EIGENVECTOR RESEARCH INCORPORATED
Research, Training, and Software

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

Using Partial Least Squares Discriminant Analysis (PLSDA) for Linear Discriminant Analysis (LDA) Part I: Application

Neal B. Gallagher, Donal O'Sullivan

2023

Key words: PLSDA, LDA, class-centroid centering, generalized weighting

Introduction: An example is shown where partial least squares discriminant analysis (PLSDA) is converted to Fisher's linear discriminant analysis (LDA) when PLSDA uses class-centroid centering followed by generalized weighting preprocessing. Part II provides the theoretical development to show thee mathematical equivalence.[1]

Example Data: The Arch data set contains XRF measurements for ten elements in obsidian samples from four quarries of known origin and several samples from archeological sites of unknown origin.[2] The data set can be found in PLS_Toolbox and Solo demo datasets.[3] For the examples shown here, the data set was reduced to three classes corresponding to three quarries: 1) Koncoti, 2) Sugar Hill and 3) Annadel.

LDA Comparisons: In LDA, the objective is to find discriminator vectors that maximize the inter-class variance, $\mathbf{\Sigma}_{\text{inter}}$, to intra-class variance, $\mathbf{\Sigma}_{\text{intra}}$.[1] For the three-class problem, the measured XRF spectra are collected into a data matrix $\mathbf{X}$ with a corresponding matrix of dummy variables $\mathbf{Y}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} \mathbf{1}_{M_1} M_1^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{M_2} M_2^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{M_3} M_3^{-1} \end{bmatrix} \quad (1)$$

where $\mathbf{1}_{M_j}$ is a $M_j \times 1$ column vector of ones and M_j is the number of samples in each class. The definition of $\mathbf{Y}$ accounts for unbalanced designs where the M_j are not equal. PLSDA if often performed with only mean-centering without explicitly accounting for intra-class variance. This model (referred to as PLSDA) is compared to PLSDA using class-centroid centering and generalized least squares weighting (GLSW)[4] that explicitly accounts for intra-class variance (referred to as PLSDAp). The class-centroid is defined as the mean of the individual class means and is not affected by the number of samples in each class. This preprocessing enables PLSDA find a good solution using fewer latent variables. Additional LDA algorithms used principal components analysis (PCA) and an eigenvalue decomposition (Eig) based on a generalized eigenvalue problem.[1,5] In each case, the XRF data were autoscaled and two latent variables were used. In addition, PLSDAp and PCA also used class-centroid centering and GLSW. The Eig model did not use weighting but the $\mathbf{\Sigma}_{\text{intra}}$ matrix was regularized to match the GLSW approach.

Results: Model fits to $\mathbf{Y}$ (class-centroid centered) provide a direct performance comparison and are listed in Table 1. RMSE for PLSDAp, PCA and Eig differ only at the fifth decimal point (PCA and Eig had identical results). Figure 1 shows fits for Class 2 vs Class 1 for PLSDA (top) and PLSDAp (bottom). Differences in fits for PCA and Eig compared to PLSDAp were not visually discernable and are not shown. Comparison shows that PLSDAp has tighter classes compared to PLSDA attributed to GLSW explicitly accounting for intra-class variance.

In a second example, 1,000 simulated samples for Annadel were added to Class 3 resulting in an unbalanced design. Figure 3 shows that the model origin is now dominated by Class 3 but Figure 2 (bottom) shows that the class-

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

centroid is not severely affected and the results are similar to those for Figure 1 (bottom).

Table 1. Root mean square error of fits to **Y**.

