to express the reconstruction error as the Frobenius norm of the difference between a full-rank and a low rank matrix, and (3) applying the Eckart-Young-Mirsky theorem to obtain an expression for the solution:

$$\begin{aligned}
\hat{\mathbf{F}}_\phi, \hat{\mathbf{D}}_\phi = \arg\min_{\mathbf{F},\mathbf{D}} \left\| \mathbf{DF} \mathbf{\Sigma}^{\frac{1}{2}} - \hat{\mathbf{W}}_\phi^{\text{OLS}} \mathbf{\Sigma}^{\frac{1}{2}} \right\|_{\mathcal{F}} &= \arg\min_{\mathbf{F},\mathbf{D}} \left\| \mathbf{DF} \mathbf{\Sigma}^{\frac{1}{2}} - \mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-\frac{1}{2}} \right\|_{\mathcal{F}} \\
\Leftrightarrow \hat{\mathbf{D}}_\phi \hat{\mathbf{F}}_\phi \mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{U}_Q \mathbf{S}_Q \mathbf{V}_Q^T &= \mathbf{U}_Q \mathbf{U}_Q^T \mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-\frac{1}{2}}
\end{aligned}$$

where $\mathbf{U}_Q, \mathbf{V}_Q^T, \mathbf{S}_Q$ are the top Q left/right singular vectors and singular values, respectively, of $\mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-\frac{1}{2}}$. Multiplying both sides by the inverse of $\mathbf{\Sigma}^{\frac{1}{2}}$ then gives us the final expression that our dPC solution must satisfy:

$$\Leftrightarrow \hat{\mathbf{D}}_\phi \hat{\mathbf{F}}_\phi = \mathbf{U}_Q \mathbf{U}_Q^T \mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-1}$$

Note that the left singular vectors $\mathbf{U}_Q$ are also the eigenvectors of the matrix $\mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\phi$ (whose eigenvalues are $\mathbf{S}_Q^2$).

Finally, we pick a solution for $\hat{\mathbf{D}}_\phi, \hat{\mathbf{F}}_\phi$ that satisfies this equation. Infinitely many solutions exist, and they all give different (but related) values for the dPC's (that is, the rows of $\hat{\mathbf{F}}_\phi$). The solution picked by [Kobak et al., 2016] is as follows[3]:

$$\begin{aligned}
\hat{\mathbf{D}}_\phi &= \mathbf{U}_Q \\
\hat{\mathbf{F}}_\phi &= \mathbf{U}_Q^T \mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-1}
\end{aligned}$$

This particular solution carries the great advantage that the values of the first $q \leq Q$ dPC's remains the same no matter what Q is: increasing Q requires only adding new columns/rows to $\hat{\mathbf{D}}_\phi$ / $\hat{\mathbf{F}}_\phi$, without modifying the ones already there. One can therefore simply set $Q = N$ and then look at the amount of variance explained by each dPC and decide which ones to pick *a posteriori*. The disadvantage to this solution is that the dPC's are not necessarily orthogonal. [Kobak et al., 2016] found that, in practice,

[3] It is straight-forward to verify that the solution for $\hat{\mathbf{F}}$ are the top Q eigenvectors of the matrix $\mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\phi^2$. Note the relationship of this solution to multi-class LDA, where the solution is given by top eigenvectors of $\mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\phi$. In fact, LDA can also be interpreted as a reduced-rank regression problem, but where the reconstruction error is defined in terms of the ϕ-specific class of each data point – i.e. a one-hot vector of dimension equal to the number of values ϕ could take on (e.g. TS in the case of $\phi = ts$) –, rather than the ϕ-specific mean deviation $d\mathbf{x}_\phi$ corresponding to the class of that that data point.

the dPC's capturing most variance tend to be close to orthogonal in the data sets that they analyzed. But nothing in the method guarantees this.

This can be guaranteed by instead picking a different solution. One such solution can be obtained by performing the singular value decomposition

$$\mathbf{U}_Q \mathbf{U}_Q^T \mathbf{\Sigma}_\phi \mathbf{\Sigma}^{-1} = \tilde{\mathbf{U}}_Q \tilde{\mathbf{S}}_Q \tilde{\mathbf{V}}_Q^T$$

and setting

$$\hat{\mathbf{D}} = \tilde{\mathbf{U}}_Q \tilde{\mathbf{S}}_Q$$
$$\hat{\mathbf{F}} = \tilde{\mathbf{V}}_Q^T$$

in which case the dPC's (the rows of $\hat{\mathbf{F}}$) are guaranteed to be orthogonal. However, this carries the disadvantage that each value of Q requires computing a different singular value decomposition, which means that all the dPC's change when Q does.

References

[Kobak et al., 2016] Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C. K. (2016). Demixed principal component analysis of neural population data. *eLife*, 5:e10989.