Plan for TissueFBA

John Whitman

August 24, 2024

Purpose: Establish the idea of the PCA cloud.

Math

We have a matrix, $\mathbf{D}$, with dimension $m \times n$, where m is the number of observations of data, and n is the number of features collected per sample. We define the PCA function as taking in a matrix of some dimension, and returning a matrix $\mathbf{B}$ of dimension $o \times n$ where o is the desired, reduced number of dimensions of the operation. Given a data vector $\vec{d}$ (a row from $\mathbf{D}$) which is dimension $1 \times n$ by definition, we can transform the point into the target space by:

$$\vec{d'} = \vec{d} \times \mathbf{B}^T \tag{1}$$

yielding a new vector of $1 \times o$ dimension.

Our goal is to measure the variability in dimensional reduction as a function of subsampling the data population. To do this, we first generate an eigenbasis using the full dataset, $\mathbf{D}$. We define this eigenbasis as $\mathbf{B}_0$. We also generate the transformed datapoints in the eigenbasis: $\{\vec{d_{i0}}' = \vec{d_i} \times \mathbf{B}_0^T : i \in \{1, \ldots, m\},\ d_i \in \mathbf{D}\}$. For absolute clarity, $\vec{d_{i0}}'$ is the transformed datapoint relating to row i of the original datamatrix, in the reduced dimensionality eigenbasis generated by the PCA of the full dataset.

Now, we generate a new datamatrix through subsampling out the population; this datamatrix is of size $l \times n$ where $l \leq m$ is the size of the subsample. Running PCA on this datamatrix (keeping sure to maintain the same target rank of the output dimensional reduction) generates a new eigenbasis $\mathbf{B'}$, with which we may generate new transformed datapoints: $\{\vec{\hat{e}_j} = \vec{e_j} \times \mathbf{B'}^T : j \in \{1, \ldots, l\},\ e_j \in \mathbf{D'}\}$.

At this point, we have up to three representations of any particular data point i: the original generation in full feature space ($\vec{d_i}$), the dimensional reduced point in $\mathbf{B}$ ($\vec{d_i'}$), and, if the data point was in the subsampled population of $\mathbf{D'}$, another dimensionally reduced point ($\vec{e_i'}$). We note that we can project $\vec{\hat{e}_i'}$ into the space of $\mathbf{B}$, since both $\mathbf{B}$ and $\mathbf{B'}$ will be orthogonal eigenbases. In plain terms, this projection provides an abstract measure of the variation in the original PCA as a function of changing the subpopulation of the projection.

We quickly calculate the transformation by moving $\vec{\hat{e}_i'}$ back into feature space and then into $\mathbf{B}$.

$$\begin{aligned} \vec{e_i} &= \vec{\hat{e}_i} \times \left(\mathbf{B'}^T\right)^{-1} \\ \vec{e_i'} &= \vec{\hat{e}_i} \times \left(\mathbf{B'}^T\right)^{-1} \times B^T \end{aligned} \tag{2}$$

where we imply, since $\mathbf{B'}$ is generally non-square, that $\mathbf{B'}^{-1}$ is the Monroe-Penrose psuedo-inverse. Interestingly, since $\mathbf{B'}$ is, by construction, an orthogonal matrix, the matrix inverse is just the transpose. So this matrix multiplication reduces to just $\mathbf{B'} \times B^T$. As a gut check, if $\mathbf{B'} = \mathbf{B}$, this would be the identity matrix and there would be no transformation.