Journal of Data Science **3**(2005), 449-464

A Monte Carlo Comparison of Two Linear Dimension Reduction Matrices for Statistical Discrimination

J. Wade Davis[1], Dean M. Young[1] and Karin B. Ernstrom-Keim[2]
[1]*Baylor University and* [2] *University of California, San Diego*

Abstract: We compare two linear dimension-reduction methods for statistical discrimination in terms of average probabilities of misclassification in reduced dimensions. Using Monte Carlo simulation we compare the dimension-reduction methods over several different parameter configurations of multivariate normal populations and find that the two methods yield very different results. We also apply the two dimension-reduction methods examined here to data from a study on football helmet design and neck injuries.

Key words: Dimension reduction, discriminant analysis, singular value decomposition.

1. Introduction

Statistical classification involves assigning a given observation $\mathbf{x}$ to one of k possible classes (or populations) based on p measured variables, also known as *features*. As the dimension of the feature space p increases, the computational complexity for the classification task can become cumbersome and time consuming. In addition, more training samples are needed to design appropriate classification rules. Therefore, one often desires to reduce the dimension of the original feature vector if possible.

In this paper we consider the topic of linear feature reduction for statistical classification. Specifically, we compare and contrast the efficacies of two linear feature-reduction methods formulated by Brunzell and Eriksson (2000) and Tubbs, Coberly, and Young (1982) using a Monte Carlo simulation. The two linear dimension-reduction methods considered resemble each other but do not, in general, give equivalent results in terms of expected probabilities of misclassification. In this paper we clarify some differences and similarities between the two methods by addressing the following questions. When and why do the methods give different or similar results? For which data characteristics is one method better than the other? Can either method improve the probability of classification compared to using the full dimension of the feature vector?

To address these questions, we perform a Monte Carlo simulation study to compare classification performance in the full feature space versus classification in a reduced space determined via the methods developed by Tubbs, Coberly, and Young (TCY) and Brunzell and Eriksson (BE). We note that BE have contrasted their linear dimension-reduction approach to that of TCY, and to other methods such as the Mahalanobis-based linear transformation, canonical variables, principal components analysis, and four variations of Fisher's discriminant. For more comparisons of pattern recognition methods in high-dimensional settings, see Aeberhard, de Vel, and Coomans (1994).

On the data sets considered in BE, BE's dimension reduction method is uniformly superior to that of TCY in terms of yielding smaller expected error rates in a reduced dimension. Our goal is to analyze the performance of these two linear feature-selection matrices over classification problems with diverse parameter configurations.

2. Linear Dimension-Reduction Matrices

In pattern recognition and statistical discriminant classification problems, one often desires to reduce the dimension of the feature space before classification. A reduced dimension can result in fewer computations, a reduction in cost and time, and even improved classification accuracy. Additionally, one typically needs fewer training observations to estimate population parameters because the necessary training sample size is directly related to the feature dimension. If the number of training observations can be reduced without a significant increase in the probability of misclassification (PMC), the classification task becomes more efficient in terms of time and cost.

Many different competing feature-reduction methods exists. The two methods we discuss are linear transformations of the feature vector $\mathbf{x} \in R_{p\times 1}$, which are of the form

$$\mathbf{x} \to \mathbf{y} = \mathbf{T}^t\mathbf{x} \tag{2.1}$$

with $\mathbf{T} \in R_{p\times q}$, where p is the original full dimension and q is the transformed reduced dimension. The matrix $\mathbf{T}$ is known as a linear feature-selection or linear dimension-reduction matrix. We desire that $1 \leq q \ll p$ and that the PMC remains essentially the same as in the full-dimension case.

Dimension-reduction methods are beneficial in the case when the ratio of the training sample size n to the dimension of the feature vector p is small ($n/p < 4$). If $1 \leq n/p < 4$, then one can encounter a problem with accurately inverting the covariance matrices due to extreme bias from small eigenvalues of the covariance matrices. Reducing the feature dimension gives a more stable estimated covariance matrix and estimated inverse covariance matrix by decreasing the number

of parameters to be estimated.

In the next two subsections, we review two linear feature-selection methods for the case of unequal covariance matrices.

2.1 Tubbs, Coberly, and Young's linear feature-selection method (TCY)

The objective of TCY is to determine a matrix to perform a linear transformation such that the PMC in the reduced q-dimensional transformed feature space is approximately the same as in the original p-dimensional feature space, or $PMC(p) \approx PMC(q)$. The following theorem describes the motivation for TCY.

