The material from this lesson is drawn from the resources I cited in my Medium article: https://medium.com/@matthew.w.brems/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

I. Eigenvalues and Eigenvectors

Recall that linear transformations are transformations that preserve additivity and scalar multiplication. More importantly, we can write any linear transformation $f(\mathbf{x})$ as the matrix multiplication $\mathbf{Ax}$.

A particular class of linear transformations of interest are those satisfying the equation $\mathbf{Ax} = \lambda\mathbf{x}$, where λ is a scalar. This implies that the matrix $\mathbf{A}$ merely stretches or shrinks $\mathbf{x}$ but does not otherwise change $\mathbf{x}$.

If $\mathbf{Ax} = \lambda\mathbf{x}$, we say that λ is an eigenvalue of $\mathbf{A}$ and that $\mathbf{x}$ is the eigenvector of $\mathbf{A}$ that corresponds to λ.

- Given a matrix $\mathbf{A}$, we showed that we can find λ by solving $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ for λ.
- Given a matrix $\mathbf{A}$ and a nonzero eigenvalue λ, we showed that we can find the corresponding eigenvector $\mathbf{x}$ by solving $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ for $\mathbf{x}$.

II. Decomposition of Matrices

We have discussed before the idea of decomposing one matrix $\mathbf{A}$ into multiple matrices.

Formally, a matrix $\mathbf{A}$ can be decomposed into matrices $\mathbf{X}$ and $\mathbf{Y}$ if $\mathbf{A} = \mathbf{XY}$.

- The spectral decomposition (also called the eigendecomposition) of a matrix $\mathbf{A}$ is where $\mathbf{A} = \mathbf{PDP}^{-1}$ with $\mathbf{D} = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $\mathbf{P}$ consisting of the eigenvectors corresponding to the eigenvalues in $\mathbf{D}$.
  - Suppose we want to find $\mathbf{A}^k$.

$$\begin{aligned}\mathbf{A}^k &= (\mathbf{PDP}^{-1})^k \\ &= (\mathbf{PDP}^{-1})(\mathbf{PDP}^{-1})\cdots(\mathbf{PDP}^{-1}) \\ &= \mathbf{PD}^k\mathbf{P}^{-1} \\ &= \mathbf{P} \times \text{diag}\{\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k\} \times \mathbf{P}^{-1}\end{aligned}$$

  - Spectral decomposition will also be very important in principal component analysis.
  - This works for any diagonalizable (also called "diagonable") matrix, which means there are n independent eigenvectors.

III. Covariance Matrix

Recall that, for any two random variables, we can calculate the covariance.

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n}$$

When working with a vector of random variables, we can construct a covariance matrix. Consider the matrix $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2 \;\; \cdots \;\; \mathbf{X}_p]$.

$$\begin{aligned}\text{Cov}(\mathbf{X}) &= \begin{bmatrix} \text{Cov}(\mathbf{X}_1, \mathbf{X}_1) & \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_1, \mathbf{X}_p) \\ \text{Cov}(\mathbf{X}_2, \mathbf{X}_1) & \text{Cov}(\mathbf{X}_2, \mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_2, \mathbf{X}_p) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(\mathbf{X}_p, \mathbf{X}_1) & \text{Cov}(\mathbf{X}_p, \mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_p, \mathbf{X}_p) \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}(\mathbf{X}_1) & \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_1, \mathbf{X}_p) \\ \text{Cov}(\mathbf{X}_2, \mathbf{X}_1) & \text{Var}(\mathbf{X}_2) & \cdots & \text{Cov}(\mathbf{X}_2, \mathbf{X}_p) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(\mathbf{X}_p, \mathbf{X}_1) & \text{Cov}(\mathbf{X}_p, \mathbf{X}_2) & \cdots & \text{Var}(\mathbf{X}_p) \end{bmatrix}\end{aligned}$$

Note that the covariance matrix is diagonalizable.

Also note that the covariance matrix stores, in each entry, information about the relationships between different variables. For example, $\text{Cov}(\mathbf{X}_1, \mathbf{X}_2)$ is a measure of how $\mathbf{X}_1$ and $\mathbf{X}_2$ change as the other changes.

