EIGENVECTOR
RESEARCH INCORPORATED
Research, Training, and Software

196 Hyacinth Road
Manson, WA 98831
www.Eigenvector.com

The Effect of Data Centering on PCA Models

Neal B. Gallagher, Donal O'Sullivan, Manuel Palacios

Key words: Principal Components Analysis, PCA, Centering, Mean-Centering

Introduction: Two common questions associated with mean-centering in principal components analysis (PCA) are the following. 1) If I don't mean-center, is the first principal component [PC 1] the mean of the data? The short answer is that PC 1 is not the mean of the data but it can point in the direction of the mean. And, 2) why does PC 1 capture so much more "variance" when I don't mean-center my data? The short answer is that PC 1 must account for sum-of-squares (ssq) due to the mean when the data aren't centered. For both questions, the extent to which centering affects the PCA model depends on the relative amount of ssq due to the mean and ssq due to variance about the mean. This white paper provides additional insight to the effect of data centering on the PCA model that can help interpretation of PCA results.

Some Definitions: For a data set $\mathbf{X}$ with M samples and N variables, the total ssq, s_{tot}^2, can be calculated as

$$s_{\text{tot}}^2 = \sum_{n=1}^{N} \sum_{m=1}^{M} x_{m,n}^2 = \text{tr}(\mathbf{X}^{\text{T}}\mathbf{X}) \qquad (1)$$

where $x_{m,n}$ is an element of $\mathbf{X}$ with $m = 1, \dots, M$ and $n = 1, \dots, N$, and tr( ) is the trace operator. The mean, $\bar{\mathbf{x}}$, can be calculated from

$$\bar{x}_n = \tfrac{1}{M} \sum_{m=1}^{M} x_{m,n} \;\; ; \;\; \bar{\mathbf{x}}^{\text{T}} = \tfrac{1}{M}\mathbf{1}^{\text{T}}\mathbf{X} \qquad (2)$$

where $\mathbf{1}$ is a vector of ones. The sum-of-squares *about the mean*, s_{err}^2, is given by

$$\begin{aligned} s_{\text{err}}^2 &= \text{tr}\left((\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\text{T}})^{\text{T}}(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\text{T}})\right) = \\ &\text{tr}(\mathbf{X}^{\text{T}}\mathbf{X}) - \text{tr}(2\mathbf{X}^{\text{T}}\mathbf{1}\bar{\mathbf{x}}^{\text{T}} - \bar{\mathbf{x}}\mathbf{1}^{\text{T}}\mathbf{1}\bar{\mathbf{x}}^{\text{T}}) \\ &= \text{tr}(\mathbf{X}^{\text{T}}\mathbf{X}) - M\bar{\mathbf{x}}^{\text{T}}\bar{\mathbf{x}} \\ s_{\text{err}}^2 &= s_{\text{tot}}^2 - s_{\text{mean}}^2 \qquad (3) \end{aligned}$$

where $s_{\text{mean}}^2 = M\bar{\mathbf{x}}^{\text{T}}\bar{\mathbf{x}}$ is defined as the sum-of-squares *due to the mean.* For clarity, total variance s_{var}^2 is proportional to s_{err}^2 and accounts for degrees of freedom. Variance, has the specific definition $s_{\text{var}}^2 = \frac{1}{M-1} s_{\text{err}}^2$. Similarly, the covariance is defined as

$$\text{cov}(\boldsymbol{X}) = \frac{1}{M-1}(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\text{T}})^{\text{T}}(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\text{T}}). \qquad (4)$$

Early applications of PCA often worked with mean-centered data and showed that the PCA eigenvalues are proportional to the "variance" captured in the covariance matrix.[1,2] Mean-centering is used to create models of multivariate data in multivariate statistical process control.[3] However, PCA is ubiquitous to multivariate analysis and applications have evolved to include data that have been center, not-centered and preprocessed in a wide variety of ways. For that reason, it is more general to say that the PCA eigenvalues are proportional to "the total sum-of-squares captured for preprocessed $\mathbf{X}$ about the data origin." (E.g., for s_{tot}^2, the data origin is zero.) The rest of the white paper will work with s_{tot}^2, s_{err}^2 and s_{mean}^2, and define the ratio

$$f = {s_{\text{mean}}^2}\Big/{s_{\text{err}}^2} = {s_{\text{mean}}^2}\Big/{(s_{\text{tot}}^2 - s_{\text{mean}}^2)}. \qquad (5)$$