Theorem 1. (Tubbs, Coberly and Young, 1982): Let Π_i be a p-dimensional multivariate normal population with a priori probability α_i, mean $\boldsymbol{\mu}_i \in R_{p\times 1}$, and symmetric nonnegative-definite covariance $\boldsymbol{\Sigma}_i, i = 1, 2, \ldots, k$, such that $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$ for at least one value of $j,\ 2 \leq j \leq k$. Let

$$\mathbf{M} = [\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 | \boldsymbol{\mu}_3 - \boldsymbol{\mu}_1 | \cdots | \boldsymbol{\mu}_k - \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 | \boldsymbol{\Sigma}_3 - \boldsymbol{\Sigma}_2 | \cdots | \boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_1 |], \quad (2.2)$$

$1 \leq i < j \leq k$, and let FG be a full-rank decomposition of $\mathbf{M}$ such that $\mathbf{M} = \mathbf{FG}$ with $\text{rank}(\mathbf{M}) = \text{rank}(\mathbf{F}) = \text{rank}(\mathbf{G}) = q,\ 1 \leq q < p$. Then the p-variate Bayes procedure assuming equal cost loss assigns $\mathbf{x}$ to Π_i if and only if the q-variate Bayes procedure assuming equal cost loss assigns $\mathbf{F}^+\mathbf{x}$ to $\Pi_i,\ i = 1, 2, \ldots, k$, where $\mathbf{F}^+$ denotes the Moore-Penrose generalized inverse of $\mathbf{F}$ (Harville, 1997, pg. 493). Moreover, q is the smallest positive integer such that there exists a $q \times p$ compression matrix preserving the Bayes assignment of $\mathbf{x}$ to Π_i.

Theorem 1 yields a linear transformation $\mathbf{F}^+ \in R_{q\times p}$ such that $PMC(p) = PMC(q)$ provided $\text{rank}(M) = q < p$. If $\text{rank}(\mathbf{M}) = p$, there exists no $q \times p$ matrix that preserves the full-feature PMC and, thus, we seek a matrix $\mathbf{T} \in R_{p\times q}$ such that $PMC(p) \approx PMC(q)$.

The parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i,\ i = 1, 2, \ldots, k$, are rarely known and, therefore, sample estimates must be obtained using the n_i training samples. An estimator of $\mathbf{M}$ is then

$$\hat{\mathbf{M}} = [\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1 | \bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_1 | \cdots | \bar{\mathbf{x}}_k - \bar{\mathbf{x}}_1 | \mathbf{S}_2 - \mathbf{S}_1 | \mathbf{S}_3 - \mathbf{S}_1 | \cdots | \mathbf{S}_k - \mathbf{S}_1 |].$$

Let n_i be the sample size for estimating the parameters of the i-th population. If $n_i \geq p$, then $\text{rank}(\hat{\mathbf{M}}) = p$ with probability one. In this case, Theorem 1 cannot be directly applied, so Tubbs, Coberly and Young (1982) use the singular value decomposition (SVD) to obtain a best approximation of $\hat{\mathbf{M}}$ (under the Frobenius norm) in a smaller dimension $q < p$.

Let $\hat{\mathbf{M}} = \mathbf{PD}_p\mathbf{Q}^t$ be the SVD of $\hat{\mathbf{M}}$, where $\mathbf{D}_p \equiv Diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ with $\lambda_i \geq \lambda_j$ for $1 \leq i \leq j \leq p$ and let $\hat{\mathbf{F}} = \mathbf{PD}_p$. Define $\mathbf{D}_p = Diag(\lambda_1, \lambda_2, \ldots, \lambda_q, 0_{q+1},$

$\ldots, 0_p)$ with $\lambda_i \geq \lambda_j$ for $1 \leq i \leq j \leq q$. Then $\tilde{\mathbf{M}}_q = \mathbf{P}\mathbf{D}_q\mathbf{Q}^t$ is a rank q approximation of $\hat{\mathbf{M}}$ and $\tilde{\mathbf{F}}_q = \mathbf{P}\mathbf{D}_q$ is a rank q approximation of $\mathbf{P}\mathbf{D}_p$. A $q \times p$ feature-reduction matrix to perform the linear transformation in equation (2.1) is then $\hat{\mathbf{F}}_q^+ = \hat{\mathbf{P}}_q^+$ where $\tilde{\mathbf{F}}_q = [\mathbf{P}_q : \mathbf{0}] \in R_{p\times p}$.

Because TCY allows for unequal means and unequal covariance matrices, $\hat{\mathbf{F}}_q^+$ should perform well when the population covariances are unequal and the number of large singular values of $\hat{\mathbf{M}}$ is small relative to p. Also, the method should perform well when n_i is large and rank($\mathbf{M}$)$= q \ll p$ because the estimators $\bar{\mathbf{x}}_i$ and $\mathbf{S}_i,\ i = 1, 2, \ldots, k$, are strongly consistent.