Finally, consider how we might find the covariance matrix of $\mathbf{X}$. Let's just look at one entry first: $\text{Cov}(\mathbf{X}_1, \mathbf{X}_2)$.

$$\text{Cov}(\mathbf{X}_1, \mathbf{X}_2) \quad = \quad \frac{\sum_{i=1}^{n}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)}{n}$$

- We're taking each observation of $\mathbf{X}_1$ and subtracting the mean of $\mathbf{X}_1$.
- We're taking each observation of $\mathbf{X}_2$ and subtracting the mean of $\mathbf{X}_2$.
- We're then multiplying the corresponding observations. (i.e. The first element of $\mathbf{X}_1$ times the first element of $\mathbf{X}_2$, the second element of $\mathbf{X}_1$ times the second element of $\mathbf{X}_2$, etc.)
- Then we add all of these products up.
- Finally, we divide by n.

IV. Principal Component Analysis

1. Take the matrix of independent variables $\mathbf{X}$ and, for each column $\mathbf{X}_1$,...,$\mathbf{X}_p$, subtract the mean of that column from each entry.
    i. This ensures that the data set is centered at $\mathbf{0}$.
2. Decide whether or not to standardize. Given the columns of $\mathbf{X}$, are features with higher variance more important than features with lower variance, or is the importance of features independent of the variance? (In this case, importance means how well that feature predicts $\mathbf{Y}$.) If the importance of features is independent of the variance of the features, then divide each observation in a column by that column's standard deviation. Call the centered (and possibly standardized) matrix $\mathbf{Z}$.
    i. If you standardized, then this, combined with step 1, standardizes each column of $\mathbf{X}$ to make sure each column has mean zero and standard deviation 1.
3. Given this new matrix $\mathbf{Z}$, calculate the covariance matrix $\text{Cov}(\mathbf{Z}) = \mathbf{Z}^T\mathbf{Z}$.
    i. Because we didn't divide by n, our covariance matrix will be off by a constant factor, but that isn't going to affect our results.
4. Decompose $\mathbf{Z}^T\mathbf{Z}$ into $\mathbf{PDP}^{-1}$ through diagonalization.
    i. Because we're working with a covariance matrix and covariance matrices have nice properties, we know we can always diagonalize.
    ii. Recall that $\mathbf{P}$ is the set of eigenvectors of your covariance matrix and that $\mathbf{D}$ is the diagonal matrix containing the eigenvalues that correspond to each eigenvector.
5. Rearrange the columns of $\mathbf{D}$ so that the eigenvectors are sorted left to right from largest to smallest. Rearrange the columns of $\mathbf{P}$ accordingly and call this sorted matrix $\mathbf{P}^*$.
    i. For example, if λ_2 is the largest eigenvalue, then take the second column of $\mathbf{P}$ and place it in the first column position.
    ii. The eigenvector with the largest eigenvalue is called the "principal component" or the "first principal component." We order our eigenvalues and their corresponding eigenvectors from largest to smallest so that we can "rank" them in order of importance!
6. Calculate $\mathbf{Z}^* = \mathbf{ZP}^*$.
    i. This new matrix, $\mathbf{Z}^*$ is our data $\mathbf{X}$ that has been centered and standardized, then transformed by our sorted matrix of eigenvectors $\mathbf{P}^*$.

7. Determine how many principal components you want or what proportion of the variance you want to explain with your model. Suppose you decide to keep k principal components. Then drop the right-most $p - k$ variables from $\mathbf{Z}^*$, leaving only the left-most (and thus most important) k variables of $\mathbf{Z}^*$.
    i. The result here is the original data, but only in terms of the most important eigenvectors, with the least important eigenvectors discarded.
    ii. Because each eigenvector is orthogonal to the others, each column of our new dataset $\mathbf{Z}$ is orthogonal to the other columns of $\mathbf{Z}$. This ensures that the assumption of independence of our features in a linear regression model is satisfied.
8. If your goal is to do principal component regression, then fit your model by regressing your dependent variable $\mathbf{Y}$ on the remaining transformed independent variables $\mathbf{Z}^*$.