Centering Example 1: A plot of Data Set 1 is shown in Figure 1. The mean vector is plotted from [0, 0] to [4, 2.3] and the *f*-ratio is 11.2. This means that s_{mean}^2 is 11.2 times the size of s_{err}^2 and it would be expected that PC 1, $\mathbf{p}_1$, would point in the general direction of the mean, $\bar{\mathbf{x}}$. The fraction of variance due to s_{mean}^2 is 91.8% and that due to $\mathbf{p}_1$, is 98.5%.

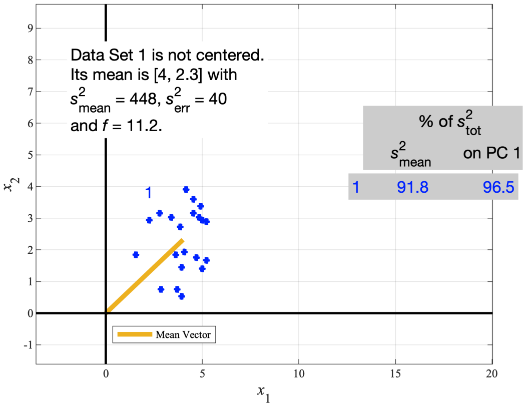

Figure 1: Non-centered data, Data Set 1 with mean drawn from the origin to [4, 2.3].

Figure 2 shows that as the data sets move towards [0, 0] the $s_{\text{mean}}^2/s_{\text{tot}}^2$ decreases as expected. The last data set, Data Set 6, is mean-centered and $s_{\text{mean}}^2/s_{\text{tot}}^2 = 0$. The fraction of s_{tot}^2 on PC1, s_1^2, is greater than due to the mean s_{mean}^2. To see why, recall that PCA finds $\mathbf{p}$ that maximizes capture of ssq. If $\mathbf{X}_\mathrm{c} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^\mathrm{T})$ is the centered data, then the maximization for uncentered data $\mathbf{X}$ for is

$$\max_{\mathbf{p}}\left\{\mathbf{p}^\mathrm{T}\mathbf{X}^\mathrm{T}\mathbf{X}\mathbf{p}\right\} = \max_{\mathbf{p}}\left\{\mathbf{p}^\mathrm{T}\mathbf{X}_\mathrm{c}^\mathrm{T}\mathbf{X}_\mathrm{c}\mathbf{p} + \mathbf{p}^\mathrm{T}\bar{\mathbf{x}}\mathbf{1}^\mathrm{T}\mathbf{1}\bar{\mathbf{x}}^\mathrm{T}\mathbf{p}\right\}. \quad (6)$$

where the ssq captured on $\mathbf{p}$ is the terms in { }. For the first PC, the term on the right-hand-side is

$$\mathbf{p}_1^\mathrm{T}\mathbf{X}_\mathrm{c}^\mathrm{T}\mathbf{X}_\mathrm{c}\mathbf{p}_1 + M\mathbf{p}_1^\mathrm{T}\bar{\mathbf{x}}\bar{\mathbf{x}}^\mathrm{T}\mathbf{p}_1 \geq M\mathbf{p}_1^\mathrm{T}\bar{\mathbf{x}}\bar{\mathbf{x}}^\mathrm{T}\mathbf{p}_1 \qquad (7)$$

where the term to the right of the inequality is the ssq due to the mean captured on $\mathbf{p}_1$. Inequality 7 says that $s_1^2 \geq s_\mathrm{mean}^2$ as shown in the Table in Figure 2.

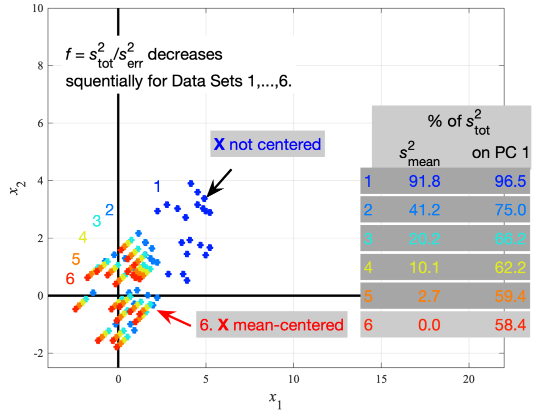

Figure 2: Examples of non-centered data with means closer to the origin as the sets progress from Data Set 1 to 6.