2.2 Brunzell and Eriksson's linear feature-selection method (BE)

Tubbs, Coberly, and Young (1982) explicitly use PMC as their dimension-reduction optimality criterion, whereas Brunzell and Eriksson implicitly consider the PMC via a derived distance measure. This distance measure

$$\Delta_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t(\alpha_i\boldsymbol{\Sigma}_i + \alpha_j\boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),$$

$1 \leq i < j \leq k$, is used to obtain an upper bound on the expected PMC denoted by $EPMC$. For two multivariate populations with prior probabilities α_1 and α_2 with $\alpha_1 + \alpha_2 = 1$, we obtain $EPMC \leq (2\alpha_1\alpha_2)/(1 + \alpha_1\alpha_2\Delta_{12})$.

In the case of two populations with equal covariance matrices and equal prior probabilities, Δ_{12} is the squared generalized Mahalanobis distance between class means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Brunzell and Eriksson (2000) introduce a generalized separation measure for the case of k populations with possibly unequal covariance matrices. Assuming the prior probabilities are equal, they obtain the separation measure

$$\prod_{1\leq i\leq j\leq k} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t(\alpha_i\boldsymbol{\Sigma}_i + \alpha_j\boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{2.3}$$

The objective of BE is to determine a linear dimension-reduction matrix with $q \ll p$ such that the full-dimension separation measure is at least approximately preserved. The following theorem provides motivation for the BE method.

Theorem 2. (Brunzell and Eriksson, 2000). Let

$$\begin{aligned}\mathbf{U} \;=\; & [(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)|\cdots|(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)|\cdots \\ & |(\boldsymbol{\Sigma}_{k-1} + \boldsymbol{\Sigma}_k)^{-1}(\boldsymbol{\mu}_{k-1} - \boldsymbol{\mu}_k)]\end{aligned} \tag{2.4}$$

for $1 \leq i < j \leq k$. The separation measure (2.3) is preserved by the transformation $\mathbf{x} \to \mathbf{y} = \mathbf{T}^t\mathbf{x}$ if $\mathbf{T}$ satisfies the condition $R(\mathbf{T}) \supseteq R(\mathbf{U})$, where $R(\cdot)$ represents a row space.

Consider now the case where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, $i = 1, 2, \ldots, k$, are unknown and must be estimated from the n_i training samples. An estimator of $\mathbf{U}$ is then

$$\hat{\mathbf{U}} = [(\mathbf{S}_1 + \mathbf{S}_2)^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \cdots | (\mathbf{S}_i + \mathbf{S}_j)^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) \cdots | (\mathbf{S}_{k-1} + \mathbf{S}_k)^{-1}(\bar{\mathbf{x}}_{k-1} - \bar{\mathbf{x}}_k)|],$$

$1 \leq i < j \leq k$. Note that $\hat{\mathbf{U}} \in R_{p \times r}$, where $r = k(k-1)/2$, and if $r > q$, then $\hat{\mathbf{U}}$ does not yield a $q \times p$ linear dimension-reduction matrix. Therefore, BE, like TCY, utilizes the SVD rank-q approximation of $\hat{\mathbf{U}}$ to obtain a linear feature-reduction matrix that compresses p-dimensional observation vectors into a q-dimensional transformed feature space where $1 \leq q < p$.

Let $\tilde{\mathbf{U}} = \mathbf{R}\mathbf{D}_p\mathbf{S}^t$ be the SVD of the matrix $\hat{\mathbf{U}}$, where $\mathbf{D}_p = Diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ with $\lambda_i \geq \lambda_j$ for $1 \leq i \leq j \leq p$ and let $\hat{\mathbf{H}} = \mathbf{R}\mathbf{D}_p$. Further, let $\tilde{\mathbf{H}} = \mathbf{R}\mathbf{D}_q$, where $\mathbf{D}_q = Diag(\lambda_1, \lambda_2, \ldots, \lambda_q, 0_{q+1}, \ldots, 0_p)$ with $\lambda_i \geq \lambda_j$ for $1 \leq i \leq j \leq q$. Then, $\tilde{\mathbf{U}}$ is a rank q approximation of $\hat{\mathbf{U}}$ and $\tilde{\mathbf{H}}$ is a rank q approximation of $\hat{\mathbf{H}}$. A $q \times p$ feature-reduction matrix to perform the linear transformation in equation (2.1) is then $\hat{\mathbf{H}}_q^t = \mathbf{R}_q$, where $\hat{\mathbf{H}}_q = \mathbf{R}_q$ and $\hat{\mathbf{H}}_1^t = [\mathbf{R}_q : \mathbf{0}] \in R_{p \times p}$.

The BE technique classifies the data based essentially on the rotated difference in the means rather than on the differences in the covariance structures. Note that separation measure (2.3) uses a type of pooled covariance matrix. By pooling the pairs of covariance matrices, Eriksson and Brunzell are not necessarily using all of the information in the differences of the covariance matrices. However, pooling is beneficial in dealing with the near singularity of $\mathbf{S}_i$, which occurs when n_i is small relative to p so that pooling $\mathbf{S}_i$ and $\mathbf{S}_j$ gives more stable values of $(\mathbf{S}_i + \mathbf{S}_j)^{-1}$ to estimate $(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)^{-1}$, $1 \leq i < j < k$. Therefore, BE should perform well relative to TCY when covariance matrices are similar, when essentially all of the discrimination information is in the means, and when n_i/p is small. However, BE may lose classificatory information by pooling acutely dissimilar pairs of covariance matrices.

