The Effect of Scaling and Mean Centering Prior to a Principal Component Analysis

Sebastian Raschka
se.raschka@gmail.com

October 2, 2014

A frequently asked question is whether data needs to be centered prior to dimensionality reduction via *Principal Component Analysis (PCA)* or not. Here, the assumption is that the PCA is computed based on the *covariance matrix*, i.e., the k principal components are the eigenvectors of the covariance matrix that correspond to the k largest eigenvalues.

1 Mean Centering Does not Affect the Covariance Matrix

Under the assumption that the PCA is obtained from covariance matrix, the resulting principal components will be the same regardless of whether mean centering was performed or not as long as the covariance matrix stays the same. The following equations will show that *mean centering* does not affect the covariance matrix. Let $\mathbf{x}$ and $\mathbf{y}$ be two random variables so that the covariance between the attributes is calculated as

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{1}$$

The centered variables can be written as

$$x' = x - \bar{x} \text{ and } y' = y - \bar{y} \tag{2}$$

And the centered covariance can be calculated as follows:

$$\sigma'_{xy} = \frac{1}{n-1}\sum_{i}^{n}(x'_i - \bar{x}')(y'_i - \bar{y}') \tag{3}$$

Mean centering results in samples means of 0, $\bar{x}' = 0$ and $\bar{y}' = 0$, thus

$$\sigma'_{xy} = \frac{1}{n-1}\sum_{i}^{n} x'_i y'_i \tag{4}$$

This is equal original covariance matrix, which can be shown by resubstitution:

$$x' = x - \bar{x} \text{ and } y' = y - \bar{y} \tag{5}$$

Even the centering of only one variable, e.g., $\mathbf{x}$, would leave the covariance matrix unaffected.

$$\sigma_{\text{xy}} = \frac{1}{n-1}\sum_{i}^{n}(x'_i - \bar{x}')(y_i - \bar{y}) \tag{6}$$

$$= \frac{1}{n-1}\sum_{i}^{n}(x'_i - 0)(y_i - \bar{y}) \tag{7}$$

$$= \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{8}$$

2 Scaling Does Affect the Covariance Matrix

However, in contrast to mean centering, *scaling* (e.g, pounds into kilogram where 1 pound = 0.453592 kg) **does** have an effect on the covariance matrix and therefore influences the results of a PCA.
Let c be the scaling factor for $\mathbf{x}$. Given that the "original" covariance is calculated as

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{9}$$

the covariance after scaling is calculated as follows:

$$\sigma'_{xy} = \frac{1}{n-1}\sum_{i}^{n}(c \cdot x_i - c \cdot \bar{x})(y_i - \bar{y}) \tag{10}$$

$$= \frac{c}{n-1}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{11}$$

$$\Rightarrow \sigma_{xy} = \frac{\sigma'_{xy}}{c} \tag{12}$$

$$\Rightarrow \sigma'_{xy} = c \cdot \sigma_{xy} \tag{13}$$

Therefore, the scaling of one attribute by a constant c will result in a rescaled covariance $c\sigma_{xy}$. E.g., if an attribute $\mathbf{x}$ measured in pounds is rescaled into kilograms, the covariance between $\mathbf{x}$ and $\mathbf{y}$ will be 0.453592 times less the original value.

3 Standardizing Does Affect the Covariance Matrix

Standardization of features also has an effect on the outcome of a PCA since standardization can be understood as a scaling the covariance between every pair of variables by the product of the standard deviations of each pair of variables. The equation for standardization of a variable is written as

$$z = \frac{x_i - \bar{x}}{\sigma} \tag{14}$$

The "original" covariance matrix:

$$\sigma_{xy} = \frac{1}{n-1} \sum_i^n (x_i - \bar{x})(y_i - \bar{y}) \tag{15}$$

And after standardizing both variables:

$$x' = \frac{x - \bar{x}}{\sigma_x} \text{ and } y' = \frac{y - \bar{y}}{\sigma_y} \tag{16}$$

$$\sigma'_{xy} = \frac{1}{n-1} \sum_i^n (x'_i - 0)(y'_i - 0) \tag{17}$$

$$= \frac{1}{n-1} \sum_i^n \left(\frac{x - \bar{x}}{\sigma_x}\right)\left(\frac{y - \bar{y}}{\sigma_y}\right) \tag{18}$$

$$= \frac{1}{(n-1) \cdot \sigma_x \sigma_y} \sum_i^n (x_i - \bar{x})(y_i - \bar{y}) \tag{19}$$

$$\Rightarrow \sigma'_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{20}$$