Graphical tools for quadratic discriminant analysis

Iain Pardoe* Xiangrong Yin† R. Dennis Cook‡

November 21, 2006

Abstract

Sufficient dimension reduction methods provide effective ways to visualize discriminant analysis problems. For example, Cook and Yin (2001) showed that the dimension reduction method of sliced average variance estimation (SAVE) identifies variates that are equivalent to a quadratic discriminant analysis (QDA) solution. This article makes this connection explicit to motivate the use of SAVE variates in exploratory graphics for discriminant analysis. Classification can then be based on the SAVE variates using a suitable distance measure. If the chosen measure is Mahalanobis distance, then classification is identical to QDA using the original variables. Just as canonical variates provide a useful way to visualize linear discriminant analysis (LDA), so SAVE variates help to visualize QDA—this would appear to be particularly useful given the lack of graphical tools for QDA in current software. Furthermore, while LDA and QDA can be sensitive to nonnormality, SAVE is more robust.

Key words: canonical variates, classification, dimension reduction, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), sliced average variance estimation (SAVE).

1 Introduction

Discriminant analysis seeks to classify a categorical outcome variable Y, an indicator with C classes, using values of a $p \times 1$ vector of features $\mathbf{X}$. Visualization and graphics can play an important role in understanding discriminant analysis. For example, a classical linear discriminant analysis (LDA) of a three-class outcome with two or more feature variables results in two canonical variates (or linear discriminants) which are linear combinations of the features, effectively reducing the feature space from p to two dimensions. A scatterplot of these two variates, with the data points marked by class, shows the effectiveness of the discrimination between classes. Many major statistical software programs can produce such plots.

Quadratic discriminant analysis (QDA) extends LDA by allowing the intraclass covariance matrices to differ between classes, so that discrimination is based on quadratic rather than linear functions of $\mathbf{X}$. With QDA, however, there are no natural canonical variates and no general methods for displaying the analysis graphically. When there are just two feature variables, it is possible to visualize the quadratic classification regions in a scatterplot, but there is no clear extension to analyses with three or more features.

In this article, we show how the dimension reduction technique of sliced average variance estimation (SAVE) provides a useful way to define reduced dimension variates when the covariance

*Department of Decision Sciences, Lundquist College of Business, University of Oregon, Eugene, OR 97403.
†Department of Statistics, 204 Statistics Building, University of Georgia, Athens, GA 30605.
‡School of Statistics, 224 Church Street S.E., University of Minnesota, Minneapolis, MN 55455.

matrices differ across classes. This method may be particularly appropriate at the outset of an analysis because it does not require a prespecified classification rule. Nevertheless, we show that classifying using Mahalanobis distance in combination with the SAVE variates leads to the usual normal theory QDA discriminant rule based on the original feature variables. Thus, SAVE can be used to construct reduced dimension scatterplots for QDA, analogous to canonical variate scatterplots for LDA. Furthermore, these SAVE variates can be useful even in nonnormal cases, because estimation of these variates does not require a classification rule and is less sensitive to normality than parametric methods.

The outline of the article is as follows. To provide further motivation for our proposal, we discuss in Sections 2 and 3 the general approaches of statistical software programs to graphics for LDA and QDA and review some other methods for discriminant analysis. In Section 4, we derive some basic results for SAVE and make an explicit connection to QDA. This section motivates the proposed graphical tools: use LDA and SAVE in concert as an easily implemented graphical diagnostic method that can be used for visualization at the outset of an analysis. Fisher's (1936) classic iris data provides a running example for Sections 2–4. Section 5 contains two further examples illustrating the use of SAVE graphics in discriminant analysis. The first example concerns classification of radar returns from the ionosphere, where SAVE identifies clear variability differences between different types of return. The second example shows how SAVE can perform well in nonnormal settings with multiple outliers. Finally, Section 6 contains a discussion, while justifications for the theoretical results are contained in the appendix.

2 Graphics for LDA

Let $\mathbf{Z}$ be the standardized vector of feature variables, $\mathbf{Z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, where $\mathrm{E}(\mathbf{X})$ is the mean vector and $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the positive definite covariance matrix. As indicated by Cook (1998, Proposition 6.1), there is no loss of generality in working on the $\mathbf{Z}$ scale, so we shall use $\mathbf{Z}$ instead of $\mathbf{X}$ throughout the rest of the article. To further facilitate presentation in this article, we shall assume equal prior class probabilities when considering classification rules—our results are unchanged but easier to read with this assumption.

In the classical form of LDA, the conditional distribution of $\mathbf{Z}|Y$ is assumed to be normal with a common intraclass covariance matrix, $\mathbf{W}$. The discriminant function is a linear function in $\mathbf{Z}$:

$$\mathcal{L}_c(\mathbf{Z}_0) = \boldsymbol{\mu}_c^T \mathbf{W}^{-1} \mathbf{Z}_0 - \frac{1}{2} \boldsymbol{\mu}_c^T \mathbf{W}^{-1} \boldsymbol{\mu}_c,$$

where $\boldsymbol{\mu}_c$ is the c-th class mean ($c = 1, \ldots, C$). LDA then classifies $\mathbf{Z}_0$ in the class that maximizes $\mathcal{L}_c$. This is equivalent to classifying using Mahalanobis distance for the first $r = \min(p, C-1)$ eigenvectors of $\mathbf{B}$ relative to $\mathbf{W}$, where $\mathbf{B}$ is the population between-class covariance matrix. These relative eigenvectors are also called (population) canonical variates or linear discriminants. It is common practice to plot sample versions of the first two of these canonical variates to allow visualization of the discriminant analysis, and it is straightforward in most statistical software programs to produce such plots.

For example, Fisher's classic iris data consist of four length measurements on 50 specimens of each of three species of iris. Figure 1 displays the first two sample canonical variates constructed by using the *R* program. Since the intraclass covariances of the canonical variates are identity matrices, this plot is scaled so that each axis covers the same range. The plot suggests that almost all of the linear discrimination between the three classes comes from the first canonical variate. The *setosa* species clearly separates from the *versicolor* and *virginica* species, but there is less separation

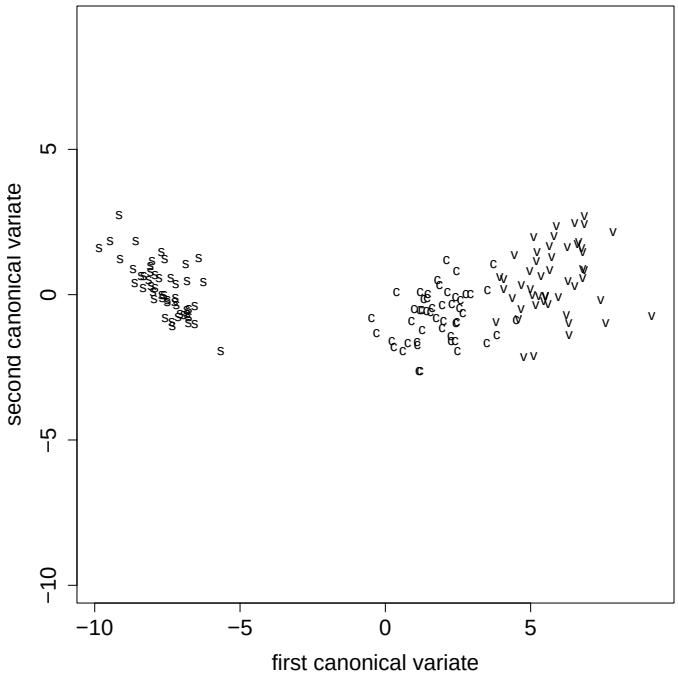

Figure 1: *R* LDA graphical display for the iris data. The scatterplot graphs the first sample canonical variate against the second, with the data points marked by class: s's are *iris setosa*, c's are *iris versicolor*, and v's are *iris virginica*.

between the latter two. Fisher's original graphical summary consisted of three histograms, one for each species, along an axis representing the first "discriminating linear function". He displaced the histogram for *versicolor* to avoid overlap with the histogram for *virginica*.

3 Graphics for QDA

3.1 Normal theory QDA

In the classical form of QDA, the distribution of $\mathbf{Z}|Y$ is assumed to be normal without requiring the intraclass covariance matrices to be constant. The full discriminant function $\mathcal{Q}_c^f$ is then a quadratic function in $\mathbf{Z}$:

$$\mathcal{Q}_c^f(\mathbf{Z}_0) = -(p/2)\log(2\pi) - (1/2)\log|\mathbf{W}_c| + \mathcal{Q}_c(\mathbf{Z}_0)$$

where

$$\mathcal{Q}_c(\mathbf{Z}_0) = \boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}\mathbf{Z}_0 - \frac{1}{2}\boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\mathbf{Z}_0^T\mathbf{W}_c^{-1}\mathbf{Z}_0,$$

and $\boldsymbol{\mu}_c$ and $\mathbf{W}_c$ are the c-th class mean vector and intraclass covariance matrix respectively. We call $\mathcal{Q}_c$ the *quadratic direction function*, because it contains all of the reductive information and can thus be used as a basis for choosing plotting directions. If all $\mathbf{W}_c$ are equal, then $\mathcal{Q}_c$ reduces to $\mathcal{L}_c$ (minus a constant). The advantage of $\mathcal{Q}_c$ in allowing $\mathbf{W}_c$ to vary is that information for discrimination may come from the intraclass covariance matrices, $\mathbf{W}_c$ $(c = 1, \ldots, C)$, in addition to the intraclass means, $\boldsymbol{\mu}_c$.