We note that BE is limited to a feature-reduction dimension that depends on the number of classes k. For $k = 2$, BE allows one to reduce the data to only one dimension regardless of the full-feature vector dimension. When $k = 2$, $\hat{\mathbf{H}}$ reduces observations to a one-dimensional reduced feature space and, therefore, one may lose discriminatory information. In general, if we have k classes, BE can reduce the feature vector to at most $q = k(k-1)/2$ dimensions because $\mathbf{U}$, given in (2.4), has $k(k-1)/2$ columns. This restriction is potentially a major disadvantage, especially in the case when k/p is small. A larger reduced dimension may be more beneficial in preserving or improving the full-dimension error rate. On the other hand, TCY allows one to reduce the original feature vector to any dimension q, $1 \leq q < p$, for any k populations.

3. A Simulation Study

We conducted a Monte Carlo simulation to compare the performance of TCY and BE using six different population configurations. We generated 1000 training and test sets from each multivariate normal distribution for each parametric configuration. We obtained estimates of the configuration parameters using the training data, and the test data were classified using the quadratic discriminant function (QDF). We computed $\hat{\mathbf{F}}_q^+$ and $\hat{\mathbf{H}}_q^+$, and found the estimated expected error rates by averaging the estimated conditional error rate over all training samples. We compared TCY and BE in terms of their estimated $EPMC$ and contrasted this with estimated $EPMC$ for the full-feature dimension. Also, we used $n_i = 2p$ and $n_i = 10p$ to determine the effect of training-sample size on the two methods.

Table 1: Description of simulation configurations

Means	Covariance Matrice	p	k	Rank **M**	Rank **U**	Non-zero singular values **M**	Non-zero singular values **U**	AGMD
Unequal	Unequal	7	2	2	1	29.01, 2.13	Rank(**U**) =1	2.19
Unequal but relatively close	Uqual	7	2	2	1	28.04, 1.92	Rank(**U**) = 1	0.51
Unequal	Unequal	6	3	2	2	9.65, 5.82	.77, 0.01	1.52
Equal	Unequal Spherical	6	3	6	2	4.24, 3.31, 1.41 1.41, 1.41,1.41	1.53, 0.92	4.52
Differ in First $p-1$ Features	Equal Elliptical	6	3	2	2	12.27, 12.13	3.15, 0.53	10.85
Differ in Last $p-1$ Features	Equal Elliptical	6	3	2	2	38.7, 24.58	0.41, 0.28	8.25

For each configuration we calculated the ranks of **M** and **U**, along with $SV(\mathbf{M})$ and $SV(\mathbf{U})$, where $SV(\mathbf{A})$ represents the set of singular values of some matrix **A**. For TCY, the number and values of the non-zero elements of $SV(\mathbf{M})$ indicate the appropriate reduced dimension q for which little classificatory information is lost. To predict the performance of BE, we calculated rank(**U**) and the *average generalized Mahalanobis distance* ($AGMD$) among the means, defined as

$$\bar{\Delta} = \sum_{1<i<j<n} \frac{2(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{n(n-1)}. \tag{3.1}$$

A relatively large value of $AGMD$ ($AGMD > 3$) indicates that most of the classificatory information lies in the means, and thus BE is more likely to perform well. In Table 1 we summarize the values of these descriptive measures for each parameter configuration.

In the following sections, we discuss the simulation results for the six configurations considered.

3.1 Rank($\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$) = 1, unequal means, $k = 2$

This configuration is composed of two multivariate normal populations with unequal population means and unequal population covariance matrices. The population parameters are $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = [3, 4, 4, 2, 2, 3, 2]^t$,

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 7 & 4 & 5 & 4 & 3 & 4 & 4 \\ 4 & 10 & 5 & 2 & 4 & 3 & 3 \\ 5 & 5 & 7 & 5 & 5 & 3 & 4 \\ 4 & 2 & 5 & 12 & 3 & 4 & 2 \\ 3 & 4 & 5 & 3 & 8 & 3 & 4 \\ 4 & 3 & 3 & 4 & 3 & 9 & 3 \\ 4 & 3 & 4 & 2 & 4 & 3 & 14 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_2 = \begin{bmatrix} 3 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 6 & 1 & -2 & 0 & -1 & -1 \\ 1 & 1 & 3 & 1 & 1 & -1 & 0 \\ 0 & -2 & 1 & 8 & -1 & 0 & -2 \\ -1 & 0 & 1 & -1 & 4 & -1 & 0 \\ 0 & -1 & -1 & 0 & -1 & 5 & -1 \\ 0 & -1 & 0 & -2 & 0 & -1 & 10 \end{bmatrix}.$$

