Submitted to Biometrika

Maximal Data Piling in Discrimination

Jeongyoun Ahn
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
Email: jyahn@email.unc.edu

J. S. Marron
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
Email: marron@email.unc.edu

December 3, 2004

Abstract

In a binary discrimination problem, a linear classifier finds a linear hyperplane that separates two classes by partitioning the data space. Especially in a High Dimension Low Sample Size (HDLSS) setting, there are linear separating hyperplanes such that the projections of the training data points onto their normal direction vectors are identically zero, or some non-zero constant. Of interest in this paper is a linear separating hyperplane such that the projections of the training data points from each class onto its normal direction vector have two distinct values, one for each class. This direction vector is uniquely defined in the subspace generated by the data. A simple formula is given to find this direction. In non-HDLSS settings, this direction vector is the same as the Fisher Linear Discrimination direction vector.

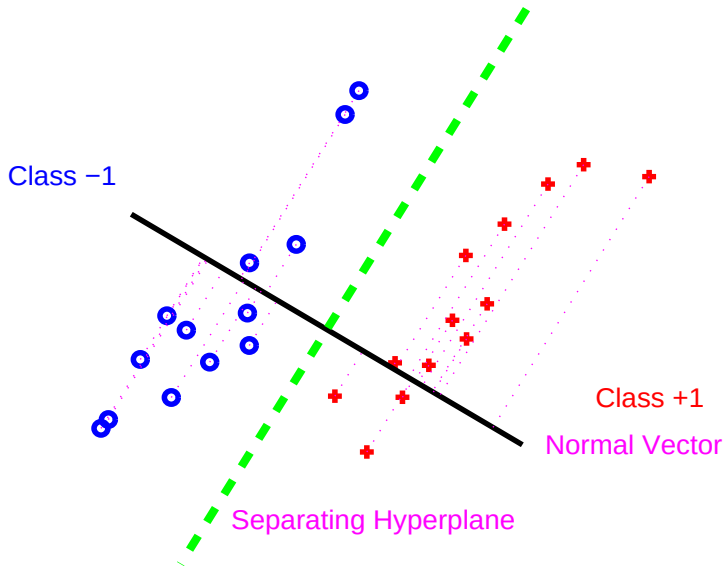

Figure 1: Separable toy data in two dimensions, with a separating hyperplane, shown as the dashed line, and projections onto the normal vector (solid line).

1 Introduction

Suppose we have a data set of sample size n and each data vector is a d-dimensional vector in Euclidean space $\mathbb{R}^d$, which we call the data space. In a binary classification (i.e. discrimination) problem, a classifier partitions the data space into two spaces. A simple partition is done by a linear classifier which creates a separating hyperplane between two linearly partitioned spaces (see Duda et al. (2000) for further overview.) We say a training data set is linearly separable if there exists a separating hyperplane that classifies all training data points into the correct partitioned space. If the underlying distribution of the data is continuous in the data space and the dimension is larger than the sample size (High Dimension Low Sample Size (HDLSS)), the data are separable with probability one. Figure 1 shows a separable toy data set in $\mathbb{R}^2$ with a separating hyperplane. It also shows the projections of the data points onto the normal vector of the hyperplane, which will be discussed shortly.

In the simplest separable case, the Support Vector Machine (SVM) linear classifica-

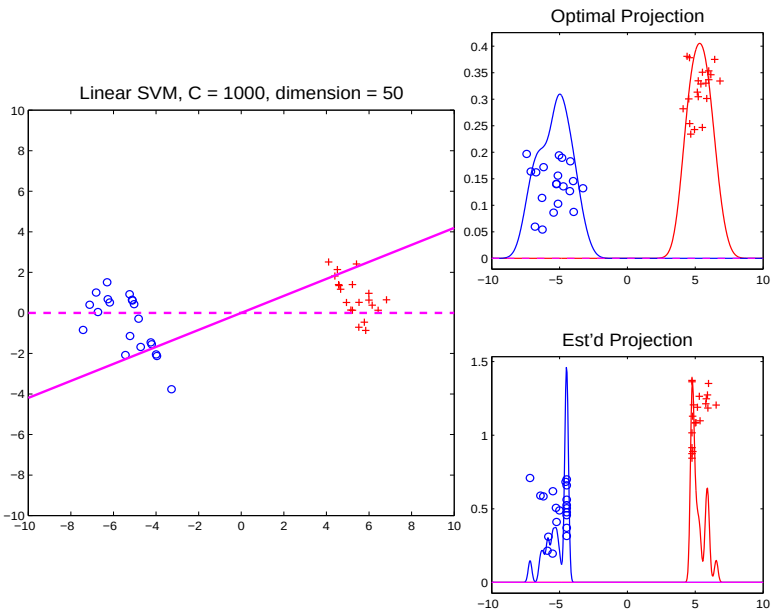

Figure 2: Toy data example illustrating data piling for the Support Vector Machine. The SVM direction vector is shown as a solid line and the optimal direction vector as a dashed line.