Historically, QDA "canonical variates" have proved elusive. Several authors have considered dimension reduction for quadratic discrimination in normal populations with different covariance

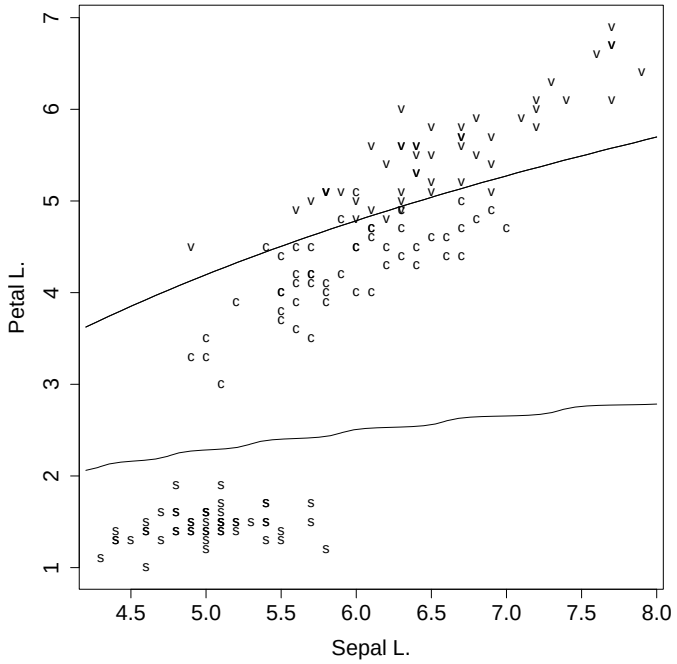

Figure 2: *R* QDA graphical display with two feature variables for the iris data. The scatterplot graphs the first feature against the second, with the data points marked by class. The curved boundaries mark the quadratic classification regions.

matrices. Nevertheless, there appears to be no standard canonical variate analysis for QDA as there is for LDA. And because there is no obvious way to define canonical variates, there are no good graphical approaches to QDA in commercial statistical software. However, it is possible to make headway by building a framework in which SAVE can serve as a reduced dimension variate analysis for QDA.

Before developing this framework in Section 4, we provide further motivation for our proposal by surveying the graphical aspects of QDA that are available in some of the major statistical software programs. While the methods available in major programs do not reflect the totality of the literature, they do reflect what might be regarded as the state of the art in "standard" practice. We again illustrate using Fisher's iris data.

3.2 Major software packages

Venables and Ripley (2002) have a `qda` function for quadratic discriminant analysis in **S-PLUS** and **R**. They note that "the boundaries of the decision regions are quadratic surfaces in [feature] space," and provide an example using two feature variables and three classes. Their code can be used to plot the quadratic surfaces whenever the number of features is two: the outcome using just two of the iris length measurements is shown in Figure 2.

The function `discrim` in S-PLUS for Microsoft Windows is based on Venables and Ripley's functions, `qda` and `lda`. The function has a plot method that displays all two-dimensional scatterplots of the original features, with the data points marked by the class to which they belong. No other graphical techniques are suggested. A "canonical" model can be specified in `discrim` and is useful for dimension reduction. However, only a homoscedastic covariance structure is permitted (giving rise to the standard LDA canonical variates).

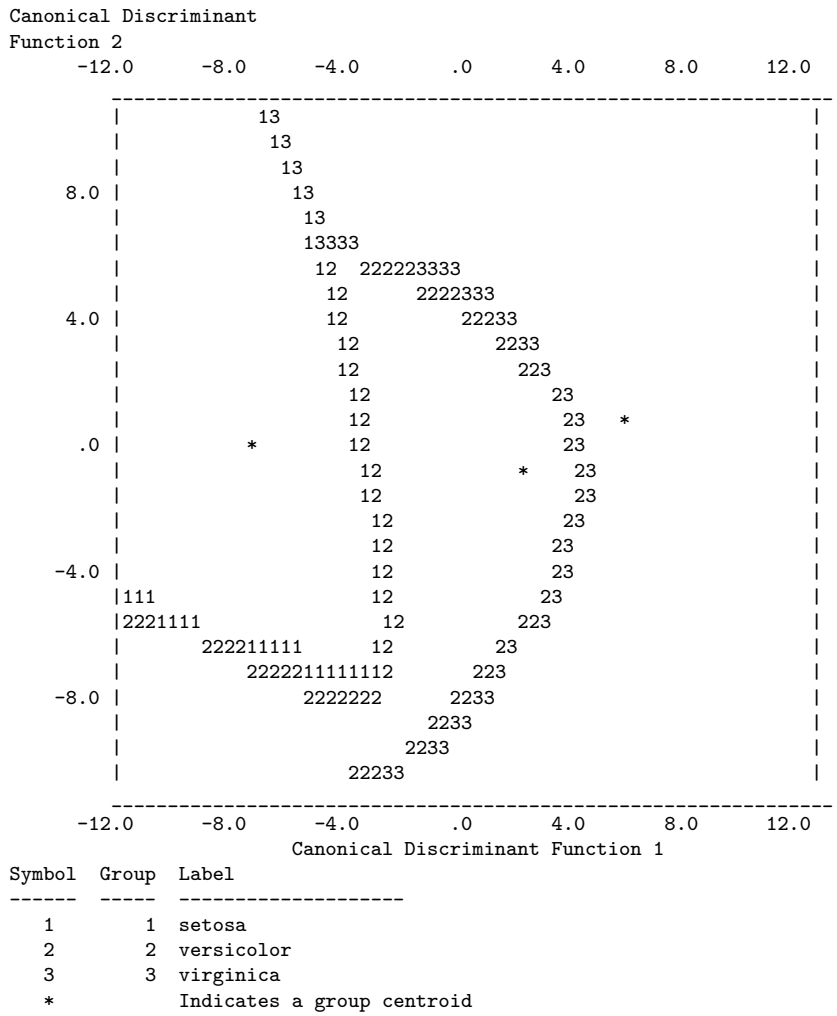

Figure 3: SPSS territorial map for the iris data. The class symbols denote the boundaries of the classification regions based on an analysis using separate class covariances for the standard canonical variates.

Khattree and Naik (2000) described how to use the `DISCRIM` procedure in **SAS** to construct a plot showing QDA decision regions when the number of feature variables is two. The resulting graphical display using just two of the iris measurements has the same essential character as Figure 2. Khattree and Naik note that "such plots are not possible for higher dimensional datasets." They also discuss canonical discriminant analysis in terms of dimension reduction. They point out, however, that this is applicable only when all populations have a common covariance structure.

SPSS does not perform classical QDA, but rather classification using intraclass covariance matrices of the (standard) canonical discriminant variates, not those of the original features. One graphical display offered using this routine is a "territorial map"—a plot showing the decision regions in the space spanned by the first two canonical variates. Figure 3 displays such a plot for the iris data. The data can also be plotted directly in a plot of the (first two) canonical variates. Somewhat confusingly, this plot displays in an identical fashion whether the intraclass covariance matrices are assumed to be common or different. Thus, for the iris data—even for the "separate-classes covariance matrices" analysis—this plot is similar to Figure 1.

Minitab offers QDA as part of its multivariate analysis routines but makes no suggestions for graphical displays. Sall et al. (2004) illustrated analysis of the iris data in **SAS JMP** using just two of the length measurements with bivariate density curves. Again, there is no suggestion for a graphical display for datasets with more than two features. **Stata** has a discriminant analysis module that appears to provide the same output and options as SPSS. **Statistica** performs only LDA, although QDA techniques are used in classification tree calculations (see Loh and Shih, 1997). **Systat** offers QDA, but makes no suggestions for graphical displays. However, canonical variates are calculated as part of the procedure and can be displayed visually. When there are two canonical variates, the result is a "canonical scores plot" in which the axes are the canonical variates and the points are the canonical variate scores marked by class; this plot is similar to Figure 1 above. With more than two canonical variates, a scatterplot matrix of the first three canonical variates results. As with SPSS, and again somewhat confusingly, these displays are the same whether the original discriminant analysis is quadratic or linear.

3.3 Other dimension reduction methods

Other dimension reduction methods for discriminant analysis include those proposed by Young et al. (1987), Fukunaga (1990), and Schott (1993), who used both mean and covariance differences to develop "symmetric methods" for classification (see Hennig, 2004). Hennig (2004) also developed "asymmetric methods" for classification, particularly aimed at two classes consisting of a homogeneous group and a nonhomogeneous group; the `fpc` package in *R* implements these methods, as well as the methods of Young et al. and Fukunaga. Zhu and Hastie (2003) developed a nonparametric method for discriminant analysis that uses a likelihood-based interpretation of Fisher's LDA criterion to find important discriminant directions. In Sections 4.3 and 4.4 we return to discuss these methods, after first developing a framework for SAVE in the context of quadratic discriminant analysis.

4 A graphical solution: SAVE

4.1 SAVE and QDA

Sliced average variance estimation (Cook and Weisberg, 1991; Cook and Yin, 2001) is a versatile dimension reduction technique that can be used to construct reduced dimension variates for QDA. The population SAVE kernel matrix is defined as $\mathbf{M} = \mathrm{E}(I - \mathbf{W}_{\mathbf{Z}|Y})^2$, where $\mathbf{W}_{\mathbf{Z}|Y}$ is the conditional covariance matrix of $\mathbf{Z}$ given Y. In discriminant analysis this is the intraclass covariance matrix, $\mathbf{W}_{\mathbf{Z}|Y=c} = \mathbf{W}_c$. Letting $\mathbf{M} = \mathbf{\Gamma D \Gamma}^T$ denote the spectral decomposition of $\mathbf{M}$, the population SAVE variates are defined as $\mathbf{\Gamma}^T\mathbf{Z}$ and are ordered by the eigenvalues on the diagonal of $\mathbf{D}$.

