The effect of scaling and mean centering of variables prior to a Principal Component Analysis

Sebastian Raschka
se.raschka@gmail.com

09/22/2014

Let us think about whether it matters or not if the variables are centered for applications such as Principal Component Analysis (PCA) if the PCA is calculated from 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

Here, the rational is: If the covariance is the same whether the variables are centered or not, the result of the PCA will be the same.

Let's assume we have the 2 variables $\mathbf{x}$ and $\mathbf{y}$ Then the covariance between the attributes is calculated as

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

Let us write the centered variables as

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

The centered covariance would then be calculated as follows:

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

But since after centering, $\bar{x}' = 0$ and $\bar{y}' = 0$ we have

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

which is our original covariance matrix if we resubstitute back the terms

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

.

Even centering only one variable, e.g., **x** wouldn't affect the covariance:

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

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

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

2 Scaling of variables does affect the covariance matrix

If one variable is scaled, e.g, from pounds into kilogram (1 pound = 0.453592 kg), it does affect the covariance and therefore influences the results of a PCA.

Let c be the scaling factor for **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 would be calculated as:

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

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

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

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

Therefore, the covariance after scaling one attribute by the constant c will result in a rescaled covariance $c\sigma_{xy}$ So if we'd scaled **x** from pounds to kilograms, the covariance between **x** and **y** will be 0.453592 times smaller.

3 Standardizing affects the covariance

Standardization of features will have an effect on the outcome of a PCA (assuming that the variables are originally not standardized). This is because we are 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:

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

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

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

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

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