Consider a collection of random variables $x_1, x_2, \ldots, x_p$. These might be explanatory variables in a linear regression context, in which case we might include the response variable y. The goal of Principal Component Analysis (PCA) is to obtain a lower-dimensional representation of the variables. If we could take p variables down to, say, 2 variables, then we could easily graph the data to look for patterns or features. The definition of the 2 derived variables could also lead to insights into data structure.

PCA is a form of **dimension reduction**. It is often used as a form of exploratory data analysis.

PCA operates by finding "principle components," variables derived from the $x_1, x_2, \ldots, x_p$. The PCs are defined to be linear combinations of the x_i such that as much of the variation in the x_i as possible is captured.

Let $x_1, x_2, \ldots, x_p$ be a collection of random variables. They don't need to be Normally distributed, but it is usually recommended that they be standardized (mean 0 and sd 1). Let $\mathbf{\Sigma}$ be the covariance matrix of the x_i. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ be the eigenvalues and eigenvectors of $\mathbf{\Sigma}$, respectively.

We want to find new variables that are defined as linear combinations of the x_i:

$$z_{ij} = \phi_{1j} x_{i1} + \phi_{2j} x_{i2} + \cdots + \phi_{pj} x_{ip}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. We want the new variables z_j to explain as much variance in the x_i as possible. The z_j are called the *scores* of the jth principal component.

The specific PCA algorithm is to solve the following optimization problem:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

The z_{ij} will have mean 0, so what we are maximizing is the sample variance of the z_{ij}. This can be solved via an eigen decomposition.

The solution to the maximization task is to set the ϕ_j equal to the eigenvectors $\mathbf{e}_j$. Thus, for example, the first PC is

$$z_{i1} = e_{11} x_{i1} + e_{21} x_{i2} + \cdots + e_{p1} x_{ip}$$

The coefficients of the linear combination $e_{11}, e_{21}, \ldots, e_{p1}$ are called the *loadings* of the first PC. The proportion of variance in the x_i that is explained by the first PC is

$$\frac{\lambda_1}{\sum_{j=1}^{p} \lambda_j}$$

Because the loadings are defined as eigenvectors, we have that $\mathbf{e}_i'\mathbf{e}_j = 0$ for all $i \neq j$ (that is, the eigenvectors are orthogonal to one another). Note also that $\sum_{j=1}^{p} e_{kj}^2 = 1$, so all the eigenvectors have length 1. The second PC can be interpreted as the linear combination of the x_i that (a) is uncorrelated with PC 1 and (b) maximizes the variance of the z_{i2}.

The loading vector $\boldsymbol{\phi}_1$ with components $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ defines a direction in feature space along which the data vary the most. If we project the n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ onto this direction, the projected values are the principal component scores $z_{11}, z_{21}, \ldots, z_{n1}$. The second PC is similarly defined in terms of a projection (orthogonal to that for PC 1) onto feature space in a direction of maximal variance. Thus, principal components projects the original data down onto the subspace spanned by $\phi_1, \phi_2, \ldots, \phi_p$.

The loadings for PC 1 are all around 0.5 or 0.6. This suggests that the first PC is an overall average or "index" of the x_i. States that have higher than average values for the three variables will have big values of PC 1. Similarly, states that have lower than average values for the three variables will have big negative values of PC 1. Notice that PC 1 accounts for about 79% of the variance in the x_i.

PC 2 is defined by a contrast, mostly between Murder and Rape. States with higher than average Murder statistics and lower than average Rape statistics (or vice versa) will be highlighted by PC 2. It only accounts for about 15% of the variance in the x_i, so PC 2 is definitely picking up on a more subtle pattern than PC 1.