The sample version of the SAVE matrix is $\widehat{\mathbf{M}} = \sum_{c=1}^{C} \frac{n_c}{n}(I - \widehat{\mathbf{W}}_c)^2$, where n_c is the sample size for group c, $n = \sum_{c=1}^{C} n_c$, and $\widehat{\mathbf{W}}_c$ is the sample covariance matrix for $\mathbf{Z}$ in class c. The sample SAVE variates $\widehat{\mathbf{\Gamma}}^T\widehat{\mathbf{Z}}$ are constructed from the spectral decomposition of $\widehat{\mathbf{M}} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{D}}\widehat{\mathbf{\Gamma}}^T$, where $\widehat{\mathbf{Z}} = \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}})$.

The following result, obtained by Cook and Critchley (2000; see also Cook and Yin, 2001), shows that the population SAVE variates depend on the same quantities as the quadratic direction function, $\mathcal{Q}_c$, from Section 3.1.

Proposition 1

$$\begin{aligned} \text{span}(\mathbf{M}) &= \text{span}\,(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C, \mathbf{W}_2 - \mathbf{W}_1, \ldots, \mathbf{W}_C - \mathbf{W}_{C-1}) \\ &= \text{span}\,(I - \mathbf{W}_c, \quad c = 1, \ldots, C). \end{aligned}$$

In what follows, we go beyond this subspace connection and link the methods of SAVE and QDA directly. Consider using the SAVE variates to calculate the corresponding Mahalanobis distances between the point to be classified and the class means. We shall prove that, apart from an unimportant constant, this is the same as using the Mahalanobis distances between the point to be classified and the class means using the original feature variables.

Suppose that we wish to classify the point $\mathbf{Z}_0$. To do this we need only the SAVE variates corresponding to the nonzero eigenvalues of $\mathbf{M}$. Thus, we partition both $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2)$, where $\boldsymbol{\Gamma}_1$ is a $p \times p_1$ matrix and $\boldsymbol{\Gamma}_2$ is a $p \times (p - p_1)$ matrix, and

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\mathbf{D}_1$ is a $p_1 \times p_1$ diagonal matrix. Let $\mathbf{U} = \boldsymbol{\Gamma}_1^T \mathbf{Z}$, so that $\mathbf{U}_0 = \boldsymbol{\Gamma}_1^T \mathbf{Z}_0$ and $\boldsymbol{\nu}_c = \boldsymbol{\Gamma}_1^T \boldsymbol{\mu}_c$. Then the Mahalanobis distance from the point $\mathbf{U}_0$ to the c-th class is

$$D_c^2 = (\boldsymbol{\nu}_c - \mathbf{U}_0)^T (\boldsymbol{\Gamma}_1^T \mathbf{W}_c \boldsymbol{\Gamma}_1)^{-1} (\boldsymbol{\nu}_c - \mathbf{U}_0).$$

Recall the definition of the quadratic direction function:

$$\mathcal{Q}_c(\mathbf{Z}_0) = \boldsymbol{\mu}_c^T \mathbf{W}_c^{-1} \mathbf{Z}_0 - \frac{1}{2} \boldsymbol{\mu}_c^T \mathbf{W}_c^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \mathbf{Z}_0^T \mathbf{W}_c^{-1} \mathbf{Z}_0.$$

Now we are ready to present the main results.

Lemma 1 *Assume that all intraclass covariance matrices, $\mathbf{W}_c$, are nonsingular and distinct. Then $D_c^2 = -2\mathcal{Q}_c + K$, where K is a constant that does not depend on c.*

Let

$$\mathcal{Q}_c^f(\mathbf{U}_0) = -(p_1/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}_1^T \mathbf{W}_c \boldsymbol{\Gamma}_1| - (1/2) D_c^2$$

be the full quadratic discriminant function corresponding to the reduced variables $\mathbf{U}$. Then, based on Lemma 1, we have the following:

Proposition 2 $\mathcal{Q}_c^f(\mathbf{Z}_0) = \mathcal{Q}_c^f(\mathbf{U}_0) + (1/2)K + [(p_1 - p)/2]\log(2\pi)$.

This proposition shows that maximizing $\mathcal{Q}_c^f(\mathbf{Z}_0)$ is equivalent to maximizing the corresponding $\mathcal{Q}_c^f(\mathbf{U}_0)$. In other words, Lemma 1 and Proposition 2 show that the population SAVE variates together with classification by Mahalanobis distance leads to the usual normal theory QDA discriminant rule based on all the feature variables. Thus, the SAVE variates can serve as QDA "reduced dimension variates." Furthermore, since SAVE does not depend on any normality assumptions, the SAVE variates may also work in nonnormal cases where QDA classification performs poorly.

While the above link between SAVE and QDA is much like the link between canonical variates and LDA, there is an important difference. Whereas QDA provides a fixed classification rule, SAVE merely performs dimension reduction to find new reduced dimension variates that are linear combinations

of the original features. We can then calculate Mahalanobis distances using the SAVE variates—equivalent to QDA—or we can use any other method of classifying based on the SAVE variates.

When $\mathbf{M}$ is full rank, no linear reduction is possible and the number of relevant reduced dimension variates is the same as the number of original features. However, the SAVE variates are ordered by the eigenvalues of $\mathbf{M}$, and useful graphical information can often be obtained by plotting the first few SAVE variates. Additional discussion along these lines is provided in the next and subsequent sections.

4.2 SAVE graphics for discriminant analysis

In our experience, plots of the first two or three SAVE variates with points marked according to class often provide valuable insights about a discrimination problem. Since SAVE, like QDA, can sometimes be less powerful than LDA in picking up location differences, a combined plot of the first one or two LDA canonical variates together with the first one or two SAVE variates can also be useful. SAVE searches over a larger space than LDA and consequently it may find discriminatory information beyond that provided by LDA, particularly when faced with strong scale effects. In other words, the first few SAVE variates sometimes focus on intraclass covariance differences and downplay location differences readily detected by LDA. In such situations, location differences may show in later SAVE variates, depending on the strength of the scale differences. This phenomenon accounts for some of the past findings on the behavior of SAVE, for example, sensitivity to outliers (Sheather and McKean, 2001) and over-emphasizing high-order information when first-order differences dominate (Zhu and Hastie, 2003; Peltonen and Kaski, 2005). Combining SAVE and LDA graphics in a single plot may be useful, since we typically wish both to gauge location differences alone and to contrast location and scale differences.

The first few SAVE variates contain relevant information but, depending on the complexity of the problem, there could also be relevant graphical information in subsequent variates. For this reason it can be helpful to infer about the number, $d = \dim(\text{span}(\mathbf{M}))$, of relevant SAVE variates, those with nonzero eigenvalues in the population. Cook and Yin (2001) described a permutation test for this purpose which generalizes a suggestion by Cook and Weisberg (1991). Large sample tests for SAVE were developed by Cook and Ni (2005), and Shao, Cook and Weisberg (2006).

A SAVE analysis of Fisher's iris data is straightforward to run in the free regression program *Arc* (see Appendix A of Cook and Weisberg, 1999) or using the `dr` package in *R* (Weisberg, 2002). The permutation test for SAVE, also available in *Arc* and *R*, indicates two significant variates:

$$\begin{aligned} \text{first SAVE variate} &= -.17\,\text{Sepal L.} - .42\,\text{Sepal W.} + .52\,\text{Petal L.} + .73\,\text{Petal W.}, \\ \text{second SAVE variate} &= .07\,\text{Sepal L.} + .02\,\text{Sepal W.} + .37\,\text{Petal L.} - .93\,\text{Petal W.} \end{aligned}$$

This contrasts with the first two canonical variates found by LDA:

$$\begin{aligned} \text{first LDA variate} &= -.21\,\text{Sepal L.} - .39\,\text{Sepal W.} + .55\,\text{Petal L.} + .71\,\text{Petal W.}, \\ \text{second LDA variate} &= -.01\,\text{Sepal L.} - .59\,\text{Sepal W.} + .25\,\text{Petal L.} - .77\,\text{Petal W.} \end{aligned}$$

The first SAVE variate is very similar to the first LDA variate, indicating that, for these data, the methods find the same dominant location differences, but the second variates are rather different.

Figure 4 shows a summary plot of the first two SAVE variates, in which SAVE finds strong location separation in its first variate and clear scale separation on the second. In particular, we can easily distinguish *iris setosa* on location alone, but we need both location and scale to distinguish the remaining two species. This latter distinction was not readily apparent from the quadratic discriminant analysis and graphics available from the other major statistical software programs considered in Section 3.2.

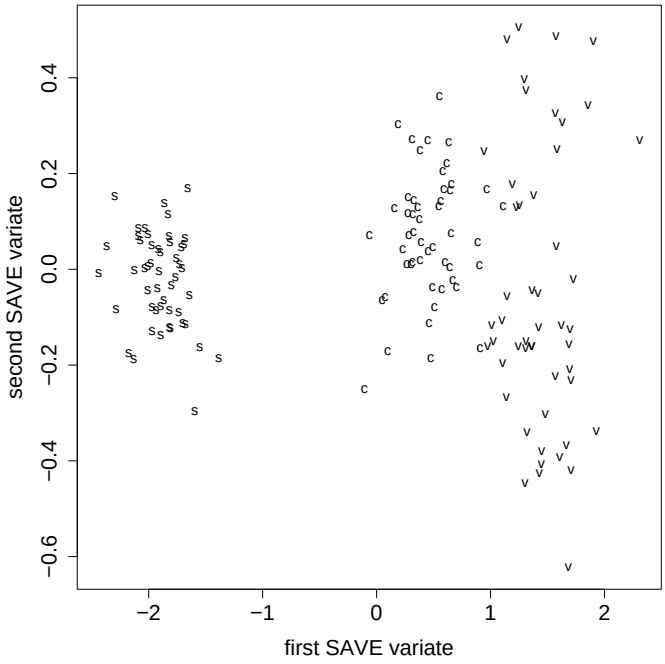

Figure 4: SAVE summary plot for the iris data: s's are *iris setosa*, c's are *iris versicolor*, and v's are *iris virginica*.

