Lab 1: PCA with R

Stat 154 with Prof. Sanchez

PCA with matrix decompositions

As we saw in lecture (and in lab), a principal components analysis boils down to performing a matrix decomposition of some data matrix:

- EVD of cross-product matrices ($\mathbf{X}^\mathsf{T}\mathbf{X}$ or $\mathbf{X}\mathbf{X}^\mathsf{T}$)
- SVD of the data matrix $\mathbf{X}$

Example

For comparison purposes, let's use the data set `USArrests` that comes in R, and perform a Principal Components Analysis with the function `prcomp()`

```r
# PCA with prcomp
pca <- prcomp(USArrests, scale. = TRUE)

names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

- `sdev` is a vector with the standard deviations of the PCs
- `rotation` is the matrix of eigenvectors $\mathbf{V}$ (*aka* loadings)
- `x` is the matrix of PCs $\mathbf{Z}$

1) PCA with EVD of correlation matrix

Recall that the EVD of a square matrix $\mathbf{M}$ is given by:

$$\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$$

where:

- $\mathbf{V}$ is an orthonormal matrix of eigenvectors.
- $\mathbf{\Lambda}$ is the a diagonal matrix of eigenvalues.

Using the EVD of the sample correlation matrix $\mathbf{R} = \frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$, the matrix of principal components $\mathbf{Z}$ is obtained as:

$$\mathbf{Z} = \mathbf{XV}$$

where the matrix of eigenvectors $\mathbf{V}$ is the matrix of *loadings*. With the PCs and the loadings, you can express $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{ZV}^\mathsf{T}$$

Your turn

- Use `scale()` to standardize the `USArrests` data. Call this object `arrests` (this will be the matrix $\mathbf{X}$)
- Compute the sample correlation matrix $\mathbf{R}$ (don't use `cor()`). Call this matrix `R`.
- Use the function `eigen()` to compute the Eigenvalue Decomposition of $\mathbf{R}$.
- Take the output of `eigen()` to create matrices $\mathbf{\Lambda}$ and $\mathbf{V}$
- Confirm that the matrix of loadings returned by `prcomp()` is equal to $\mathbf{V}$
- Compute the product $\mathbf{Z} = \mathbf{XV}$ and check that it's equal to the principal components returned by `prcomp(R)`

2) PCA with SVD of the data matrix

As you know, the SVD of a matrix $\mathbf{A}$ is given by:

$$\mathbf{A} = \mathbf{UDV}^\mathsf{T}$$

where:

- $\mathbf{U}$ is an orthonormal matrix of left singular vectors.
- $\mathbf{D}$ is the a diagonal matrix of singular values.
- $\mathbf{V}$ is an orthonormal matrix of right singular vectors.

Using the SVD of $\mathbf{X}$, the matrix of principal components $\mathbf{Z}$ is usually obtained as:

$$\mathbf{Z} = \mathbf{UD}$$

Using all the extracted components $\mathbf{Z}$, the data matrix can be expressed as:

$$\mathbf{X} = \mathbf{ZV}^\mathsf{T}$$

Your turn

- Use the function `svd()` to compute the Singular Value Decomposition of `arrests`.
- Take the output of `svd()` to create matrices $\mathbf{U}, \mathbf{D}, \mathbf{V}$
- Compute the product $\mathbf{Z} = \mathbf{UD}$ and check that it's equal to the principal components returned by `prcomp()`
- Confirm that the matrix of loadings returned by `prcomp()` is equal to $\mathbf{V}$

3) PCA with EVD of association matrix

Instead of using the correlation matrix, you can also use the *association* matrix $\mathbf{XX}^\mathsf{T}$ (sum-of-squares and cross-products of rows) to perform a PCA.

From the matrix decompositions of sections 1) and 2), how would you obtain the principal components $\mathbf{Z}$, based on $\mathbf{XX}^\mathsf{T}$?