tion rule seeks a separating hyperplane that maximizes the *margin*, the minimum distance between the two convex hulls of data points from each class (see Hastie et al. (2001), Cristianini and Shawe-Taylor (2000) for a more detailed introduction.) In many applications of SVM in HDLSS settings, we observe that a large portion of the data are support vectors, i.e. the data points lie on the margin boundaries (Section 2). Thus, if we project the data points onto the normal vector of the SVM hyperplane, then many of the projections are identical, which is what we call *data piling*. Data piling is usually not a desirable property for a classifier since it indicates that the separating hyperplane may be unduly influenced by noise artifacts in the data. Figure 2 illustrates data piling for the SVM method. The toy data set is of size 40, out of which 20 belong to each class. They are generated from a spherical, unit variance Gaussian distribution with dimension 50, and mean 0, except that the first coordinate has mean $+5.2$ for Class $+1$, -5.2 for Class -1. The left plot shows the projections of the data onto the 2-d plane generated by the normal vector of the SVM hyperplane and the normal vector of the optimal hyperplane, which is $(1, 0, \cdots, 0)^{\mathrm{T}}$.

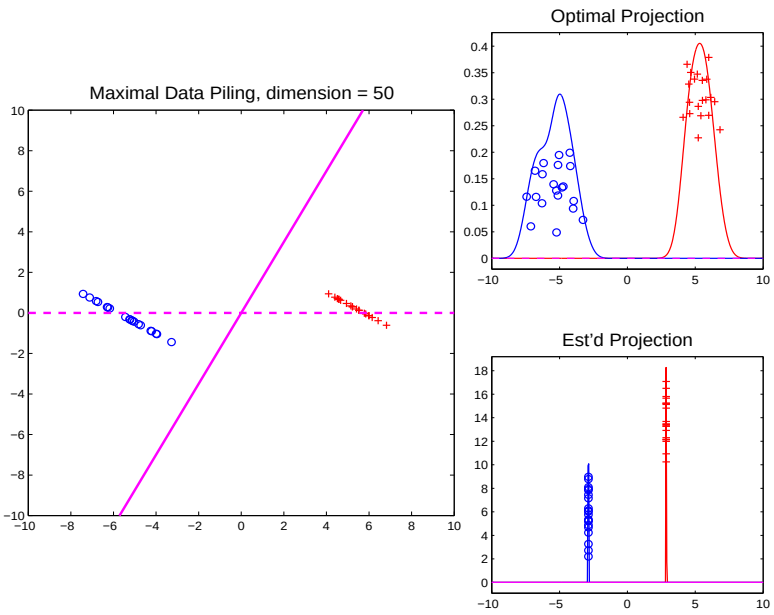

Figure 3: Toy data example illustrating data piling for the Maximal Data Piling. The MDP direction vector shown as a solid line and the optimal direction vector as a dashed line.

The upper right plot shows the projections onto the optimal direction, shown with the dashed line in the left panel, and the bottom right one shows the projections onto the SVM direction, shown with the solid line in the left panel. Each plot is represented as a "jitter plot," (Tukey and Tukey (1990)) with a random vertical coordinate for visual separation of the data points. Also kernel density estimation curves are drawn to show how the projections are distributed for each case. The bottom right plot shows that there are (perhaps) too many support vectors at the margin boundaries, i.e. there is severe data piling.

Suppose that $d > n$. In this setting, there are direction vectors where the data collapse, i.e. the projections of the data points onto those direction vectors are identical. For example, since the data generate an n-dimensional subspace, the projections of the data points onto any direction vector in the $d-n$ dimensional orthogonal subspace are all zeros. Furthermore, the projections of the data points onto any vectors orthogonal to the hyperplane generated by the data are possibly non-zero constants. It is seen below that

in the discrimination problem, there usually is a direction vector such that the two classes project to two different values. This vector is called the *Maximal Data Piling* (MDP) direction vector. The term "Maximal" is used to indicate that the MDP direction vector maximizes the amount of data piling, as discussed in the previous paragraph. Note that the linear combinations of any vector orthogonal to the hyperplane generated by the data and the MDP direction vector can maximize the amount of data piling as well, however, it is seen in Section 3.1 that the MDP direction vector is uniquely defined within the subspace generated by the data.