4.3 SAVE and the likelihood ratio

Zhu and Hastie (2003) proposed a marginal log-likelihood-ratio statistic for choosing discriminant directions and showed that, with normal populations, their approach results in sequentially maximizing the objective function (Zhu and Hastie, 2003, eq. (4.4))

$$\mathrm{L}(\boldsymbol{\alpha}) = -\sum_{c=1}^{C} \frac{n_c}{n} \log \boldsymbol{\alpha}^T \widehat{\mathbf{W}}_c \boldsymbol{\alpha} \tag{1}$$

over the $p \times 1$ vector $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$. Their basic algorithm starts with the direction $\boldsymbol{\alpha}_1$ that maximizes $\mathrm{L}(\boldsymbol{\alpha})$ and then chooses the second direction $\boldsymbol{\alpha}_2$ to maximize $\mathrm{L}(\boldsymbol{\alpha})$ subject to being orthogonal to $\boldsymbol{\alpha}_1$, continuing until the desired number of directions is reached. Zhu and Hastie compared discriminant directions obtained using this objective function to those from SAVE, concluding that "... SAVE seems to over-emphasize second-order differences among the classes."

Let $\mathrm{L}^*(\boldsymbol{\alpha}) = -E(\log \boldsymbol{\alpha}^T \mathbf{W}_c \boldsymbol{\alpha})$ denote the population version of (1), where the expectation is taken over the classes, and let $\{\boldsymbol{\alpha}_j^*\}$ denote the directions that sequentially maximize L^*. Then,

Proposition 3 $\mathrm{span}(\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, \ldots, \boldsymbol{\alpha}_d^*) = \mathrm{span}(\mathbf{M})$.

In other words, the Zhu-Hastie procedure and SAVE produce the same subspace in the population, but differ in their methods of constructing an ordered basis, a basis for $\mathrm{span}(\mathbf{M})$ with ordered vectors that correspond to discriminant directions. Hernández and Velilla (2005) recently made the general point that the Zhu-Hastie method is not directly motivated by the notion of dimension reduction subspaces; Proposition 3 suggests that dimension reduction subspaces are in fact at the heart of the method.

Thinking of the basic graphical method as a marked plot of the first few reduced dimension variates, the choice of an ordered basis is not crucial if d is small, as may often happen in practice,

since then it is possible to construct fully informative plots of the first d variates, which are the only ones that are relevant. The choice of an ordered basis for graphics can be an issue otherwise. We see span($\mathbf{M}$) and the determination of an ordered basis for it as two distinct issues. The subspace is the more fundamental construction, while external criterion might be used to determine an ordered basis. With relatively large d, it could be useful to consider low dimensional plots based on different ordered bases.

To reinforce these points, consider using the marginal likelihood ratio to select d vectors simultaneously instead of sequentially. The resulting population multi-directional objective function is $\mathrm{L}^*(\mathbf{A}) = -E(\log|\mathbf{A}^T\mathbf{W}_c\mathbf{A}|)$, which is to be maximized over the $p \times d$ semi-orthogonal matrix $\mathbf{A}$, $\mathbf{A}^T\mathbf{A} = I_d$. A value of $\mathbf{A}$ that maximizes $\mathrm{L}^*(\mathbf{A})$ is not unique because, for any $d \times d$ orthonormal matrix $\mathbf{O}$, $\mathrm{L}^*(\mathbf{A}) = \mathrm{L}^*(\mathbf{AO})$. This means that in the multi-directional version of L^*, all orthogonal bases for span($\mathbf{M}$) are equivalent, and an external criterion must be used to obtain ordered discriminant directions. Zhu and Hastie obtained such directions using successive maximizations of the single-direction likelihood ratio, and this may be useful when contributions to the likelihood are of interest. SAVE constructs an ordered basis using the eigenstructure of the kernel $\mathbf{M}$ and this seems appropriate when it is desired to emphasize second-order differences in graphics. SAVE may be preferable when second-order differences dominate first-order differences, a possibility that is supported by the results of an exploratory simulation. It is also possible to construct ordered bases to emphasize first-order differences. This is in line with the previous suggestion to combine LDA and SAVE variates in graphics.

In short, we see little fundamental difference between methods based on marginal likelihood ratios and SAVE when $\mathbf{X}$ is normal within class. As long as we operate within span($\mathbf{M}$), the choice of reduced variates for graphical constructions is largely irrelevant when d is small. Otherwise, the choice can be made based on external criteria, depending on the applications context and the interests of the investigator.

4.4 Comparisons with other dimension reduction methods

Dimension reduction methods for quadratic discriminant analysis include those described in Fukunaga (1990), Hennig (2004), Schott (1993), Young et al. (1987), the more general nonparametric methods in Zhu and Hastie (2003), as well as SAVE. Of these methods, SAVE would appear to be the most fully developed and its development is continuing. In particular, SAVE:

- has no limitation on the number of classes;
- gives easily implemented plotting directions;
- is available in software packages *R* and *Arc*;
- does not require smoothing or tuning;
- provides a graphical counterpart to QDA in the same way that we have graphical counterparts to LDA;
- comes with tests for the number of reduced dimension variates containing discriminatory information, without which we may not know whether we are missing useful information in a plot of the first few variates (such tests come with the usual warning that useful information may be missed if levels of significance are over-interpreted);
- responds predictably and usefully to outliers and mixtures (see Cook and Critchley, 2000, and Sheather and McKean, 2001);

- has a recently developed sparse version (Li, 2006. This paper also develops a sparse version of LDA; see also Li and Nachtsheim, 2006);
- has been studied for 15 years now and its operating characteristics, advantages, and limitations are well-understood. SAVE has also been used in many other areas of application, such as pattern recognition (Ling et al., 2005) and micro-array data analysis (Bura and Pfeiffer, 2003).

Although the methods of Fukunaga (1990) and Hennig (2004) are available in *R*, they are restricted to just two classes. In our experience, SAVE performs as well as these methods for two-class problems. A referee pointed out that methods restricted to just two classes can be adapted to multi-class problems by, for example, producing a discriminant plot for each class.

It can be shown that the Schott (1993) and Young et al. (1987) methods estimate the same population subspace as SAVE (cf. Proposition 1), but the methodology for doing so is different. According to Schott (personal communication), software is not available to implement his method.

The column dimension of the Young et al. kernel matrix grows as the number of classes increases, while for the SAVE matrix it is always p, the number of features, whatever the number of classes. Young et al. estimate dimensionality empirically using singular value decompositions and the resulting dimension reduction depends on a Bayes classification rule assuming normality. By contrast, there are specific tests for determining dimensionality using SAVE, and dimension reduction for mean and covariance differences is possible without a particular classification rule or normality assumptions. Using Hennig's `fpc` package and Weisberg's `dr` package in *R*, graphs from SAVE and the Young et al. method are often similar when the number of classes is small and there are no major outliers. However, there is no dimension test or sparse version available for the Young et al. method, it has unknown behavior in the presence of outliers, and computing time increases with the number of classes (for particularly large problems the method can fail completely).

QDA, SAVE, and the methods of Schott, Young et al., and Zhu and Hastie are all linked at the population level by the SAVE subspace, span($\mathbf{M}$). They differ in their underlying assumptions and methods of estimation, but at their core they all rely on this subspace.

There are a large number of nonparametric methods for discriminant analysis available in the literature. For instance, whereas Section 4.3 discusses the method of Zhu and Hastie (2003) in the context of QDA, their more general method offers a flexible nonparametric approach to discriminant analysis. When computational effort is not an issue, nonparametric methods have the potential to give improved performance, outperforming SAVE in specific examples (for example, see Peltonen and Kaski, 2005). However, Hand (2006) and his discussants give a prudent view of nonparametric methods, suggesting that their potential gains over simple projective methods are often small. SAVE provides a reliable second-order method to complement LDA as the first step in a discriminant analysis, and as a useful way to explore data visually before possibly committing to the additional effort of a more complex method.

5 Further examples

5.1 Ionosphere radar returns

In ionospheric research, radar returns from the ionosphere can be classified as either suitable for further analysis or not, a time-consuming task that has typically required human intervention. Researchers at the Applied Physics Laboratory of Johns Hopkins University developed the "Ionosphere database" to study classification techniques in this context (Sigillito et al., 1989). We use this dataset here to illustrate the potential for SAVE to recover diagnostic discriminatory information.

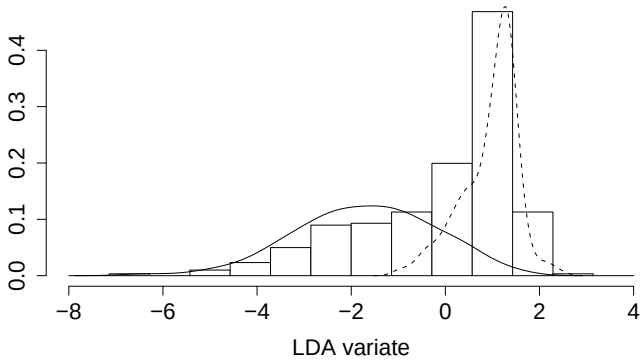

Figure 5: Histogram of the LDA canonical variate for the ionosphere data along with Gaussian kernel density estimates for the bad (left density) and good (right density) cases.

The ionosphere data were collected by 16 high-frequency antennas in Goose Bay, Labrador, Canada. "Good" radar returns show evidence of some type of structure in the ionosphere; "bad" returns have signals that simply pass through the ionosphere. 351 received signals were processed using a function of two attributes for each of 17 pulse numbers which describe the complex electromagnetic signal; thus there are 34 continuous-valued feature variables (although one is constant for this dataset). 225 of the signals were good returns, while the remaining 126 were bad.