Table 2: Rank($\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$) = 1 with unequal means ($k = 2$)

		EPMC (standard errors)	
	dim	$n = 14$	$n = 70$
Full Dimension	7	0.232 (0.001)	0.126 (0.0003)
TCY	3	0.183 (0.001)	0.130 (0.0003)
TCY	2	0.168 (0.002)	0.131 (0.0003)
TCY	1	0.160 (0.002)	0.130 (0.0003)
BE	1	0.195 (0.001)	0.143 (0.0004)

Here, rank($\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$) = 1, which implies rank($\mathbf{M}$) = 2 because $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ is not contained in span($\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$). The $SV(\mathbf{M})$ indicate that almost all classificatory information can be captured when $q = 1$ since the second singular value is small relative to the first. We expect TCY to perform well because $\mathbf{M}$ has unit rank. Due to the relatively small value of $AGMD = 2.19$, one might expect the performance of BE to be inferior to TCY.

The results are given in Table 2 and show that TCY outperforms BE for this configuration, but BE does surprisingly well.

The average PMC is actually reduced by both dimension-reduction methods when $n = 14$. The main reason for this phenomenon is that when n is small

relative to p, not enough data is available to estimate the $p(p+3)/2$ parameters for each population. By reducing the full-feature dimension p to the reduced dimension q, we can increase the ratio of n to $p(p+3)/2$ and thus obtain improved estimates for the reduced set of parameters.

3.2 Rank($\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$) = 1, unequal but relatively close means, $k = 2$

In this setting the two populations have the same covariance matrices as the configuration in 3.1, but the means are now closer together. The population mean parameters are $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = [2, 1, 0, 0, 0, 0, 0]^t$. Again, almost all classificatory information can be captured with $q = 1$ because of the relative sizes of the elements of $SV(\mathbf{M})$. A relatively small $AGMD$ indicates that a majority of the classificatory information is in the covariance matrices. Thus, BE should perform relatively poorly because it considers only the discriminatory information in the means. The results for this configuration are shown in Table 3.

As expected, BE does not perform as well in the reduced dimension $q = 1$ as TCY. For the small training-sample size, the classification results for TCY are slightly better than the full-dimension results.

Table 3: Rank($\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$) = 1 with unequal but relatively close means ($k = 2$)

		Estimated EPMC (standard errors)	
	dim	$n = 14$	$n = 70$
Full Dimension	7	0.338 (0.001)	0.228 (0.001)
TCY	3	0.323 (0.001)	0.257 (0.0006)
TCY	2	0.310 (0.001)	0.261 (0.0005)
TCY	1	0.295 (0.002)	0.259 (0.0003)
BE	1	0.375 (0.001)	0.323 (0.0005)

3.3 Unequal means, unequal covariance matrices

The third configuration we consider was earlier studied by Young, Marco, and Odell (1987). We generated training data of dimension $p = 6$ for each of the three classes using the following population means and population covariance matrices: $\boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\mu}_2 = \mathbf{e}, \boldsymbol{\mu}_3 = 2\mathbf{e}$, where $\mathbf{e}$ is a vector of ones, $\boldsymbol{\Sigma}_1 = \mathbf{I}$, $\boldsymbol{\Sigma}_2 = (1-\rho)\mathbf{I} + \rho\mathbf{e}^t\mathbf{e}$,

where $1/(1-p) \leq \rho \leq 1$ and

$$\mathbf{\Sigma}_3 = \begin{bmatrix} 2 & 1 & 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 \\ 0 & 0 & 7 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 & 1 & 1 \\ 1 & 1 & 0 & 1 & 2 & 1 \\ 1 & 1 & 0 & 1 & 1 & 2 \end{bmatrix}.$$

Because $\text{rank}(\mathbf{M}) = 2$, Theorem 1 guarantees that if the parameters are known, TCY preserves the classification information in the data when $q = 2$. Therefore, if the parameters are adequately estimated, the optimal reduced dimension should still be $q = 2$. The set $SV(\mathbf{U})$ in Table 1 indicates that $q = 1$ is the best reduced dimension for BE.

The value $AGMD = 1.52$ suggests that most of the classificatory information is in the covariance matrices, which should diminish the performance of BE. Note that the covariance matrices are relatively different: one is spherical while the other two are elliptical. Thus, the pooled estimates of the covariance matrices $(\mathbf{S}_i + \mathbf{S}_j)$ used in BE significantly differ from the individual covariance matrices, $\mathbf{S}_i,\ i = 1, 2, \ldots, k$. The simulation results are presented in Table 4.