The MDP direction vector exists if the data vectors are linearly independent, which is satisfied with probability one when the underlying distribution of the data is continuous in $\mathbb{R}^d$. The MDP result for the same toy example in Figure 2 is shown in Figure 3. Note that the data projections completely pile up at two points and the distance between the two data piling points is smaller than the distance for SVM, as shown in Figure 2. This is not surprising, since SVM seeks to maximize this distance. Also note that the angle between the MDP direction vector and the optimal direction vector is larger than the SVM - optimal angle shown in Figure 2, which means the MDP has a worse discrimination error rate than SVM for this Gaussian example. We show in Section 3.1 that the MDP direction vector lies within the hyperplane generated by all the training samples, yet is orthogonal to both of the hyperplanes generated by the separate training samples from each class.

In addition, the MDP direction vector is characterized as the product of the generalized inverse of the global sample covariance matrix and the mean difference vector. Here, the global sample covariance matrix is obtained by using the global sample mean calculated from all the samples instead of respective sample means from each class. That is, it is the Fisher Linear Discrimination (FLD) direction vector with the global sample covariance matrix instead of the pooled one. It turns out that the MDP and FLD direction vectors are actually identical in non-HDLSS settings (Section 3.2).

A comparison to other simple linear classification methods such as Mean Difference (MD) (i.e. centroid method, as in Hastie et al. (2001)), FLD, and Naive Bayes (Bickel and Levina (2003)) in terms of the classification performances is done by a simulation study in Section 4. In the simulation we consider various combinations of the sample size, dimension, and the distance between two classes in a Gaussian setting. It is seen that the error rates of MDP and FLD are largest when the dimension is close to the sample size and this phenomenon is discussed in detail there.

2 Data Piling in a Real Example

In this section a real data example is presented to demonstrate data piling. The data are microarray gene expressions from the UNC breast cancer data base. See Perou et al. (2000) for the details regarding this data set. As for many other microarray data sets, this is a HDLSS setting with 5,705 genes and 105 breast cancer patients, of whom 71 survived and 34 died. Here we consider a linear classification problem using this data with the mortality as the target variable and apply the SVM, Distance Weighted Discrimination (DWD) (Marron et al. (2004)), and the MDP. DWD is a natural method for HDLSS discrimination because it aims to avoid the data piling problem of SVM.

The projections of the data points onto the MDP, SVM, DWD, and MD direction vectors are shown on the diagonal panels of Figure 4, and the projections onto the 2-d planes generated by each pair of the direction vectors are shown on the off-diagonal panels of the same figure. In each plot, the circles represent the samples that survived and the plusses are for the samples that died. The first diagonal panel shows that the projections of each class onto the MDP direction vector pile up completely at two points, one for each class, and the distance between them is 9.00. The second diagonal panel shows a very large amount of data piling for the SVM direction vector. The projections of the group that died piles up completely near 6 and most of the group that survived piles up around -3, with the distance between the piling points 9.09. The projections of each group onto the DWD direction vector shown in the third diagonal panel, on the other hand, shows no piling at all, and the distance between the two peaks is a little more than ten. The 2-d projections on the off-diagonal panels highlight relationships between these directions. In particular, MDP and SVM are quite similar to each other (e.g. the angle between them is small), with only a slight rotation being the difference between substantial data piling (SVM) and complete data piling (MDP). DWD is rather close to both of those, in terms of small angle. The fact that the MD direction is very different is reflected in the much larger relative angles.

The ten-fold cross validation error rates of the four methods are given in Table 1. Consistent with the above discussions, MDP has a substantially worse error rate. MD is also substantially worse. The SVM shows slightly better performance than DWD. Hall et al. (2002) pointed out a need for improvement of DWD with unequal sample sizes of each class, which may be the reason for the inferior performance of DWD. Since this data have a large difference in the samples from each class, an improved version of DWD (Zhang et al. (2004)) may be useful in this problem, which is not pursued further in this