A histogram of the LDA canonical variate, along with Gaussian kernel density estimates for the good and bad cases, is given in Figure 5. The densities show mean separation with the variance for the bad cases appearing larger than that for the good cases. We turn to SAVE to see if there is additional discriminatory information in the data. A plot of the first two SAVE variates is shown in the left panel of Figure 6. A three-dimensional plot of the first three SAVE predictors with points marked by class shows a spherical cloud of bad points with a dense concentration of good points at its center (See Cook, 1998, for background on interpretation of three-dimensional plots in this context). In effect, application of SAVE demonstrates immediately that the bad cases are substantially more variable than the good cases, and that this variability difference contains useful information for discrimination. A plot of the first SAVE variate versus the LDA variate is shown in the right panel of Figure 6. A three-dimensional plot of the first two SAVE variates and the LDA variate shows a relatively large cloud of bad points with all of the good points concentrated in a thin pencil that extends from the center of the cloud out to one side. From these first graphics we see that there is considerable discriminatory information in the first three SAVE variates that might complement or even dominate the information from LDA.

To carry the analysis a step further, we used the permutation test to estimate the number of SAVE variates that carry significant discriminatory information. That number turned out to be quite large, probably near 30. Moreover, the information contributed by the LDA variate is concentrated in SAVE variates 25–30, as judged by using OLS fits. SAVE did not miss the location separation found by LDA, but did judge it to be relatively unimportant. The reason for this can be found by inspecting marked three-dimensional plots for successive sets of SAVE variates. These plots all have the same general characteristics as the three-dimensional plot for the first three SAVE variates described previously: a spherical cloud of bad points with a concentration of good points at its center. Among the bad points, the standard deviation of the SAVE variates was observed to

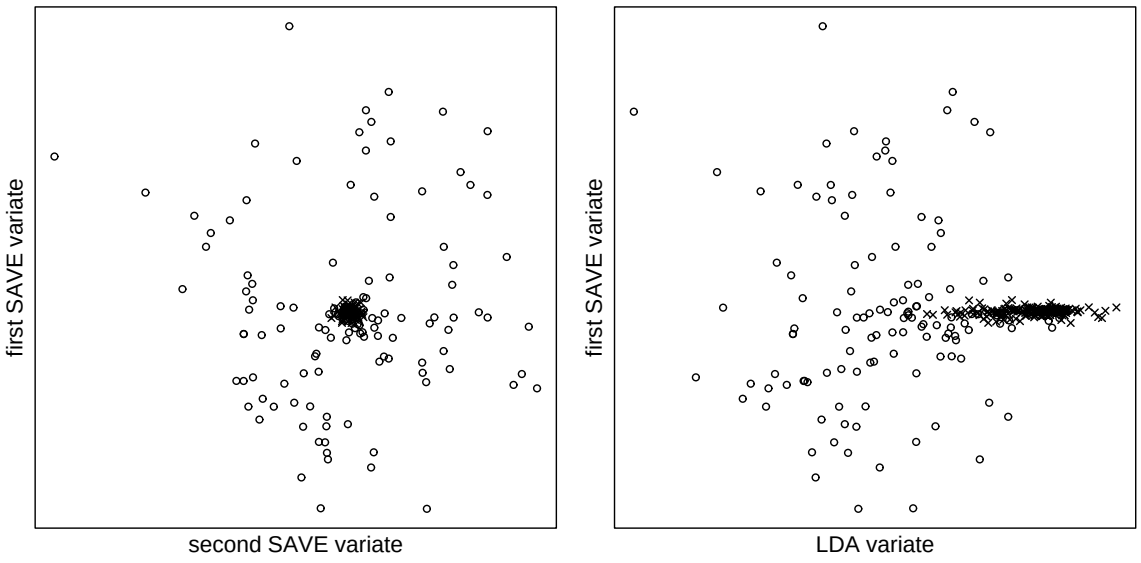

Figure 6: The left-hand graph shows a scatterplot of the first two SAVE variates for the ionosphere data. The 225 good cases are indicated with crosses which all fall near the center of the plot. The right-hand graph shows a scatterplot of the first SAVE variate and LDA variate.

be relatively constant, about 1.65. However, among the good points, the standard deviation of the SAVE variates increased monotonically, from .06 for the first SAVE variate to .55 for the 25th SAVE variate. The progression of three-dimensional plots reflected this, with the size of the central mass of good points increasing gradually relative to the cloud of bad points.

These characteristics of the SAVE variates can lead to reliable discrimination. As a qualitative illustration of this possibility, suppose that the first six SAVE variates are drawn from a spherical normal centered at 0 with standard deviation 1.65 for the bad points and .14 for the good points. The chance that a randomly selected bad point will fall in a hyper square centered at 0 with side $3 \times .14$ is quite small, about $.2^6$.

We conclude from our diagnostic analysis based on SAVE that the most useful discriminatory information likely comes from differences in scale and that differences in location are secondary. Because the various scatterplots we have observed appear approximately normal, and the good and bad point clouds seem approximately elliptical, classification by QDA seems a reasonable choice. If this option was used, then the discriminatory information would come primarily from differences in variation. The next example illustrates that SAVE can provide useful discriminatory information in nonnormal cases where LDA and QDA may not be robust.

5.2 Constructed robustness example

LDA can perform poorly in nonnormal cases, even when the variance matrices are the same (see Krzanowski, 1977). Similarly, Gnanadesikan (1989) mentioned that quadratic classification rules seem to perform poorly under nonnormality. On the other hand, our experience suggests that normality is not necessary for SAVE, since this method simply seeks to find reduced dimension variates without distributional assumptions. Guided by an initial graphical analysis of LDA and SAVE variates, any classification rule can then be applied. As shown in Section 4, using the Mahalanobis distance is then equivalent to QDA, but plotting the LDA and SAVE variates can often suggest a classification rule that may outperform QDA.

To illustrate, consider the following simulated example. We generated a sample size of 40 on ten predictors in each of three classes, $Y = 1, \ldots, 3$. Only the first two predictors (x_1, x_2) discriminate between classes. Letting x_{kci} denote the i-th observation on predictor k in class c, we generated the data as follows. For $i = 1, \ldots, 40$:

- $x_{1ci} \sim \mu_{1c} + \sigma_{1c} t_i$, where μ_{1c} is $\{.37, .36, -.57\}$ and σ_{1c} is $\{.33, .18, .13\}$ for $c = 1, \ldots, 3$, and t denotes a t random variable with 3 degrees of freedom;
- $x_{2ci} \sim \exp(\mu_{2c} + \sigma_{2c}\varepsilon + I_c x_{1ci})$, where μ_{2c} is $\{-.36, .48, .99\}$ and σ_{2c} is $\{.14, .27, .19\}$ for $c = 1, \ldots, 3$, ε is an independent standard normal random variable, and $I_c = 1$ for $c = \{1, 3\}$ and $I_c = -1$ otherwise;
- x_{3ci} and x_{4ci} are independent t random variables, each with 3 degrees of freedom;
- $(x_{5ci}, \ldots, x_{10ci})^T \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$.

The constants for the μ's and σ's were initially sampled from uniform distributions on $(-1, 1)$ and $(.1, .4)$ respectively. From this structure we see that two linear combinations of $\mathbf{X}$ are required for discrimination, and that the subspace, span($\mathbf{M}$), is spanned by $(1, 0, \ldots, 0)^T$ and $(0, 1, 0, \ldots, 0)^T$.

We generated 1000 such datasets, and ran three analyses on each: SAVE, the Young et al. (1987) method, and LDA. We calculated the vector correlation coefficient, q (Hotelling, 1936), between the two-dimensional subspace estimated by each method and the subspace span($\mathbf{M}$). SAVE resulted in a higher value of q than the Young et al. method in 84 percent of the simulations, while it was higher than q for LDA in 82 percent of the simulations. Counts of the times that q was between 0 and 0.2, between 0.2 and 0.4, between 0.4 and 0.6, between 0.6 and 0.8, or between 0.8 and 1 for SAVE versus the Young et al. method and for SAVE versus LDA are given in Table 1. Cells below the diagonal indicate SAVE outperformed the competing method, so that, for example, in 132 of the replications q was greater than 0.8 for SAVE but less than 0.2 for the Young et al. method.

Figure 7 shows one example from the 1000 simulations to illustrate. Despite the presence of outliers both in the discriminating variables (e.g., one in class 1 with low values of x_1 and x_2 and one in class 3 with a high value of x_2) as well as two of the noise variables (not shown in the plots), SAVE recovers the structure of the classes very well. By contrast, the Young et al. method is unable to recover this structure. Unsurprisingly, the first LDA variate recovers the location separation of the third class reasonably well, but the second LDA variate provides no information on the different intraclass covariances. In this dataset, the angle between span($\mathbf{M}$) and the SAVE subspace is 7 degrees, while the angles for the Young et al. method and LDA are 70 and 89 degrees, respectively.