RMSE	Konocti	Sugar Hill	Annadel
PLSDA	0.18622	0.11901	0.10092
PLSDAp	0.05406	0.04736	0.05746
PCA	0.05408	0.04735	0.05748
Eig	0.05408	0.04735	0.05748

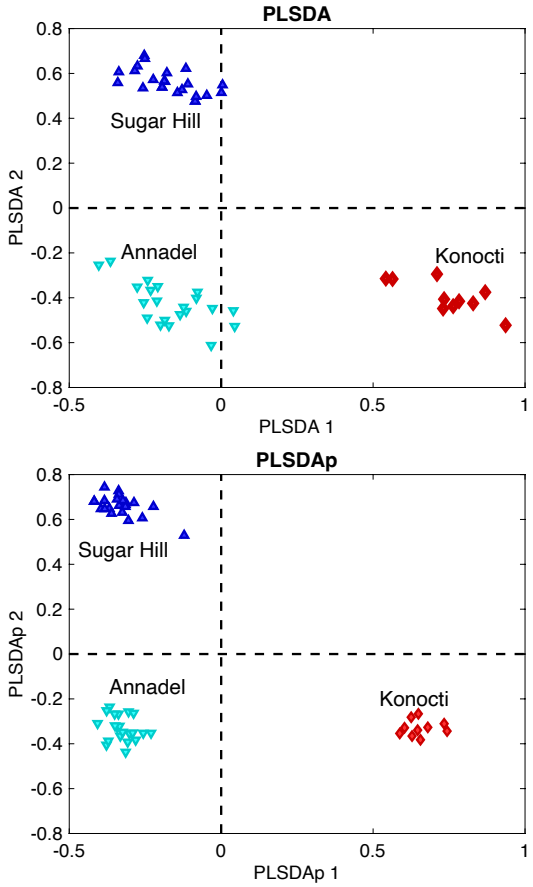

Figure 1. Fits for Class 2 vs Class 1: (top) PLSDA, (bottom) PLSDAp.

Conclusion:

The example showed that PLSDA can be converted to Fisher's LDA when using class-centroid centering followed by generalized weighting preprocessing. Results showed that algorithms PLSDAp, PCA and Eig yield similar results for LDA.

References:

[1] Using Partial Least Squares Discriminant Analysis (PLSDA) for Linear Discriminant Analysis (LDA) Part II: Theory. White Paper (2003).

[2] Kowalski, BR, Schatzki, TF, Stross, FH, *Anal. Chem.*, **44**(13), 2176–2180 (1972).

[3] PLS_Toolbox and Solo Version 9.2.1, Manson, WA USA, Eigenvector Research, Inc.

[4] Martens, H, Høy, M, Wise, BM, Bro, R, Brockhoff, PB, *J. Chemom.*, **17**(3), 153-165 (2003). doi: 10.1002/cem.780.

[5] MATLAB Version 9.14 (2023a), Natick, MA USA, The MathWorks Inc. (see the "eig" function).

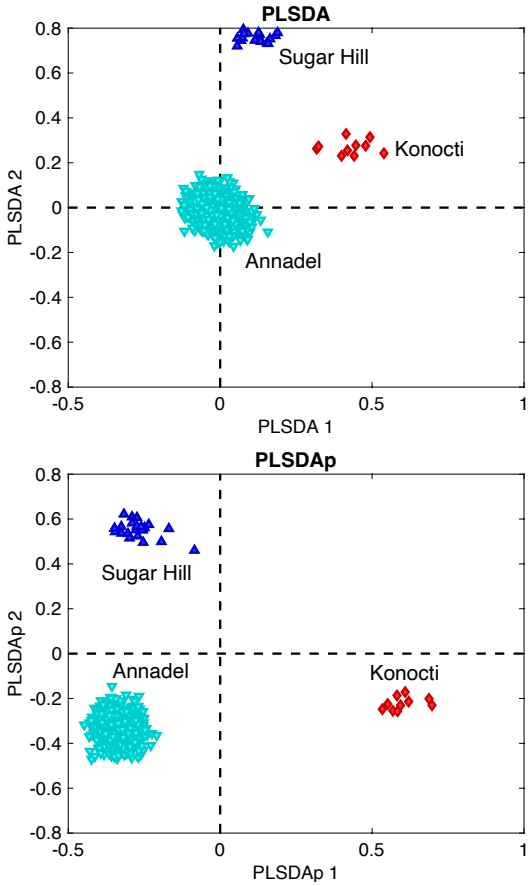

Figure 2. Fits for Class 2 vs Class 1 with an unbalanced design: (top) PLSDA, (bottom) PLSDAp.