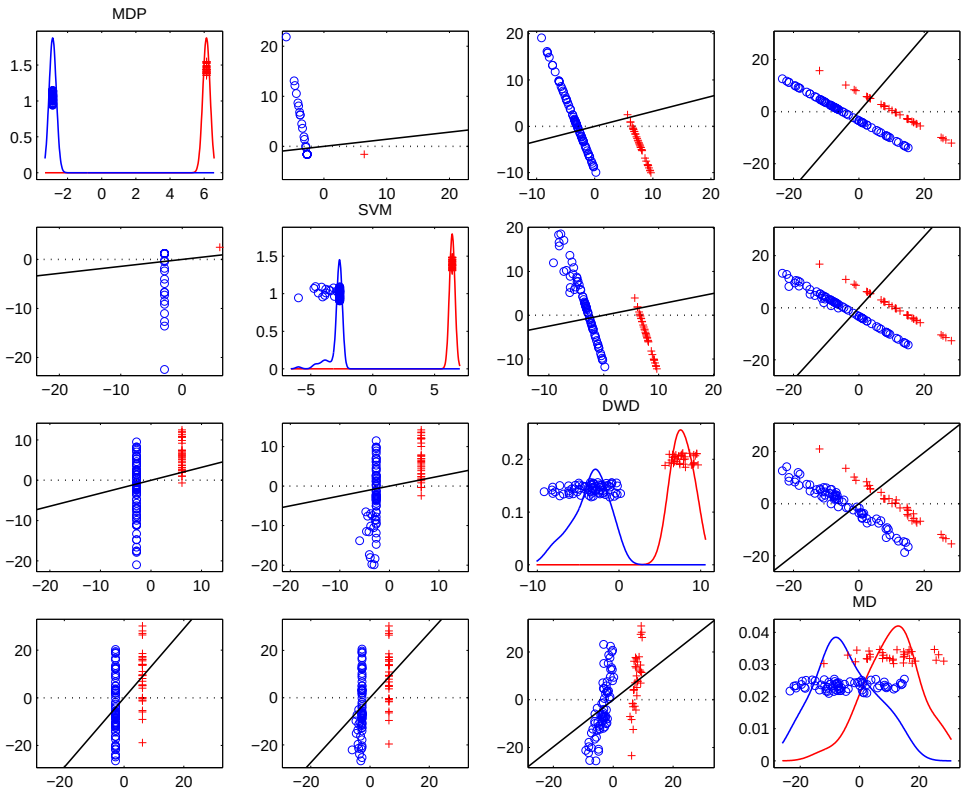

Figure 4: Projections of microarray data onto MDP, SVM, DWD, and MD direction vectors, with 1-d projections onto each direction vectors on the diagonal panels and 2-d projections onto the plane generated by each pair of the direction vectors on the off-diagonal panels

paper.

MDP	SVM	DWD	MD
0.4999	0.2811	0.2879	0.3561

Table 1: Misclassification error rates of MDP, SVM, DWD, and MD

For this particular data set, the accuracy of labels is an issue. There are likely to be some mislabelled samples in the sense that the patients who are about to die were categorized as survived when their gene expression may be more closely connected with the patients who died. Johnson et al. (2004) developed a classification method using DWD in order to deal with gene expression survival data with possible mislabeling error.

3 Maximal Data Piling

In this section we express the MDP direction vector in a closed form and we show it is uniquely defined in the data space. Also we show it is the same as the FLD direction vector when the dimension d is less than the sample size -1, i.e., $d < n - 1$.

Let $X = (\mathbf{x}_1, \cdots, \mathbf{x}_{n_1})$ be the matrix of the training data from Class $+1$, where $\mathbf{x}_i$'s are iid from a continuous probability distribution in $\mathbb{R}^d$, and define $Y = (\mathbf{y}_1, \cdots, \mathbf{y}_{n_2})$ for Class -1 in a similar way. Let $n = n_1 + n_2$. Define the $d \times n$ combined data matrix as

$$Z := [X, Y], \tag{1}$$

where $[X, Y]$ denotes the horizontal concatenation of X and Y. Denote the global mean vector of the combined data by $\bar{\mathbf{z}}$. Then the sample covariance matrix of Z is

$$\widetilde{\Sigma} := \frac{1}{n-1}[(Z - \bar{Z})(Z - \bar{Z})^{\mathrm{T}}]. \tag{2}$$

Here, $\bar{Z} = \bar{\mathbf{z}}\mathbf{1}_n^{\mathrm{T}}$, where $\mathbf{1}_n$ is a column vector of ones of length n. We will call this matrix the "global sample covariance matrix" and denote it by $\widetilde{\Sigma}$.

Definition 1. *The Maximal Data Piling (MDP) direction vector is defined as*

$$\mathbf{v}_{\mathrm{MDP}} := \frac{\tilde{\Sigma}^{-}(\bar{\mathbf{x}} - \bar{\mathbf{y}})}{\|\tilde{\Sigma}^{-}(\bar{\mathbf{x}} - \bar{\mathbf{y}})\|}, \tag{3}$$

where A^{-} is the Moore-Penrose generalized inverse of a matrix A.

3.1 Specification of MDP Direction Vector in the Data Space

The data vectors in the matrix Z, i.e. the columns of Z, generate a subspace in $\mathbb{R}^d$ and this subspace is expressed as

$$\mathcal{S}_Z = \{Z\mathbf{w} : \mathbf{w} \in \mathbb{R}^n\}. \tag{4}$$

In other words, $\mathcal{S}_Z$ is the set of all linear combinations of the data vectors. Note that $\mathcal{S}_Z$ has dimension n. If we let $\widetilde{\mathcal{H}}_Z$ be the hyperplane generated by Z, then we can write

$$\widetilde{\mathcal{H}}_Z = \{Z\mathbf{u} : \mathbf{u}^{\mathrm{T}}\mathbf{1}_n = 1, \mathbf{u} \in \mathbb{R}^n\}. \tag{5}$$

Note that $\widetilde{\mathcal{H}}_Z$ is a set of linear combinations of data points of which the sum of the coefficients is 1. The parallel subspace can be found by shifting the hyperplane so that it goes through the origin. A natural shift is via the point in $\widetilde{\mathcal{H}}_Z$ that is closest to the origin which is calculated in the following lemma.