SAVE can be robust in finding reduced dimension variates because it does not require normality to extract information from the sample intraclass covariance matrices, $\widehat{\mathbf{W}}_c$. Thus, as long as the

Table 1: Simulation comparison of SAVE, the method of Young, et al., and LDA

	q for Young et al. (LDA)				
q for SAVE	0–0.2	0.2–0.4	0.4–0.6	0.6–0.8	0.8–1
0–0.2	35 (28)	26 (12)	10 (16)	1 (12)	7 (11)
0.2–0.4	32 (17)	15 (10)	12 (13)	5 (11)	2 (15)
0.4–0.6	23 (12)	22 (16)	12 (10)	10 (10)	1 (20)
0.6–0.8	28 (20)	23 (15)	13 (12)	10 (15)	4 (16)
0.8–1	132 (213)	124 (164)	95 (135)	102 (113)	256 (84)

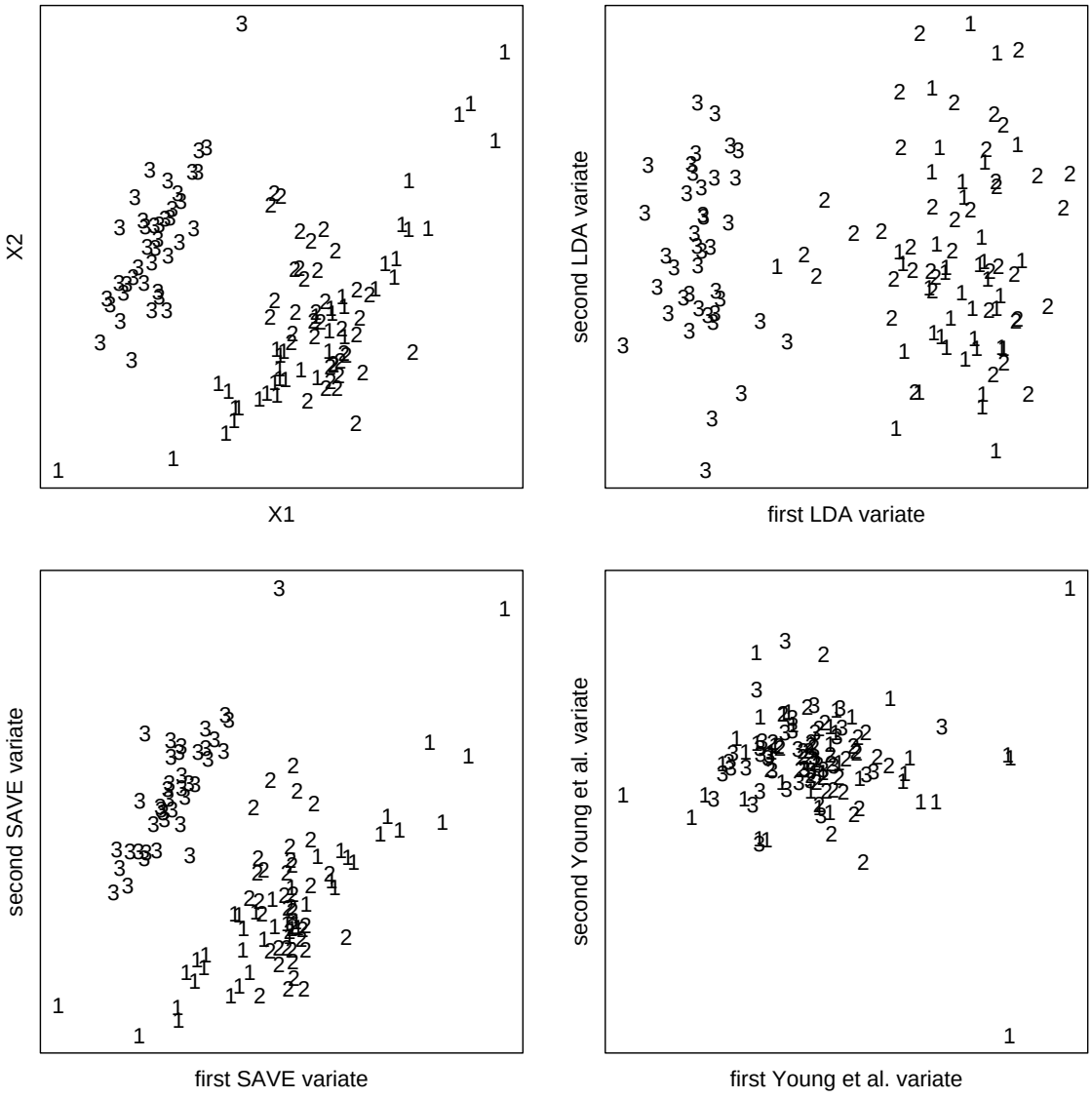

Figure 7: Scatterplots for the constructed robustness example. The upper left plot shows the first two feature variables representing the true dimension reduction subspace; the upper right plot shows the first two LDA canonical variates; the lower left plot shows the first two SAVE variates; the lower right plot shows the first two variates found using the method of Young et al.

$\widehat{\mathbf{W}}_c$'s are relatively distinct, SAVE should be able to find relevant reduced dimension variates. On the other hand, SAVE can lose power relative to LDA when the $\mathbf{W}_c$'s are similar. In such a case, most of the discrimination information is contained in the class means, and so LDA should be sufficient. Alternatively, another robust dimension reduction technique, sliced inverse regression (SIR) (Li, 1991; see also Chen and Li, 2001), can be useful in nonnormal cases when the $\mathbf{W}_c$'s are similar. However, while SIR offers an approach to discriminant analysis problems where graphical tools already exist (i.e., LDA canonical variates), our proposed approach using SAVE applies to problems where there are currently no general graphical methods available.

6 Discussion

LDA and QDA are widely regarded as first methods in discriminant and classification analysis. While there are graphical tools for LDA (McLachlan, 1992), it seems remarkable to us that there is a paucity of graphical tools in current statistical software for QDA. We have shown that using SAVE variates and then calculating Mahalanobis distance is equivalent to QDA on the original feature variables. Thus, SAVE provides a graphical foundation for QDA, and can facilitate selection of a classification rule. Just as there is a canonical variate approach to LDA, we believe that SAVE serves a similar purpose for QDA by providing useful graphical displays for discriminant analysis when intraclass covariances are not constant. In essence, SAVE provides a reduced dimension variate approach to QDA and can also provide additional insights when a normality assumption is questionable.

Judging from the survey of major programs presented in Section 3, the methods of Schott (1993), Young et al. (1987), and Zhu and Hastie (2003) seem to have had little general impact on standard graphical methodology in multi-class QDA problems. This article shows how SAVE provides an easily implemented graphical method that can supplement the standard graphics that come with LDA. To implement the proposed methodology in practice, a test of dimension suggests the number of SAVE variates to consider in scatterplots with points marked according to class. Two or three dimensions are easily handled in two or three dimensional scatterplots, while in cases where there are more than three dimensions, techniques such as linked plots, scatterplot matrices, or multipanel conditioning can be useful. In our experience, cases with more than three dimensions are infrequent, but do occur as the ionosphere data illustrate. LDA and SAVE variates can be plotted together to check for cases in which LDA finds mean differences more readily than SAVE, or they can be combined more formally using the method of Ye and Weiss (2003, Section 3.3).

One useful aspect of Figure 2, which shows the R QDA graphical display with two feature variables for the iris data, is the presence of curved boundaries marking the quadratic classification region. Similar boundaries could be a useful addition to SAVE plots also. However, just as the regions in Figure 2 are based on a specific classifier, in this case QDA, any region marked on a SAVE plot must be based on a classifier. By itself, SAVE is not a classifier, but is proposed here as an easy diagnostic method for visualization at the outset of a discriminant analysis. The resulting plots will often suggest a suitable classification rule that might lead to boundaries that could be added to a plot, but the SAVE methodology by itself does not provide these boundaries directly.

A related application of SAVE is to fitting mixtures of multivariate distributions. Yin and Cook (2003) use SAVE in an application with multivariate normal mixtures, while Cook and Critchley (2000) discuss more general theory on identifying mixtures graphically using inverse regression.

SAVE is a method originally intended to deal with dimension reduction in regression (Cook and Weisberg, 1991), where it is assumed that the distribution of $Y|\mathbf{X}$ depends on $\mathbf{X}$ only through the variates $\boldsymbol{\beta}^T\mathbf{X}$, where $\boldsymbol{\beta}$ is an unknown $p \times d$ matrix with $d < p$. More specifically, it is assumed that Y and $\mathbf{X}$ are independent given $\boldsymbol{\beta}^T\mathbf{X}$ and the analytic goal is to infer about d and the *central subspace*, span($\boldsymbol{\beta}$) (see Cook 1994a,b, 1996, 1998). Basic properties of the relationship between SAVE in regression and discriminant analysis were given by Cook and Yin (2001).

Acknowledgments

The authors are grateful to the Associate Editor and the Referees for helpful comments on earlier versions of this article. This work was supported in part by National Science Foundation Grant DMS-0405360 awarded to R. Dennis Cook.

A Appendix

A.1 Justification of Lemma 1

First, rewrite the following:

$$\begin{aligned} D_c^2 &= (\boldsymbol{\mu}_{\mathbf{U}_c} - \mathbf{U}_0)^T(\boldsymbol{\Gamma}_1^T\mathbf{W}_c\boldsymbol{\Gamma}_1)^{-1}(\boldsymbol{\mu}_{\mathbf{U}_c} - \mathbf{U}_0) \\ &= (\boldsymbol{\mu}_c - \mathbf{Z}_0)^T\boldsymbol{\Gamma}_1(\boldsymbol{\Gamma}_1^T\mathbf{W}_c\boldsymbol{\Gamma}_1)^{-1}\boldsymbol{\Gamma}_1^T(\boldsymbol{\mu}_c - \mathbf{Z}_0) \\ &= \boldsymbol{\mu}_c^T\boldsymbol{\Gamma}_1(\boldsymbol{\Gamma}_1^T\mathbf{W}_c\boldsymbol{\Gamma}_1)^{-1}\boldsymbol{\Gamma}_1^T\boldsymbol{\mu}_c - 2\boldsymbol{\mu}_c^T\boldsymbol{\Gamma}_1(\boldsymbol{\Gamma}_1^T\mathbf{W}_c\boldsymbol{\Gamma}_1)^{-1}\boldsymbol{\Gamma}_1^T\mathbf{Z}_0 \\ &\quad + \mathbf{Z}_0^T\boldsymbol{\Gamma}_1(\boldsymbol{\Gamma}_1^T\mathbf{W}_c\boldsymbol{\Gamma}_1)^{-1}\boldsymbol{\Gamma}_1^T\mathbf{Z}_0 \\ &= F + G + H. \end{aligned}$$

