EIGENVECTOR RESEARCH INCORPORATED
Research, Training, and Software

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

Generalized Weighting to Account for Sampling Artifacts

Neal B. Gallagher and Jeremy M. Shaver

Key words: Variable weighting, pattern recognition, generalized least squares

Introduction: Variable weighting strategies can be used to enhance pattern recognition. The most common method is autoscaling where each variable has its mean removed and the result is scaled by the standard deviation. Each variable (column of $\mathbf{X}$) is scaled to unit variance. Autoscaling is a data preprocessing that can be represented by the following transformation

$$\hat{\mathbf{X}} = \left(\mathbf{I} - \tfrac{1}{M}\mathbf{1}\mathbf{1}^T\right)\mathbf{X}\mathbf{W}^{-1/2}$$

where $\mathbf{X}_{M\times N}$ is the original data matrix and $\hat{\mathbf{X}}_{M\times N}$ is the preprocessed data. The term in parentheses mean centers the data. For autoscaling $\mathbf{W}^{1/2}$ is a diagonal matrix with the standard deviation of each variable on the diagonal. The effect of autoscaling is to give each variable equal weighting–even if that variable contains only noise.

Another useful weighting strategy weights each variable by its noise level. This can be used when the measurement noise is white (random and uncorrelated). It can also be used if the noise has different levels for each variable. In this case, $\mathbf{W}$ corresponds to a diagonal matrix with the noise *variance* of each variable on the diagonal. Replicate measurements are required to allow characterization of the different noise levels.

Generalized weighting can be used to account for correlated noise and interferences (the term "clutter" includes both interference and random noise). In this case, $\mathbf{W}$ is no longer diagonal and the weighting is directly related to that used in generalized least squares (GLS). Replicate measurements are required to characterization the clutter. For example, assume there are J different objects and a set of M_j replicates for each object is measured with $j = 1,\mathrm{K}\ ,J$ and $M = \sum_{j=1}^{J} M_j$. The weighting matrix $\mathbf{W}$ is estimated as

$$\mathbf{W} = \sum_{j=1}^{J} \tfrac{1}{M_j-1}\mathbf{X}_j^T\left(\mathbf{I} - \tfrac{1}{M_j}\mathbf{1}\mathbf{1}^T\right)\mathbf{X}_j \ .$$

The term $\mathbf{W}^{-1/2}$ de-weights directions with high correlation and high variance more than directions with low correlation and low variance. An example is shown with the ARCH data set.

Experimental: The data set $\mathbf{X}_{75\times 10}$ consists of X-ray fluorescence measurements for ten elements (ppm) in obsidian samples (see the ARCH data set in PLS_Toolbox and Ref. 1). Samples 1 to 63 are from known quarries labeled K, BL, SH, and ANA. Each quarry has multiple replicates used to estimate $\mathbf{W}$ with $J = 4$. Samples 64 to 75 are samples from unknown quarries that we wish to classify.

Results and Discussion: Principal components analysis (PCA) results for autoscaled data are shown in Figure 1 for scores on principal component (PC) 2 versus PC 1. The unknowns were not included during calibration and were projected onto the PCA model after the PCA decomposition. The figure shows that the quarries cluster and some classification of the unknowns could be made using this plot (note: Figure 1 ignores other PCs and Q residuals which can significantly alter the classification).

Figure 2 shows the preprocessing GUI used to apply GLS weighting. The preprocessing assumes that each class is a single object and that multiple class members are replicates. Figure 3 shows the PCA results after GLS weighting. The quarries now cluster along PCs 1 and 2, and the clusters are tighter. It's easier to see where the unknowns lie with respect to each cluster.

In another example the data are normalized using a 1-norm prior to weighting (Figure 4) and the results are shown in Figure 5. In this case, the clustering is more easily viewed in a 3D plot.