Table 4: Unequal means and unequal covariance matrices ($k = 3$)

		Estimated EPMC (standard errors)	
	dim	$n = 12$	$n = 60$
Full Dimension	6	0.161 (0.001)	0.155 (0.0004)
TCY	3	0.244 (0.001)	0.195 (0.0005)
BE	3	0.295 (0.001)	0.233 (0.0007)
TCY	2	0.212 (0.002)	0.187 (0.0005)
BE	2	0.295 (0.002)	0.244 (0.0006)
TCY	1	0.243 (0.002)	0.243 (0.0005)
BE	1	0.305 (0.002)	0.258 (0.0005)

As expected, TCY performs better than BE regardless of the training-sample size and the reduced-dimension size q. This result is mainly due to the fact that TCY uses classificatory information in the covariance matrices that is unused in BE. Neither dimension-reduction method performs as well as the full-feature dimension. However, both methods perform at least as well at $q = 2$ dimensions than at $q = 3$ dimensions. The reason for this phenomenon is that using more dimensions as necessary results in adding "noise" or additional variability into the dimension-reduction approximation. Thus, this additional noise yields a linear

feature-selection matrix for $q = 3$ that is worse than the linear feature-selection matrix at $q = 2$ in terms of the average PMC. Also, notice that the performance of BE is fairly constant for all the reduced dimensions. The reason is that the population means are aligned in one dimension and, therefore, BE gains no information from additional dimensions.

3.4 Unequal means and unequal spherical covariance matrices

The next configuration we analyzed was considered by Friedman (1989). Compared to configuration 3.3, the population means are not aligned in a one-dimensional subspace, but the covariance matrices have a similar structure. The population mean for class Π_1 is at the origin, and the means for classes Π_2 and Π_3 are shifted from the zero vector in orthogonal directions. In addition, the parameter configurations are $\boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\mu}_2 = [3, 0, 0, 0, 0, 0]^t, \boldsymbol{\mu}_3 = [0, 4, 0, 0, 0, 0]^t, \boldsymbol{\Sigma}_1 = \mathbf{I}, \boldsymbol{\Sigma}_2 = 2\mathbf{I}$, and $\boldsymbol{\Sigma}_3 = 3\mathbf{I}$, where $\mathbf{I}$ is the six-dimensional identity matrix.

All six elements of $SV(\mathbf{M})$ are non-zero and relatively large indicating that some classificatory information may be lost with TCY when $q \leq 5$. Note that all of the covariance matrices are spherical so BE should not lose information from pooling because the covariance matrices span the same space. Also, note that $AGMD = 4.52$, which indicates that most of the discriminatory information is in the means and that the BE method should be superior. The results for this configuration are shown in Table 5.

Table 5: Unequal means and unequal spherical covariance matrices ($k = 3$)

		EPMC (standard errors)	
	dim	$n = 12$	$n = 60$
Full Dimension	6	0.092 (0.001)	0.033 (0.0003)
TCY	3	0.154 (0.001)	0.088 (0.0003)
BE	3	0.147 (0.001)	0.090 (0.0003)
TCY	2	0.186 (0.002)	0.101 (0.0008)
BE	2	0.142 (0.001)	0.094 (0.0003)
TCY	1	0.251 (0.002)	0.271 (0.001)
BE	1	0.219 (0.002)	0.184 (0.001)

For this configuration BE performs somewhat better than TCY. One reason is that BE gains from pooling the individual covariance matrices because they are proportional. Thus, all of the classification information is in the differences of the means. Therefore, TCY is actually adding noise or variability to the reduced-dimension representation by including the differences in the covariance matrices.

3.5 Means differ in first $p-1$ features, equal elliptical covariance matrices

Three classes were generated from populations with the same highly elliptic covariance matrix but different means. The eigenvalues of the common population covariance matrix $\mathbf{\Sigma}$ are $e_i = [9(i-1)/(p-1)+1]^2,\ 1 \leq i \leq p$, and the population means are $\boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\mu}_{2i} = 2.5(p-i)(\sqrt{e_i/p})/(p/2-1)$, and $\boldsymbol{\mu}_{3i} = (-1)^i \boldsymbol{\mu}_{2i}, 1 \leq i \leq p$. This configuration was initially considered by Friedman (1989).

The first components of the feature vector are the most informative. Because the common covariance matrix is highly ellipsoidal, the estimated group means differ in a low-variance space but vary in a high-variance space. Note that the elements in $SV(\mathbf{M})$ indicate that $q = 2$ should be the optimal reduced dimension for TCY. The BE method should perform well because the population covariance matrices are equal and, thus, pooling the sample covariance matrices is beneficial. Also, all of the discriminatory information is in the difference of the means as summarized by the fact that $AGMD = 10$. The results for this configuration are shown in Table 6.

Here, BE is far superior to TCY, as expected. For TCY we see that the estimated average PMC is very high for $q \leq 3$ due to the variability of the differences of the sample means. That is, the vector space spanned by $[\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1]$ can greatly vary. Therefore, a conditional reduced-feature space can be drastically different from the optimal reduced-feature space.