Lemma 2. *Let $\mathbf{v}_Z$ be the point in $\widetilde{\mathcal{H}}_Z$ that is nearest to the origin. Then,*

$$\mathbf{v}_Z = \frac{Z(Z^TZ)^{-1}\mathbf{1}_n}{\mathbf{1}_n^T(Z^TZ)^{-1}\mathbf{1}_n}.$$

Proof. Since $\mathbf{v}_Z$ is on the hyperplane $\widetilde{\mathcal{H}}_Z$, it can be expressed in the form $\mathbf{v}_Z = Z\mathbf{u}$, where $\mathbf{u}^{\mathrm{T}}\mathbf{1}_n = 1, \mathbf{u} \in \mathbb{R}^n$ by (5). The squared distance from the origin to $\mathbf{v}_Z$ is $\mathbf{u}^{\mathrm{T}}Z^{\mathrm{T}}Z\mathbf{u}$ and we need to find $\mathbf{u}$ that minimizes this distance. The Lagrangian (see Chapter 5 in Cristianini and Shawe-Taylor (2000)) of this minimization problem is

$$L(\mathbf{u}) = \frac{1}{2}\mathbf{u}^{\mathrm{T}}Z^{\mathrm{T}}Z\mathbf{u} - \alpha(\mathbf{1}_n^{\mathrm{T}}\mathbf{u} - 1),$$

where $\alpha > 0$ is the Lagrangian multiplier. From $\partial L(\mathbf{u})/\partial\mathbf{u} = 0$,

$$\mathbf{u} \;=\; \alpha(Z^{\mathrm{T}}Z)^{-1}.$$

From $\mathbf{u}^{\mathrm{T}}\mathbf{1}_n = 1$,

$$\alpha \;=\; \frac{1}{\mathbf{1}_n^{\mathrm{T}}(Z^{\mathrm{T}}Z)^{-1}\mathbf{1}_n}.$$

Thus,

$$\mathbf{v}_Z = \frac{Z(Z^{\mathrm{T}}Z)^{-1}\mathbf{1}_n}{\mathbf{1}_n^{\mathrm{T}}(Z^{\mathrm{T}}Z)^{-1}\mathbf{1}_n}.$$

□

Now let us shift the hyperplane $\widetilde{\mathcal{H}}_Z$ so that it contains the origin and call the new shifted hyperplane $\mathcal{H}_Z$. Because $\mathcal{H}_Z = \widetilde{\mathcal{H}}_Z - \mathbf{v}_Z$,

$$\begin{aligned}\mathcal{H}_Z &= \left\{ Z\mathbf{u}^* : \mathbf{u}^* = \mathbf{u} - \frac{(Z^{\mathrm{T}}Z)^{-1}\mathbf{1}_n}{\mathbf{1}_n^{\mathrm{T}}(Z^{\mathrm{T}}Z)^{-1}\mathbf{1}_n}, \mathbf{u}^{\mathrm{T}}\mathbf{1}_n = 1, \mathbf{u} \in \mathbb{R}^n \right\} \\ &= \left\{ Z\mathbf{u}^* : \mathbf{u}^{*\mathrm{T}}\mathbf{1}_n = 0, \mathbf{u}^* \in \mathbb{R}^n \right\}. \end{aligned} \tag{6}$$

Note that $\mathcal{H}_Z$ is a subspace of $\mathbb{R}^d$ with dimension $n-1$ and we can decompose $\mathcal{S}_Z$ into an orthogonal sum of $\mathcal{H}_Z$ and $\{\mathbf{v}_Z\}$, i.e. $\mathcal{S}_Z = \mathcal{H}_Z \oplus \{\mathbf{v}_Z\}$.
In the same fashion we can define subspaces parallel to the hyperplanes of X and Y, call them $\mathcal{H}_X$ and $\mathcal{H}_Y$, respectively. They have the following expressions:

$$\begin{aligned}\mathcal{H}_X &= \left\{ X\boldsymbol{\nu}_1^* : \boldsymbol{\nu}_1^* = \boldsymbol{\nu}_1 - \frac{(X^{\mathrm{T}}X)^{-1}\mathbf{1}_{n_1}}{\mathbf{1}_{n_1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\mathbf{1}_{n_1}}, \boldsymbol{\nu}_1^{\mathrm{T}}\mathbf{1}_{n_1} = 1, \boldsymbol{\nu}_1 \in \mathbb{R}^{n_1} \right\} \\ &= \left\{ X\boldsymbol{\nu}_1^* : \boldsymbol{\nu}_1^{*\mathrm{T}}\mathbf{1}_{n_1} = 0, \boldsymbol{\nu}_1^* \in \mathbb{R}^{n_1} \right\}, \end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{H}_Y &= \left\{ Y\boldsymbol{\nu}_2^* : \boldsymbol{\nu}_2^* = \boldsymbol{\nu}_2 - \frac{(Y^{\mathrm{T}}Y)^{-1}\mathbf{1}_{n_2}}{\mathbf{1}_{n_2}^{\mathrm{T}}(Y^{\mathrm{T}}Y)^{-1}\mathbf{1}_{n_2}}, \boldsymbol{\nu}_2^{\mathrm{T}}\mathbf{1}_{n_2} = 1, \boldsymbol{\nu}_2 \in \mathbb{R}^{n_2} \right\} \\ &= \left\{ Z\boldsymbol{\nu}_2^* : \boldsymbol{\nu}_2^{*\mathrm{T}}\mathbf{1}_{n_2} = 0, \boldsymbol{\nu}_2^* \in \mathbb{R}^{n_2} \right\}. \end{aligned} \tag{8}$$

We can show these two subspaces of X and Y are actually the subspace of $\mathcal{H}_Z$ in the following lemma.

Lemma 3. *Let $\mathcal{H}_Z$, $\mathcal{H}_X$, and $\mathcal{H}_Y$ be as defined in (6), (7), and (8), respectively. Then both $\mathcal{H}_X$ and $\mathcal{H}_Y$ are subspaces of $\mathcal{H}_Z$.*

Proof. This can be shown by setting the $\mathbf{u}^*$ in (6) to $\mathbf{u}^* = [\boldsymbol{\nu}_1^*; \mathbf{0}_{n_2}]$ for $\mathcal{H}_X$ and $\mathbf{u}^* = [\mathbf{0}_{n_1}; \boldsymbol{\nu}_2^*]$ for $\mathcal{H}_Y$, where ";" denotes the vertical concatenation of two vectors. □

Now we have a theorem regarding where the MDP direction vector $\mathbf{v}_{\mathrm{MDP}}$(as defined at (3)) lies within the data space.

Theorem 4. *The maximal data piling vector $\mathbf{v}_{\mathrm{MDP}}$ is a member of $\mathcal{H}_Z$ and orthogonal to the subspaces $\mathcal{H}_X$ and $\mathcal{H}_Y$. i.e.*

$$\mathcal{H}_Z = \{\mathcal{H}_X + \mathcal{H}_Y\} \oplus \{\mathbf{v}_{\mathrm{MDP}}\}.$$

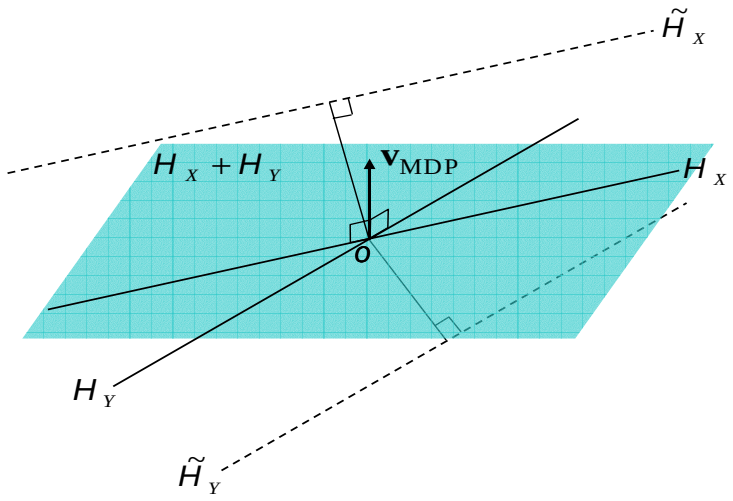

Figure 5: The illustration of $\mathcal{H}_X$, $\mathcal{H}_Y$, and $\mathbf{v}_{\mathrm{MDP}}$ when $d = 3$, $n_1 = 2$, and $n_2 = 2$.