Figure 3 shows that as the data move away from [0, 0] (e.g., from Data Set 6 to 1 in Figure 2), PC 1 points in the direction of the mean: $\mathbf{p}_1 \rightarrow \bar{\mathbf{x}}\|\bar{\mathbf{x}}\|^{-1}$ where $\|\bar{\mathbf{x}}\|^2 = \bar{\mathbf{x}}^\mathrm{T}\bar{\mathbf{x}}$. *For this example*, the angle between $\mathbf{p}_1$ and $\bar{\mathbf{x}}$ is slightly more than 1 degree when $\boldsymbol{f} = 2.1$. Figure 4 shows this more clearly for the example data sets.

Conclusions: A common misconception in PCA is that PC 1 is the mean of the data when the data are not centered. It was shown in this paper that PC 1 is not the mean of the data but it can point in the direction of the mean. The extent to which PC 1 points in the direction of the mean depends on how far away the data set mean is from the origin i.e., when the sum-of-squares due to the mean, s_mean^2, dominates the total sum-of-squares, s_tot^2. For this example, PC 1 pointed in the direction of the mean when $f = s_\mathrm{tot}^2/(s_\mathrm{tot}^2 - s_\mathrm{mean}^2)$ was greater than 2.1. See `ComparePC1toMeandemo` in PLS_Toolbox version 8.9 or higher.[4]

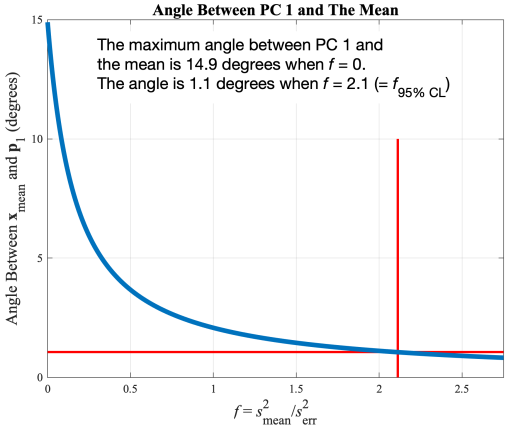

Figure 3: Plot of the angle between $\mathbf{p}_1$ and $\bar{\mathbf{x}}$ as $\boldsymbol{f}$ increases. *For this data*, the approximate 95% limit is 2.1.

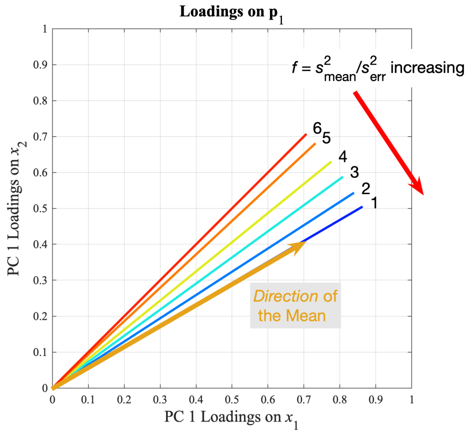

Figure 4: Plot of the $\mathbf{p}_1$ compared to $\bar{\mathbf{x}}$ as $\boldsymbol{f}$ increases.

References:

[1] Jackson JE. A User's Guide to Principal Components, John Wiley & Sons: New York, NY,1991.

[2] Malinowski, E.R., "Factor Analysis in Chemistry", Second Edition, John Wiley & Sons, New York, NY, 1991.

[3] Wise, B.M. and Gallagher, N.B., "The Process Chemometrics Approach to Chemical Process Monitoring and Fault Detection," *J. Proc. Cont* **6**(6), 329-348 (1996).

[4] PLS_Toolbox 8.2.2 (2020). Eigenvector Research, Inc., Manson, WA USA 98831; software available at http://www.eigenvector.com.