EIGENVECTOR RESEARCH INCORPORATED
Research, Training, and Software

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

Introduction to Preprocessing Calibration and Application

Neal B. Gallagher and Jeremy M. Shaver

Key words: Preprocessing, mean-centering, autoscaling

Introduction: Data preprocessing is employed in multivariate analysis but it is often unclear why and in what order specific preprocessing should be employed. The topic gets more confusing when the large number of available methods is considered. In short, the objective of data preprocessing is to separate the signal of interest from clutter where clutter is defined as signal attributable to interferences and noise. Therefore the appropriate preprocessing method depends on the data analysis objective, and on how the signal and clutter manifest in the data. Obviously, this topic can get complicated. However, it is the intention here to provide a brief introduction to the preprocessing objective using simple preprocessing methods as examples. More thorough discussions of the objective, theory and math of each preprocessing procedure are generally included when discussing specific methods [e.g., see Martens, et al. (2003) and Gallagher, et al (2005)].

Preprocessing is typically performed prior to data analysis methods such as principal components analysis (PCA) or partial least squares regression (PLS). Recall, that PCA maximizes the capture of sum-of-squares with factors or principal components (PCs) within a single block of data, and PLS is slightly more complicated method that finds linear relationships between two blocks of data. This introduction will use PCA and two of the simplest examples of preprocessing; mean-centering and autoscaling.

Mean-Centering: Imagine that the objective is to perform exploratory analysis of an *M*x*N* data matrix $\mathbf{X}$ using PCA. Recall that samples (or objects) correspond to the rows of $\mathbf{X}$ and variables correspond to the columns. If no preprocessing is applied to $\mathbf{X}$ prior to the PCA decomposition, then the first principal component (PC) will point in the direction that captures the most sum-of-squares about zero (variance about zero).

Next, define the *N*x*1* mean of data matrix $\mathbf{X}$ as $\bar{\mathbf{x}}$. The mean is calculated down the rows of $\mathbf{X}$ so that for the n^{th} column of $\mathbf{X}$ (i.e., the n^{th} element of the vector $\bar{\mathbf{x}}$), the mean is a scalar given by

$$\bar{x}_n = \frac{1}{M}\sum_{m=1}^{M} x_n \qquad (1)$$

The mean centered data $\mathbf{X}_{mncn}$ is then calculated by subtracting the column mean from the corresponding column so that

$$x_{m,n,mncn} = x_{m,n} - \bar{x}_n \quad \text{for} \quad n = 1,...,N;\ m = 1,...,M$$

$$\mathbf{x}_{n,mncn} = \mathbf{x}_n - \mathbf{1}\bar{x}_n \quad \text{for} \quad n = 1,...,N$$

$$\mathbf{X}_{mncn} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T \qquad (2)$$

where $\mathbf{1}$ is a *M*x*1* vector of ones (typically it is assumed that $\mathbf{1}$ is of appropriate size) and ^{T} is the transpose operator. (The notations in Equation 2 provide identical results, but the simplicity of the last form shows why the linear algebra notation is often preferred.) The first step in mean-centering, represented by Equation 1, is to calculate the mean of each column of $\mathbf{X}$. This procedure can be considered "calibration" of the mean-centering preprocessing and it consists of estimating the mean from the "calibration" data $\mathbf{X}$. The second step, represented by Equation 2, subtracts the mean from the data. This procedure can be considered "applying" the centering to the data $\mathbf{X}$. The first PC of a PCA model of $\mathbf{X}_{mncn}$ will then capture the most sum-of-squares about the mean $\bar{\mathbf{x}}$ (variance about the mean or simply 'variance'). The mean is now a part of the overall PCA model "calibrated" on the "calibration" data $\mathbf{X}$ and the mean-centering operation has changed what sum-of-squares is captured by the first PC. In other words the preprocessing has changed the data to get the PCA model to focus on a different type of variance. As a result, the PCA model must be interpreted differently during the exploratory analysis.