Figure 5 shows the geometric relationship among $\mathcal{H}_X$, $\mathcal{H}_Y$, and $\mathbf{v}_{\mathrm{MDP}}$ when the dimension of the data is three and the number of data points is two for each group, i.e. $d = 3$, $n_1 = 2$, and $n_2 = 2$. Note that the subspaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ are actually one-dimensional straight lines and they are not necessarily orthogonal to each other. The hyperplanes $\widetilde{\mathcal{H}}_X$ and $\widetilde{\mathcal{H}}_Y$ are shifted to $\mathcal{H}_X$ and $\mathcal{H}_Y$ so they meet each other at the origin. Note that they do not necessarily cross each other before the shift.

To prove this theorem we need the following lemma:

Lemma 5. *Let A be $d \times m (m < d)$ matrix with rank$(A) = m - 1$ such that the sum of each row is zero. And let A^- denote the Moore-Penrose generalized inverse of A. Then*

$$A^- A = I_m - \frac{1}{m} J_m,$$

where $J_m = \mathbf{1}_m \mathbf{1}_m^T$.

Proof. Consider the singular value decomposition of A:

$$A = UDV^{\mathrm{T}},$$

where

$$\begin{aligned} U_{(d\times m)} &= (\mathbf{u}_1, \cdots, \mathbf{u}_m), \\ D_{(m\times m)} &= \mathrm{diag}(d_1, \cdots, d_{m-1}, 0), \text{and} \\ V_{(m\times m)} &= (\mathbf{v}_1, \cdots, \mathbf{v}_m). \end{aligned}$$

Note that the columns of U and V form an orthonormal basis in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively. Especially, V is obtained from the eigenvalue decomposition of $A^{\mathrm{T}}A$:

$$A^{\mathrm{T}}A = VD^2V^{\mathrm{T}}.$$

Here the columns of the eigenvector matrix V span the row space of A. Hence the sum of the coefficients of $\mathbf{v}_i, (i = 1, \cdots, m-1)$ is zero and the last column $\mathbf{v}_m$ is $(m^{-1/2}, \cdots, m^{-1/2})^{\mathrm{T}}$ to satisfy orthogonality condition on columns. Now, since

$$A^- = (\mathbf{v}_1, \cdots, \mathbf{v}_{m-1}) \begin{pmatrix} d_1^{-1} & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & d_{m-1}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{u}_{m-1}^{\mathrm{T}} \end{pmatrix},$$

we have

$$A^-A = (\mathbf{v}_1, \cdots, \mathbf{v}_{m-1}) \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_{m-1}^{\mathrm{T}} \end{pmatrix}$$

by the orthogonality of the matrix U. It follows from the fact $VV^{\mathrm{T}} = I_m$, that

$$\begin{aligned} A^-A &= \mathbf{v}_1\mathbf{v}_1^{\mathrm{T}} + \cdots + \mathbf{v}_{m-1}\mathbf{v}_{m-1}^{\mathrm{T}} \\ &= I_m - \mathbf{v}_m\mathbf{v}_m^{\mathrm{T}} \\ &= I_m - \frac{1}{m}J_m. \end{aligned}$$

□

Proof of Theorem 4. Note that each member of $\mathcal{H}_Z$ is a linear combination of data points where the sum of the coefficients is zero. Since

$$(Z - \bar{Z}) = Z\ \left(I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right),$$

and

$$(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = Z\ \left(\frac{1}{n_1}, \cdots, \frac{1}{n_1}, -\frac{1}{n_2}, \cdots, -\frac{1}{n_2}\right)^{\mathrm{T}},$$

$\mathbf{v}_{MDP}$, defined in (3), is in $\mathcal{H}_Z$.

Now let $\boldsymbol{\nu}$ be $[\boldsymbol{\nu}_1^*; \boldsymbol{\nu}_2^*]$ with $\boldsymbol{\nu}_1^*$ and $\boldsymbol{\nu}_2^*$ as defined in (7) and (8), respectively, and let $\widetilde{Z}$ be the centered version of Z, $(Z - \bar{Z})$. To show that $\{\mathbf{v}_{\text{MDP}}\} \perp \mathcal{H}_X$ and $\{\mathbf{v}_{\text{MDP}}\} \perp \mathcal{H}_Y$, it suffices to show that the inner product of $Z\,\boldsymbol{\nu}$ and $\mathbf{v}_{\text{MDP}}$ is zero.

$$\begin{aligned} < Z\boldsymbol{\nu}, \mathbf{v}_{\text{MDP}} > &= \boldsymbol{\nu}^{\text{T}} Z^{\text{T}} \mathbf{v}_{\text{MDP}} \\ &\propto \boldsymbol{\nu}^{\text{T}} Z^{\text{T}} (\widetilde{Z}\widetilde{Z}^{\text{T}})^{-} (\bar{\mathbf{x}} - \bar{\mathbf{y}}) \\ &\propto \boldsymbol{\nu}^{\text{T}} Z^{\text{T}} (\widetilde{Z}\widetilde{Z}^{\text{T}})^{-} Z[\mathbf{1}_{n_1}; -\mathbf{1}_{n_2}] \\ &= \boldsymbol{\nu}^{\text{T}} \widetilde{Z}^{\text{T}} (\widetilde{Z}\widetilde{Z}^{\text{T}})^{-} \widetilde{Z}[\mathbf{1}_{n_1}; -\mathbf{1}_{n_2}]. \end{aligned}$$