By Proposition 1, we know that $\text{span}(I - \mathbf{W}_c) = \text{span}(\text{E}(I - \mathbf{W}_c)^2) = \text{span}(\mathbf{M})$, where $\mathbf{M}$ is the SAVE kernel matrix.
Next, write the spectral decomposition of $\mathbf{M}$ as $\mathbf{M} = \boldsymbol{\Gamma}\mathbf{D}\boldsymbol{\Gamma}^T$ and let $\boldsymbol{\Gamma}_1 = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_{p_1})$ so that

$$\begin{aligned} I - \mathbf{W}_c &= (\boldsymbol{\Gamma}_1(c), \boldsymbol{\Gamma}_2(c)) \begin{pmatrix} \mathbf{D}_1(c) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Gamma}_1(c)^T \\ \boldsymbol{\Gamma}_2(c)^T \end{pmatrix}, \\ \text{where } \boldsymbol{\Gamma}_1(c) &= (\boldsymbol{\gamma}_1(c), ..., \boldsymbol{\gamma}_{p_1}(c)) \\ \text{and } \boldsymbol{\gamma}_j(c) &= \sum_{k=1}^{p_1} a_{jk}(c)\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_{p_1}) \begin{pmatrix} a_{j1}(c) \\ \cdot \\ a_{jp_1}(c) \end{pmatrix} = \boldsymbol{\Gamma}_1\mathbf{A}_j(c), \end{aligned}$$

and $\mathbf{A}_j(c) = (a_{j1}(c), ..., a_{jp_1}(c))^T$ is a $p_1 \times 1$ vector.
Note that if $j \neq k$, then $\boldsymbol{\gamma}_j(c)^T\boldsymbol{\gamma}_k(c) = 0$, so $\mathbf{A}_j(c)^T\mathbf{A}_k(c) = 0$. Also, $\boldsymbol{\gamma}_j(c)^T\boldsymbol{\gamma}_j(c) = 1$, so $\mathbf{A}_j(c)^T\mathbf{A}_j(c) = 1$.
Next, define a $p_1 \times p_1$ matrix

$$\mathbf{A}(c) = (\mathbf{A}_1(c), ..., \mathbf{A}_{p_1}(c)).$$

Because $\mathbf{W}_c$ is nonsingular, none of the elements of $\mathbf{D}_1(c)$ is 1. We then have

$$\begin{aligned} \boldsymbol{\Gamma}_1(c) &= \boldsymbol{\Gamma}_1\mathbf{A}(c) \\ \boldsymbol{\Gamma}_1^T\boldsymbol{\Gamma}_1(c) &= \mathbf{A}(c) \\ \boldsymbol{\Gamma}_1^T\mathbf{W}_c\boldsymbol{\Gamma}_1 &= \boldsymbol{\Gamma}_1^T(I - \boldsymbol{\Gamma}_1(c)\mathbf{D}_1(c)\boldsymbol{\Gamma}_1(c)^T)\boldsymbol{\Gamma}_1 \\ &= I - \mathbf{A}(c)\mathbf{D}_1(c)\mathbf{A}(c)^T. \end{aligned}$$

Since $\mathbf{W}_c = I - \boldsymbol{\Gamma}_1(c)\mathbf{D}_1(c)\boldsymbol{\Gamma}_1(c)^T$, then

$$\begin{aligned} \mathbf{W}_c^{-1} &= \boldsymbol{\Gamma}(c) \begin{pmatrix} (I - \mathbf{D}_1(c))^{-1} & 0 \\ 0 & I \end{pmatrix} \boldsymbol{\Gamma}(c)^T \\ &= I + \boldsymbol{\Gamma}_1(c)(I - \mathbf{D}_1(c))^{-1}\boldsymbol{\Gamma}_1(c)^T - \boldsymbol{\Gamma}_1(c)\boldsymbol{\Gamma}_1(c)^T \\ &= I + \boldsymbol{\Gamma}_1(c)(I - (I - \mathbf{D}_1(c)))(I - \mathbf{D}_1(c))^{-1}\boldsymbol{\Gamma}_1(c)^T \\ &= I + \boldsymbol{\Gamma}_1(c)\mathbf{D}_1(c)(I - \mathbf{D}_1(c))^{-1}\boldsymbol{\Gamma}_1(c)^T, \\ \text{and } \boldsymbol{\Gamma}_1^T\mathbf{W}_c^{-1}\boldsymbol{\Gamma}_1 &= I + \mathbf{A}(c)\mathbf{D}_1(c)(I - \mathbf{D}_1(c))^{-1}\mathbf{A}(c)^T. \end{aligned}$$

Now it is straightforward to see that

$$\begin{aligned} (\boldsymbol{\Gamma}_1^T\mathbf{W}_c\boldsymbol{\Gamma}_1)(\boldsymbol{\Gamma}_1^T\mathbf{W}_c^{-1}\boldsymbol{\Gamma}_1) &= (I - \mathbf{A}(c)\mathbf{D}_1(c)\mathbf{A}(c)^T)(I + \mathbf{A}(c)\mathbf{D}_1(c)(I - \mathbf{D}_1(c))^{-1}\mathbf{A}(c)^T) \\ &= I + \mathbf{A}(c)[\mathbf{D}_1(c)(I - \mathbf{D}_1(c))^{-1} - \mathbf{D}_1(c) - \mathbf{D}_1(c)^2(I - \mathbf{D}_1(c))^{-1}]\mathbf{A}(c)^T \\ &= I. \end{aligned}$$

In other words, $(\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1)^{-1} = \mathbf{\Gamma}_1^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_1$.
Since by Proposition 1 the class means, $\boldsymbol{\mu}_c$, are in span($\mathbf{M}$), i.e., $(\mathbf{\Gamma}_1\mathbf{\Gamma}_1^T)\boldsymbol{\mu}_c = \boldsymbol{\mu}_c$, then

$$\begin{aligned} F &= \boldsymbol{\mu}_c^T\mathbf{\Gamma}_1(\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1)^{-1}\mathbf{\Gamma}_1^T\boldsymbol{\mu}_c \\ &= \boldsymbol{\mu}_c^T\mathbf{\Gamma}_1(\mathbf{\Gamma}_1^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_1)\mathbf{\Gamma}_1^T\boldsymbol{\mu}_c \\ &= \boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}\boldsymbol{\mu}_c. \end{aligned}$$

Next, since $\mathbf{\Gamma} = (\mathbf{\Gamma}_1, \mathbf{\Gamma}_2)$ is nonsingular, there exists a $\mathbf{V}_0$ such that $\mathbf{Z}_0 = (\mathbf{\Gamma}_1, \mathbf{\Gamma}_2)\mathbf{V}_0$. By the fact that $\boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_2 = \boldsymbol{\mu}_c^T\mathbf{\Gamma}_2 = \boldsymbol{\mu}_c^T\mathbf{\Gamma}_1\mathbf{\Gamma}_1^T\mathbf{\Gamma}_2 = 0$, then

$$\begin{aligned} G &= -2\boldsymbol{\mu}_c^T\mathbf{\Gamma}_1(\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1)^{-1}\mathbf{\Gamma}_1^T\mathbf{Z}_0 \\ &= -2\boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_1\mathbf{\Gamma}_1^T\mathbf{Z}_0 \\ &= -2\boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_1\mathbf{\Gamma}_1^T(\mathbf{\Gamma}_1, \mathbf{\Gamma}_2)\mathbf{V}_0 \\ &= -2\boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}(\mathbf{\Gamma}_1, 0)\mathbf{V}_0 \\ &= -2\boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}(\mathbf{\Gamma}_1, \mathbf{\Gamma}_2)\mathbf{V}_0 \\ &= -2\boldsymbol{\mu}_c^T\mathbf{W}_c^{-1}\mathbf{Z}_0. \end{aligned}$$

Finally, using $\mathbf{\Gamma}_2^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_1 = 0$ and $\mathbf{\Gamma}_2^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_2 = I$, we have

$$\begin{aligned} H &= \mathbf{Z}_0^T\mathbf{\Gamma}_1(\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1)^{-1}\mathbf{\Gamma}_1^T\mathbf{Z}_0 \\ &= \mathbf{V}_0^T\begin{pmatrix} \mathbf{\Gamma}_1^T \\ \mathbf{\Gamma}_2^T \end{pmatrix}\mathbf{\Gamma}_1\mathbf{\Gamma}_1^T\mathbf{W}_c^{-1}\mathbf{\Gamma}_1\mathbf{\Gamma}_1^T(\mathbf{\Gamma}_1, \mathbf{\Gamma}_2)\mathbf{V}_0 \\ &= \mathbf{V}_0^T\begin{pmatrix} \mathbf{\Gamma}_1^T \\ 0 \end{pmatrix}\mathbf{W}_c^{-1}(\mathbf{\Gamma}_1, 0)\mathbf{V}_0 \\ &= \mathbf{V}_0^T\begin{pmatrix} \mathbf{\Gamma}_1^T - 0 \\ \mathbf{\Gamma}_2^T - \mathbf{\Gamma}_2^T \end{pmatrix}\mathbf{W}_c^{-1}(\mathbf{\Gamma}_1 - 0, \mathbf{\Gamma}_2 - \mathbf{\Gamma}_2)\mathbf{V}_0 \\ &= \mathbf{Z}_0^T\mathbf{W}_c^{-1}\mathbf{Z}_0 - \mathbf{V}_0^T\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}\mathbf{V}_0 \end{aligned}$$