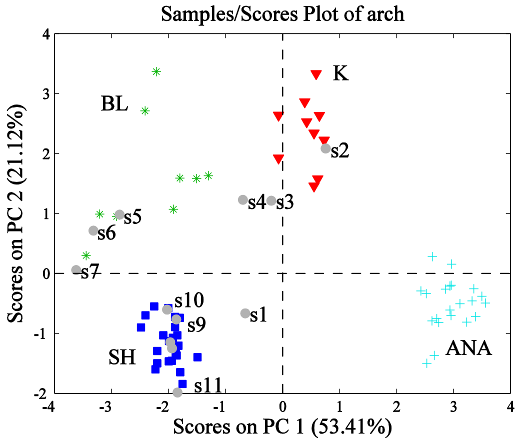
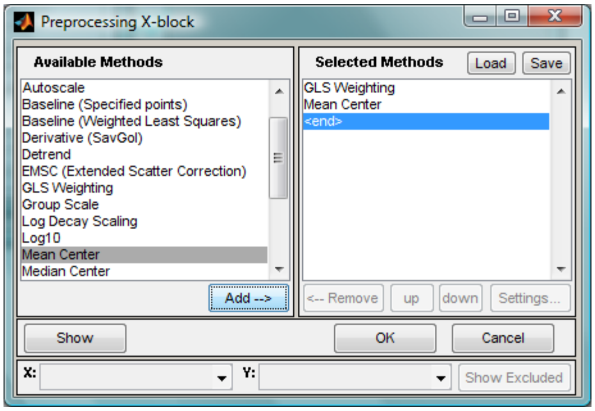


Figure 1: Scores on PC 2 versus PC 1 for autoscaled data. Known quarry samples are labeled (*) BL, (⬆) K, (◇) SH, (+) ANA, (✧) unknowns.



GLS Weighting
Mean Center
<end>

Figure 2: Preprocessing order for GLS weighting.

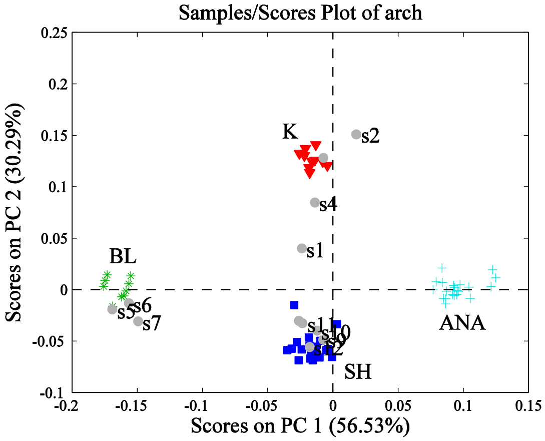
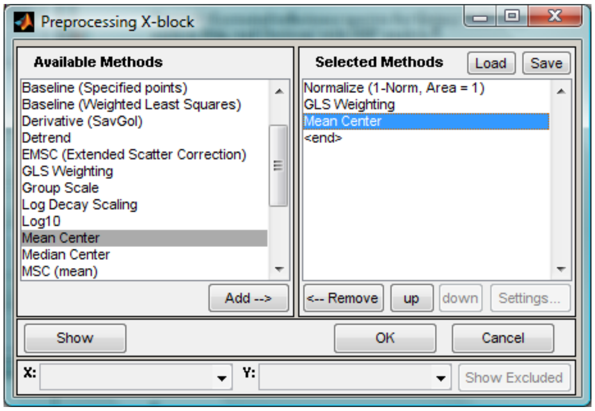


Figure 3: Scores on PC 2 versus PC 1 for GLS weighted data. Known quarry samples are labeled (*) BL, (⬆) K, (◇) SH, (+) ANA, (✧) unknowns.



Normalize (1-Norm, Area = 1)
GLS Weighting
Mean Center
<end>

Figure 4: Preprocessing order for normalization followed by GLS weighting.

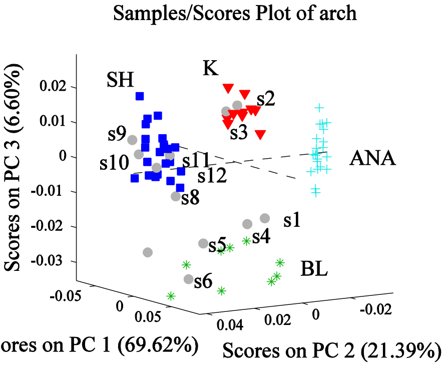


Figure 5: Scores on PC 2 versus PC 1 for normalized data followed by GLS weighting. Known quarry samples are labeled (*) BL, (⬆) K, (◇) SH, (+) ANA, (✧) unknowns.

Conclusions: Generalized least squares weighting can be used to enhance "signal-to-clutter" making pattern recognition easier. However, the practitioner must be careful not to include signal in the weighting matrix and to validate their results.

References:

[1] B. R. Kowalski, T. F. Schatzki, F. H. "Stross "Classification of archaeological artifacts by applying pattern recognition to trace element data," *Anal. Chem.* 1972; **44**(13): 2176–2180.