**PCA**

Principal component analysis is a technique for *feature extraction* — so it combines our input variables in a specific way, then we can drop the “least important” variables while still retaining the most valuable parts of all of the variables! As an added benefit, each of the “new” variables after PCA are all independent of one another.

**When should I use PCA?**

1. Do you want to reduce the number of variables, but aren’t able to identify variables to completely remove from consideration?
2. Do you want to ensure your variables are independent of one another?
3. Are you comfortable making your independent variables less interpretable?

If you answered “yes” to all three questions, then PCA is a good method to use. If you answered “no” to question 3, you should not use PCA.

By identifying which “directions” are most “important,” we can compress or project our data into a smaller space by dropping the “directions” that are the “least important.” By projecting our data into a smaller space, we’re reducing the dimensionality of our feature space… but because we’ve transformed our data in these different “directions,” we’ve made sure to keep all original variables in our model!

In fact, every principal component will ALWAYS be orthogonal (a.k.a. official math term for perpendicular) to every other principal component. Because our principal components are orthogonal to one another, they are statistically independent of one another… which is why our columns of Z\* are independent of one another!

Finally, we need to determine how many features to keep versus how many to drop. There are three common methods to determine this, discussed below and followed by an explicit example:

Method 1: We arbitrarily select how many dimensions we want to keep. Perhaps I want to visually represent things in two dimensions, so I may only keep two features. This is use-case dependent and there isn’t a hard-and-fast rule for how many features I should pick.

Method 2: Calculate the proportion of variance explained (briefly explained below) for each feature, pick a threshold, and add features until you hit that threshold. (For example, if you want to explain 80% of the total variability possibly explained by your model, add features with the largest explained proportion of variance until your proportion of variance explained hits or exceeds 80%.)

Method 3: This is closely related to Method 2. Calculate the proportion of variance explained for each feature, sort features by proportion of variance explained and plot the cumulative proportion of variance explained as you keep more features. (This plot is called a scree plot, shown below.) One can pick how many features to include by identifying the point where adding a new feature has a significant drop in variance explained relative to the previous feature, and choosing features up until that point. Once we’ve dropped the transformed variables we want to drop, we’re done! That’s PCA.

**But, like, \*why\* does PCA work?**

While PCA is a very technical method relying on in-depth linear algebra algorithms, it’s a relatively intuitive method when you think about it.

* First, the covariance matrix ***Z***ᵀ***Z*** is a matrix that contains estimates of how every variable in ***Z*** relates to every other variable in ***Z***. Understanding how one variable is associated with another is quite powerful.
* Second, eigenvalues and eigenvectors are important. Eigenvectors represent directions. Think of plotting your data on a multidimensional scatterplot. Then one can think of an individual eigenvector as a particular “direction” in your scatterplot of data. Eigenvalues represent magnitude, or importance. Bigger eigenvalues correlate with more important directions.
* Finally, we make an assumption that more variability in a particular direction correlates with explaining the behavior of the dependent variable. Lots of variability usually indicates signal, whereas little variability usually indicates noise. Thus, the more variability there is in a particular direction is, theoretically, indicative of something important we want to detect.

**Thus, PCA is a method that brings together:**

* A measure of how each variable is associated with one another. (Covariance matrix.)
* The directions in which our data are dispersed. (Eigenvectors.)
* The relative importance of these different directions. (Eigenvalues.)

PCA combines our predictors and allows us to drop the eigenvectors that are relatively unimportant.

**Extensions of PCA**

Some of the most frequently seen extensions is principal component regression, where we take our untransformed Y and regress it on the subset of Z\* that we didn’t drop. This is where the independence of the columns of Z\* comes in; by regressing Y on Z\*, we know that the required independence of independent variables will necessarily be satisfied. However, we will need to still check our other assumptions.