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

Next, assume that a new M_2xN data matrix $\mathbf{X}_2$ was available where $M_2 \geq 1$. To apply the PCA model calibrated above to the new data, the new data set must first be centered to the mean of the calibration data. The preprocessing is "applied" to the new data $\mathbf{X}_2$ using a procedure analogous to Equation 2 as follows

$$\mathbf{x}_{2,n,mncn} = \mathbf{x}_{2,n} - \mathbf{1}\bar{x}_n \,. \qquad (3)$$

Autoscaling: Autoscaling of the data is treated in a manner very similar to mean-centering but the preprocessing includes an additional step. During calibration, Equation 1 is first used to estimate the mean $\bar{\mathbf{x}}$ of the calibration data $\mathbf{X}$. Next, the standard deviation of each column is calculated using

$$s_n = \left[\frac{1}{M-1} \sum_{m=1}^{M} \left(x_n - \bar{x}_n \right)^2 \right]^{1/2} . \qquad (4)$$

Equations 1 and 3 correspond to "calibration" of the autoscaling preprocessing procedure. After calibration, the mean-centered columns are divided by the corresponding standard deviation as follows.

$$\mathbf{x}_{n,auto} = \mathbf{x}_{n,mncn} / s_n = \left(\mathbf{x}_n - \mathbf{1}\bar{x}_n \right) / s_n \qquad (5)$$

Autoscaling includes mean-centering and division by the standard deviation and Equation 6 corresponds to "applying" the preprocessing to the calibration data. Equation 6 shows how the preprocessing is applied to new data $\mathbf{X}_2$.

$$\mathbf{x}_{2,n,auto} = \mathbf{x}_{2,n,mncn} / s_n = \left(\mathbf{x}_{2,n} - \mathbf{1}\bar{x}_n \right) / s_n \qquad (6)$$

In summary, the autoscaling preprocessing parameters $\bar{x}_n$ and $\bar{s}_n$ for $n = 1,...,N$ were estimated from the calibration data $\mathbf{X}$, and the application step used these parameters to center and scale both $\mathbf{X}$ and new data $\mathbf{X}_2$. It should be clear that the calibration data should be sufficiently representative of what is expected in the future if the estimated preprocessing parameters will adequately represent the mean and standard deviation of new data. Also, variables (columns) with large standard deviation are now down-weighted relative to variables with small standard deviation. This changes the relative sum-of-squares for the preprocessed data and the first PC will now capture the largest sum-of-squares relative to the mean of the weighted matrix $\mathbf{X}_{auto}$.

Wrapping it all together: Although outside of the scope of the present introduction, it should be noted that some preprocessing methods do not operate down the rows but instead operate across the columns. For these methods, estimates such as the mean and standard deviation might not be estimated from the data. However, these methods most often include settings or parameters that dictate how they operate and it is important that all the data are treated similarly. As a result, the preprocessing settings are a part of the model just like the estimated means and standard deviations. Therefore, estimated preprocessing parameters and settings for the preprocessing are all a part of the model established during the "calibration" step, and these parameters and settings are stored as a part of the model. Subsequently, during the model application step the preprocessing parameters are applied to new data. The two step model "calibration" and model "application" includes preprocessing as well as data modeling such as PCA. It should also be clear that preprocessing can change the focus of the data modeling procedure. For example, PCA always captures the most sum-of-squares in the first PC. However, the different preprocessing methods examined above changed what sum-of-squares was the focus of the PCA decomposition. It is in this way that preprocessing can be used to tune what variance is captured by the PCA or PLS model.

This brief introduction described how preprocessing is calibrated (based on calibration data) and applied (to both the calibration and new test data). A more detailed discussion of mean-centering and autoscaling for PCA can be found in Wise and Gallagher (1996).

H. Martens, M. Høy, B.M. Wise, R. Bro and P.B. Brockhoff, "Pre-whitening of data by covariance-weighted pre-processing," *Journal of Chemometrics*, **17**(3), 153-165, March 2003.

Gallagher, N.B, Blake, T.A., Gassman, P.L., "Application of Extended Inverse Scatter Correction to mid-Infrared Reflectance Spectroscopy of Soil," *J. Chemometr.*, **19**(5-7), 271-281 (2005).

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).