The last term in the above expression is constant for all c, and hence

$$D_c^2 = -2\mathcal{Q}_c + K.$$

□

A.2 Justification of Proposition 2

The result holds if $\log|\mathbf{W}_c| = \log|\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1|$.
Let λ_i and $\mathbf{b}_i$ $(i = 1, \ldots, p)$ be the eigenvalues and eigenvectors of $\mathbf{W}_c$, respectively, so that $|\mathbf{W}_c| = \prod_{i=1}^p \lambda_i$ and $\mathbf{W}_c\mathbf{b}_i = \lambda_i\mathbf{b}_i$.
Since from Section A.1, $\mathbf{W}_c = I - \mathbf{\Gamma}_1(c)\mathbf{D}_1(c)\mathbf{\Gamma}_1(c)^T$, then $\lambda_i\mathbf{b}_i = (I - \mathbf{\Gamma}_1(c)\mathbf{D}_1(c)\mathbf{\Gamma}_1(c)^T)\mathbf{b}_i$.
Next, since $\mathbf{\Gamma}_1^T\mathbf{\Gamma}_1(c)\mathbf{D}_1(c)\mathbf{\Gamma}_1(c)^T = \mathbf{A}(c)\mathbf{D}_1(c)\mathbf{A}(c)^T\mathbf{\Gamma}_1^T$, then

$$\begin{aligned} \lambda_i\mathbf{\Gamma}_1^T\mathbf{b}_i &= \mathbf{\Gamma}_1^T\mathbf{b}_i - \mathbf{\Gamma}_1^T\mathbf{\Gamma}_1(c)\mathbf{D}_1(c)\mathbf{\Gamma}_1(c)^T\mathbf{b}_i \\ &= (I - \mathbf{A}(c)\mathbf{D}_1(c)\mathbf{A}(c)^T)\mathbf{\Gamma}_1^T\mathbf{b}_i \\ &= (\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1)\mathbf{\Gamma}_1^T\mathbf{b}_i. \end{aligned}$$

Thus $|\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1| = \prod_{i=1, \mathbf{\Gamma}_1^T\mathbf{b}_i \neq 0}^p \lambda_i$.

Next, since $I - \mathbf{W}_c = \mathbf{\Gamma}_1 \mathbf{A}(c)\mathbf{D}_1(c)\mathbf{A}(c)^T\mathbf{\Gamma}_1^T$, then for $\mathbf{\Gamma}_1^T\mathbf{b}_i = 0$ we have $\mathbf{b}_i = \mathbf{W}_c\mathbf{b}_i = \lambda_i\mathbf{b}_i$ and so $\lambda_i = 1$.
Therefore, $|\mathbf{W}_c| = \prod_{i=1}^p \lambda_i = \prod_{i=1,\mathbf{\Gamma}_1^T\mathbf{b}_i\neq 0}^p \lambda_i = |\mathbf{\Gamma}_1^T\mathbf{W}_c\mathbf{\Gamma}_1|$.
□

A.3 Justification of Proposition 3

Let $P_{\mathbf{M}}$ denote the projection onto span($\mathbf{M}$) and let $Q_{\mathbf{M}} = I - P_{\mathbf{M}}$. Then $\mathbf{W}_c Q_{\mathbf{M}} = Q_{\mathbf{M}}$ for $c = 1, \ldots C$, since a vector $\mathbf{v} \in \text{span}^{\perp}(\mathbf{M})$ if and only if $(I - \mathbf{W}_c)\mathbf{v} = 0$ for all c. To facilitate presentation, consider minimizing $-\text{L}^*(\boldsymbol{\alpha}) = E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha}))$. Then using these results we have

$$\begin{aligned} E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha})) &= E(\log(\boldsymbol{\alpha}^T(P_{\mathbf{M}} + Q_{\mathbf{M}})\mathbf{W}_c(P_{\mathbf{M}} + Q_{\mathbf{M}})\boldsymbol{\alpha})) \\ &= E(\log(\boldsymbol{\alpha}^T P_{\mathbf{M}}\mathbf{W}_c P_{\mathbf{M}}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T Q_{\mathbf{M}}\mathbf{W}_c Q_{\mathbf{M}}\boldsymbol{\alpha})) \\ &= E(\log(\boldsymbol{\alpha}^T P_{\mathbf{M}}\mathbf{W}_c P_{\mathbf{M}}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T Q_{\mathbf{M}}\boldsymbol{\alpha})). \end{aligned}$$

If $\boldsymbol{\alpha} \in \text{span}^{\perp}(\mathbf{M})$ then $E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha})) = 0 = \max_\alpha E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha}))$. In other words, an $\boldsymbol{\alpha} \in \text{span}^{\perp}(\mathbf{M})$ maximizes rather than minimizes $E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha}))$.

To see this recall that $E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha})) \leq \log(\boldsymbol{\alpha}^T E(\mathbf{W}_c)\boldsymbol{\alpha})$. Because $E(\mathbf{W}_c)$ is of the form $E(\mathbf{W}_c) = I - \mathbf{V} > 0$ with $\mathbf{V} \geq 0$, it follows that the eigenvalues of $E(\mathbf{W}_c)$ are between 0 and 1, and thus $\max_\alpha \boldsymbol{\alpha}^T E(\mathbf{W}_c)\boldsymbol{\alpha} = 1$.

This says that to minimize $E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha}))$, $\boldsymbol{\alpha}$ must have a nontrivial projection onto span($\mathbf{M}$). Consequently we can write

$$E(\log(\boldsymbol{\alpha}^T\mathbf{W}_c\boldsymbol{\alpha})) \geq E(\log(\boldsymbol{\alpha}^T P_{\mathbf{M}}\mathbf{W}_c P_{\mathbf{M}}\boldsymbol{\alpha})),$$

with strict inequality when $\boldsymbol{\alpha}^T Q_{\mathbf{M}}\boldsymbol{\alpha} > 0$. The objective function achieves its lower bound when $\boldsymbol{\alpha} \in \text{span}(\mathbf{M})$, and consequently the first vector $\boldsymbol{\alpha}_1^*$ must be in the SAVE subspace. The logic for subsequent vectors is the same under the constraint that they are orthogonal.
□

References

Bura, E., and Pfeiffer, R.M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, 19, 1252–1258.

Chen, C.-H., and Li, K.-C. (2001). Generalization of Fisher's linear discriminant analysis via the approach of sliced inverse regression. *Journal of the Korean Statistical Society*, 30, 193–218

Cook, R. D. (1994a). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89, 177–190.

Cook, R. D. (1994b). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section on Physical Engineering Sciences*, Washington.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association* 91, 983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas For Studying Regressions Through Graphics*. New York: Wiley.

Cook, R. D. and Critchley, F. (2000). Identifying outliers and regression mixtures graphically. *Journal of the American Statistical Association*, 95, 781–794

Cook, R. D. and Lee, H. (1999). Dimension reduction in regressions with a binary response. *Journal of the American Statistical Association,* 94, 1187–1200.

Cook, R. D. and Ni, L. (2005). Dimension reduction with inverse regression: a minimum discrepancy approach. *Journal of the American Statistical Association,* 100, 410–428.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, 86, 328–332.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics.* New York: Wiley.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian and New Zealand Journal of Statistics*, 43, 147–199.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition.* 2nd ed. Academic Press, Boston.

Gnanadesikan, R. (1989). Discriminant analysis and clustering. *Statistical Science* 4, 34–69.

Hand, D. J. (2006). Classifier technology and the illusion of progress (with discussion). *Statistical Science*, 21, 1–34.

Hennig, C. (2004). Asymmetric linear dimension reduction for classification. *Journal of Computational and Graphical Statistics*, 13, 930–945.

Hernández, A. and Velilla, S. (2005). Dimension reduction in nonparametric kernel discriminant analysis. *Journal of Computational and Graphical Statistics,* 14, 847–866.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika,* 28, 321–377.

Khattree, R. and Naik, D. (2000). Multivariate Data Reduction and Discrimination with SAS Software. SAS Press and John Wiley Sons Inc.

Krzanowski, W. J. (1977). The performance of Fisher's linear discriminant Function under non-optimal conditions. *Technometrics*, Vol, 19, No. 2, 191–200.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316–342.

Li, L. (2006). A note on sparse sufficient dimension reduction. North Carolina State University Technical Report.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, 48, 503–510.

Ling, Y. Bhandarkar, S. Yin, X. and Lu, Q. (2005). Saveface and Sirface: Appearance-based recognition of faces and facial expressions. *Proceeding of the IEEE international conference on Image Processing ICIP, Genova, Italy*, Sept. 11–14.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7, 815–840.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

Peltonen, J. and Kaski, S. (2005). Discriminative components of data. *IEEE Transactions on Neural Networks,* 16, 68–83.

Sall, J., Creighton, L., and Lehman, A. (2004). *JMP Start Statistics*, third edition. SAS Press.

Shao, Y., Cook, R.D. and Weisberg, S. (2006). Marginal tests with sliced average variance estimation. *Biometrika,* to appear.

Schott, J. R. (1993). Dimension reduction in quadratic discriminant analysis. *Computational Statistics and Data Analysis* 16, 161–174.

Sheather, S. J. and McKean, J. W. (2001). Discussion of Cook and Yin (2001). *Australian and New Zealand Journal of Statistics*, 43, 185–190.

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 10, 262–266.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, fourth edition. New York: Springer-Verlag.

Weisberg, S. (2002). Dimension reduction regression in R. *Journal of Statistical Software,* 7(1).

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98, 968–979.

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika,* 90, 113–125.

Young, D. M., Marco, V. R. and Odell, P. L. (1987). Quadratic discrimination: some results on optimal low-dimensional representation. *Journal of Statistical Planning and Inference*, 17, 307–319.

Zhu, M. and Hastie, T. J. (2003). Feature extraction for nonparametric discriminant analysis. *Journal of Computational and Graphical Statistics,* 12, 101–120.