Table 6: Means differ in first $p-1$ features with equal elliptical covariance matrices ($k = 3$)

		EPMC (standard errors)	
	dim	$n = 12$	$n = 60$
Full Dimension	6	0.040 (0.001)	0.013 (0.0001)
TCY	3	0.313 (0.002)	0.257 (0.002)
BE	3	0.070 (0.0008)	0.034 (0.0002)
TCY	2	0.366 (0.002)	0.346 (0.002)
BE	2	0.064 (0.0007)	0.034 (0.0001)
TCY	1	0.401 (0.003)	0.401 (0.003)
BE	1	0.107 (0.001)	0.072 (0.0003)

3.6 Means differ in the Last $p-1$ features, equal elliptical covariance matrices

This example was also studied by Friedman (1989). We modeled the three populations using the same elliptic covariance matrix as in Section 3.5. However, in this configuration the means differ in a high-variance space. The parametric configuration is $\boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\mu}_{2i} = 2.5(i-1)\sqrt{e_i}/(0.5p\sqrt{p}-1)$, and $\boldsymbol{\mu}_{3i} = (-1)^i\boldsymbol{\mu}_{2i},\ 1 \le i \le p$.

All classificatory information can be captured with two dimensions. Again, we expect BE to perform better than TCY since the population covariance matrices are equal and all of the classificatory information is in the means.

For this configuration BE and TCY perform similarly (Table 7). The TCY method performs considerably better than in configuration 3.5 because the sample means now differ in a high-variance subspace and vary in a low-variance subspace. For this configuration BE benefits from the pair-wise pooling of the covariance matrices while TCY benefits from increased stability in the sample means.

Table 7: Means differ in last $p-1$ features and equal elliptical covariance matrices ($k=3$)

		EPMC (standard errors)	
	dim	$n=12$	$n=60$
Full Dimension	6	0.049 (0.001)	0.015 (0.0001)
TCY	3	0.068 (0.0008)	0.040 (0.0002)
BE	3	0.074 (0.0008)	0.036 (0.0002)
TCY	2	0.091 (0.002)	0.055 (0.0007)
BE	2	0.079 (0.001)	0.036 (0.0003)
TCY	1	0.211 (0.003)	0.191 (0.002)
BE	1	0.206 (0.003)	0.167 (0.002)

4. A Parametric Bootstrap Simulation

In the following simulation, we use a real data set to estimate the population means and covariance matrices. We perform our Monte Carlo simulation with a parametric bootstrap using three populations: $N(\bar{\mathbf{x}}_1, \mathbf{S}_1), N(\bar{\mathbf{x}}_2, \mathbf{S}_2)$, and $N(\bar{\mathbf{x}}_3, \mathbf{S}_3)$.

The data set considered is from a preliminary study by G. R. Bryce and R.M. Barker at Brigham Young University (Rencher, 1995) on a possible link between football helmet design and neck injuries. Six different head measurements were

taken on each individual, and the study included three classes with thirty subjects in each class. The three classes are

Π_1 High-school football players,
Π_2 College football players,
Π_3 Non-football players.

The six head measurements are

1. Head-width at widest dimension,
2. Head circumference,
3. Front-to-back measurement at eye level,
4. Eye-to-top-of-head measurement,
5. Ear-to-top-of-head measurement,
6. Jaw width.

The estimated means and covariance matrices used as parameters in our parametric bootstrap are

$$\bar{\mathbf{x}}_1 = [15.42, 57.38, 19.80, 10.08, 13.45, 11.94]^t,$$
$$\bar{\mathbf{x}}_2 = [15.20, 58.94, 20.11, 13.08, 14.73, 12.27]^t,$$
$$\bar{\mathbf{x}}_3 = [15.58, 57.77, 19.81, 10.95, 13.70, 11.80]^t,$$

$$\mathbf{S}_1 = \begin{bmatrix} .545 & .541 & .172 & .233 & .176 & .247 \\ .541 & 4.21 & 1.43 & .780 & ,860 & .720 \\ .172 & 1.43 & .706 & .211 & .414 & .233 \\ .233 & .779 & .211 & 1.09 & .540 & .175 \\ .176 & .860 & .414 & .540 & .892 & .082 \\ .247 & .720 & .233 & .175 & .082 & .478 \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} .407 & .618 & .195 & -.232 & .113 & .255 \\ .618 & 2.88 & .929 & .195 & .094 & .308 \\ .195 & .929 & .552 & -.063 & -.001 & .128 \\ -.232 & .195 & -.063 & 1.15 & .087 & -.157 \\ .113 & .094 & -.001 & .087 & .570 & -.008 \\ .255 & .308 & .128 & -.157 & -.008 & .377 \end{bmatrix}$$

$$\mathbf{S}_3 = \begin{bmatrix} .333 & .575 & .107 & .251 & .085 & .182 \\ .575 & 2.39 & .700 & .985 & .066 & .487 \\ .107 & .700 & .380 & .083 & -.027 & .116 \\ .251 & .985 & .083 & 1.46 & .317 & .109 \\ .085 & .066 & -.027 & .317 & .392 & -.047 \\ .182 & .487 & .116 & .109 & -.047 * ** & .271 \end{bmatrix}.$$

