Principal Components Analysis (PCA)

Mathematical Tools for Neuroscience (NEU 314) Fall, 2021

lecture 11

Jonathan Pillow
Summary of prev (online-only) lecture

• outer product (review)
• SVD as a sum of weighted outer products
• optimal low-rank matrix approximation using SVD
• Frobenius norm (Euclidean norm for matrices)
quick review: **outer product**

\[ \vec{a} \vec{b}^\top = \mathbf{C} \]

- produces a rank-1 matrix

\[
\begin{bmatrix}
    1 \\
    k \\
\end{bmatrix} = \begin{bmatrix}
    1 & k \\
\end{bmatrix}
\]

• produces a rank-1 matrix
SVD as a sum of outer-products

\[ A = USV^T \]

\[ = S_1 u_1 v_1^T + S_2 U_2 V_2^T + \ldots + S_n U_n V_n^T \]

(rank 1 matrix)
matrix approximation

• the best rank-K approximation to $A$ (in terms of squared error) is given by truncating the SVD after $K$ terms.

\[
\begin{align*}
&= \begin{bmatrix} S_1 & \vdots & \vdots \\
\vdots & \ddots & \vdots \\
\vdots & \vdots & S_k \\
u_1 & \ldots & u_k \\
\end{bmatrix} + \ldots + \begin{bmatrix} S_K \\
\vdots & \ddots & \vdots \\
\vdots & \vdots & \vdots \\
u_K & \ldots & v_k \\
\end{bmatrix} \\
&= \begin{bmatrix} u_1 & \ldots & u_k \\
\end{bmatrix} \begin{bmatrix} v_1 & \ldots & v_k \\
\end{bmatrix}
\]

(this is an easier way to compute it)
**Fraction of variance** accounted for (by the rank-K approximation):

\[
\frac{\sum_{i=1}^{K} s_i^2}{\sum_{j=1}^{N} s_j^2}
\]

- sum of squared first K singular values
- sum of squares of all singular values
applications to neural data

(adapted from Williams et al, Neuron 2018)
Frobenius norm
(the Euclidean norm for matrices)

\[ \|A\|_F = \sqrt{\sum_{ij} a_{ij}^2} = \sqrt{\sum_{i} s_i^2} \]

sum of squared elements of A

sum of squared singular values of A

(see notes for proof)
PCA
PCA summary

the data

\[ X = \begin{bmatrix}
    \tilde{x}_1 \\
    \tilde{x}_2 \\
    \vdots \\
    \tilde{x}_N \\
\end{bmatrix} \]

N

d

2nd moment matrix

\[ C = X^\top X \]

SVD

\[ C = U S U^\top \]

first k PCs:

\( \{ u_1, \ldots, u_k \} \)

sum of squares of data within subspace:

\( s_1 + \cdots + s_k \)
1 The raw data

Suppose someone hands you a stack of \( N \) vectors, \( \{ \tilde{x}_1, \ldots, \tilde{x}_N \} \), each of dimension \( d \). For example, we might imagine we have made a simultaneous recording from \( d \) neurons, so each vector represents the spike counts of all recorded neurons in a single time bin, and we have \( N \) time bins total in the experiment.

Let's think of the data arranged in an \( N \times d \) matrix that we'll call \( X \). Each row of this matrix is a data vector representing the response from \( d \) neurons to a single stimulus:

\[
X = \begin{bmatrix}
- \tilde{x}_1 \\
- \tilde{x}_2 \\
\vdots \\
- \tilde{x}_N 
\end{bmatrix}
\]

We suspect that these vectors not "fill" out the entire \( d \)-dimensional space, but instead be confined to a lower-dimensional subspace. (For example, if two neurons always emit the same number of spikes, then their responses live entirely along the 1D subspace corresponding to the \( x_i = x_j \) line).

Can we make a mathematically rigorous theory of dimensionality reduction that captures how much of the "variance" in the data is captured by a low-dimensional projection? (Yes: it turns out the tool we are looking for is PCA!)

2 Finding the best 1D subspace (first PC)

Let's suppose we wish to find the best 1D subspace, i.e., the one-dimensional projection of the data that captures the largest amount of variability. We can formalize this as the problem of finding the unit vector \( \tilde{v} \) that maximizes the sum of squared linear projections of the data vectors:

\[
\text{Sum of squared linear projections} = \sum_{i=1}^{N} (\tilde{x}_i \cdot \tilde{v})^2 = ||X \tilde{v}||^2 = (X \tilde{v})^T (X \tilde{v}) = \tilde{v}^T X^T X \tilde{v}
\]

\[
\text{fraction of sum of squares: } \frac{s_1 + \cdots + s_k}{s_1 + \cdots + s_N}
\]

\[
\text{fraction of sum of squares: } \frac{s_1 + \cdots + s_k}{s_1 + \cdots + s_N}
\]

\[
\text{SVD} \quad C = U S U^T
\]

\[
\text{2nd moment matrix} \quad C = X^T X
\]
1 The raw data

Suppose someone hands you a stack of $N$ vectors, \{\tilde{x}_1, \ldots, \tilde{x}_N\}, each of dimension $d$. For example, we might imagine we have made a simultaneous recording from $d$ neurons, so each vector represents the spike counts of all recorded neurons in a single time bin, and we have $N$ time bins total in the experiment.

Let's think of the data arranged in an $N \times d$ matrix that we'll call $X$. Each row of this matrix is a data vector representing the response from $d$ neurons to a single stimulus:

\[
X = \begin{bmatrix}
\tilde{x}_1 \\
\tilde{x}_2 \\
\vdots \\
\tilde{x}_N
\end{bmatrix}
\]

We suspect that these vectors not "fill" out the entire $d$-dimensional space, but instead be confined to a lower-dimensional subspace. (For example, if two neurons always emit the same number of spikes, then their responses live entirely along the 1D subspace corresponding to the $x_i = x_j$ line).

Can we make a mathematically rigorous theory of dimensionality reduction that captures how much of the "variance" in the data is captured by a low-dimensional projection? (Yes: it turns out the tool we are looking for is PCA!)

2 Finding the best 1D subspace (first PC)

Let's suppose we wish to find the best 1D subspace, i.e., the one-dimensional projection of the data that captures the largest amount of variability. We can formalize this as the problem of finding the unit vector $\tilde{v}$ that maximizes the sum of squared linear projections of the data vectors:

\[
\text{Sum of squared linear projections} = \sum_{i=1}^{N} (\tilde{x}_i \cdot \tilde{v})^2 = \|X \tilde{v}\|_2^2 = (X \tilde{v})^\top (X \tilde{v}) = \tilde{v}^\top X^\top X \tilde{v}
\]

\[
\frac{s_1 + \cdots + s_k}{s_1 + \cdots + s_N}
\]

\[
2nd \ moment \ matrix \ C = X^\top X \\
\text{SVD} \quad C = USU^\top
\]

first k PCs: \{u_1, \ldots, u_k\}
Two equivalent formulations:

1. $\hat{B}_{pca} = \arg\max_B \|XB\|_F^2$
   such that $B^\top B = I$

   find subspace that preserves \textit{maximal} sum-of-squares

2. $\hat{B}_{pca} = \arg\min_B \|X - XBB^\top\|_F^2$
   such that $B^\top B = I$

   \textit{minimize} sum-of-squares of orthogonal component

reconstruction of $X$ in subspace spanned by $B$