For this data set, $\text{rank}(\mathbf{M}) = 6$ and $SV(\mathbf{M}) = \{5.05, 1.82, 0.68, 0.50, 0.280.21\}$, which indicates that the first two dimensions contain almost all of the dis-

criminatory information. This observation suggests $q = 2$ is the best reduced-dimension choice for TCY. Also, we have that $AGMD = 2.90$ and $SV(\mathbf{U}) = \{2.91, 0.95, 0.31\}$. Since $AGMD$ is moderately large, BE may perform reasonably well for this configuration. Given that there is little information in the third reduced dimension, it is unlikely that BE would benefit by adding a third reduced dimension.

The results of the parametric bootstrap simulation are presented in Table 8. The two methods perform similarly for this configuration. As predicted, TCY performs better when $q = 2$, and BE performs surprisingly well, considering the moderate value of $AGMD = 2.90$. However, neither method improves the misclassification error when compared to the full dimension. The gain in the training-sample size to parameter-dimension ratio is offset by a loss of information in the reduced-feature space.

Table 8: Football helmet study ($k = 3$)

		EPMC (standard errors)	
	dim	$n = 12$	$n = 60$
Full Dimension	6	0.132 (0.001)	0.100 (0.0004)
TCY	3	0.235 (0.001)	0.198 (0.0007)
BE	3	0.224 (0.001)	0.169 (0.0006)
TCY	2	0.218 (0.001)	0.175 (0.0007)
BE	2	0.224 (0.001)	0.160 (0.0006)
TCY	1	0.251 (0.002)	0.206 (0.001)
BE	1	0.243 (0.002)	0.214 (0.001)

We first note that at $q = 3$ neither linear dimension-reduction technique yields a $EPMC$ close to the full feature $EPMC$. Also, in view of the moderate value of $AMGD$, the BE linear feature-selection method is, somewhat surprisingly, roughly equivalent to the TCY linear feature-selection method in terms of the reduced-space average PMCs.

5. Concluding Remarks

We first remark that BE benefits from pooling the pairs of covariance matrices when they are similar. The performance of BE is enhanced if most of the classificatory information is contained in the means. This is achieved through the rotation of the pairs of means by $(\mathbf{S}_i - \mathbf{S}_j)^{-1}$ into a feature space that preserves or nearly preserves the $AGMD$.

Also, TCY performs well when rank($\mathbf{M}$) is relatively small so that $q \ll p$. We note that TCY does not depend on which population is labeled Π_1. In deciding whether to use TCY or BE as a feature-reduction method, a researcher may choose to apply both transformations to the data and select the one that performs the best on the specific data set in terms of yielding the smallest estimated conditional error rate. In configurations 3.1 and 3.2, classification is actually enhanced by feature reduction when the sample size is close to the number of parameters to be estimated. Last, we remark that unlike the results in Brunzell and Eriksson (2000), our results demonstrate that for certain combinations of parameter configurations and sample sizes, TCY can be significantly superior to BE.

Finally, we note that the simulation studies indicate that the performance of the TCY and BE can be reasonably predicted when one considers the values of the configuration parameters and the sets $SV(\mathbf{M})$ and $SV(\mathbf{M})$, and the $AGMD$.

Acknowledgements

The authors would like to thank an anonymous referee for many insightful comments that helped to improve the exposition of this paper.

References

Aeberhard, S., de Vel, O., and Coomans, D. (1994). Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition* **27**, 1065-1077.

Brunzell, H. and Eriksson, J. (2000). Feature reduction for classification of multidimensional data. *Pattern Recognition* **33**, 1741-1748.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84**, 165-175.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. John Wiley and Sons.

Tubbs, J. D., Coberly, W., and Young, D. M. (1982). Linear dimension reduction and Bayes classification with unknown population parameters. *Pattern Recognition* **15**, 167-172.

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.

Received August 24, 2004; accepted November 24, 2004.

J. Wade Davis (corresponding author)
Department of Statistical Science
Baylor University
Waco, TX 76798-7140, USA
Wade_Davis@Baylor.edu

Dean M. Young
Department of Statistical Science
Baylor University
Waco, TX 76798-7140, USA
Dean_Young@Baylor.edu

Jeanne S. Hill
Department of Statistical Science
Baylor University
Waco, TX 76798-7140, USA
Jeanne_Hill@Baylor.edu

Karin B. Ernstrom-Keim
Department of Family and Preventive Medicine
University of California
San Diego, CA 92093-0717, USA
Kernstrom@ucsd.edu