By p.222 in Searle (1982),

$$\widetilde{Z}^{\text{T}} (\widetilde{Z}\widetilde{Z}^{\text{T}})^{-} \widetilde{Z} = \widetilde{Z}^{-} \widetilde{Z}.$$

Now by Lemma 5,

$$\begin{aligned} < Z\boldsymbol{\nu}, \mathbf{v}_{\text{MDP}} > &= \boldsymbol{\nu}^{\text{T}} \widetilde{Z}^{-} \widetilde{Z}[\mathbf{1}_{n_1}; -\mathbf{1}_{n_2}] \\ &= \boldsymbol{\nu}^{\text{T}} (I_n - \frac{1}{n} J_n)[\mathbf{1}_{n_1}; -\mathbf{1}_{n_2}] \\ &= 0. \end{aligned}$$

□

3.2 Relation to the Fisher Linear Discrimination

In this section we compare the MDP direction vector with the Fisher Linear Discrimination (FLD) direction vector when $d < n - 1$. The FLD direction vector is defined as follows:

$$\mathbf{v}_{\text{FLD}} := \frac{\widehat{\Sigma}^{-}(\bar{\mathbf{x}} - \bar{\mathbf{y}})}{\|\widehat{\Sigma}^{-}(\bar{\mathbf{x}} - \bar{\mathbf{y}})\|}, \tag{9}$$

where A^{-} is the Moore-Penrose generalized inverse of the matrix A and $\widehat{\Sigma}$ is the *pooled* sample covariance matrix, i.e.

$$\widehat{\Sigma} = \frac{1}{n-2} \left[(X - \bar{X})(X - \bar{X})^{\text{T}} + (Y - \bar{Y})(Y - \bar{Y})^{\text{T}} \right], \tag{10}$$

where $\bar{X} = \bar{\mathbf{x}}\mathbf{1}_{n_1}^{\text{T}}$ and $\bar{Y} = \bar{\mathbf{y}}\mathbf{1}_{n_2}^{\text{T}}$. Note that the only difference between the MDP and FLD direction vectors, is whether one uses the *global* sample covariance matrix $\widetilde{\Sigma}$, defined in (2), or the *pooled* sample covariance matrix $\widehat{\Sigma}$, defined in (10).

The following theorem says the two direction vectors are actually the same in non-HDLSS settings.

Theorem 6. *Let* $\mathbf{v}_{\text{MDP}}$ *and* $\mathbf{v}_{\text{FLD}}$ *be as defined in (3) and (9), respectively. If* $d < n - 1$ *and the data matrix* Z *in (1) is full rank, then* $\mathbf{v}_{\text{MDP}} = \mathbf{v}_{\text{FLD}}$.

To prove this theorem we need the following lemma, which is a slight variation of exercise 5.16 in Searle (1982). Note that the Moore-Penrose generalized inverse operation is equivalent to the ordinary matrix inverse when the matrix is nonsingular.

Lemma 7. *Let* A *and* B *be matrices with the same number of rows and let* c *be a constant. As long as the following inverse matrices make sense,*

$$(B + cAA^{\tau})^{-1}A = B^{-1}A(I + cA^{\tau}B^{-1}A)^{-1}.$$

Proof.

$$\begin{aligned}
(B + cAA^{\mathrm{T}})B^{-1}A(I + cA^{\mathrm{T}}B^{-1}A)^{-1} &= (I + cAA^{\mathrm{T}}B^{-1})A(I + cA^{\mathrm{T}}B^{-1}A)^{-1} \\
&= (A + cAA^{\mathrm{T}}B^{-1}A)(I + cA^{\mathrm{T}}B^{-1}A)^{-1} \\
&= A(I + cA^{\mathrm{T}}B^{-1}A)(I + cA^{\mathrm{T}}B^{-1}A)^{-1} \\
&= A.
\end{aligned}$$

□

Proof of Theorem 6. Note that under the assumption $d < n - 1$, both the global sample covariance matrix $\widetilde{\Sigma}$ and the pooled sample covariance matrix $\widehat{\Sigma}$ are nonsingular so that the Moore-Penrose generalized inverse matrices are actually the inverse matrices. Let p and q be the proportion of samples with target $+1$ and -1, respectively, i.e. $p = n_1/n$ and $q = n_2/n$. Then the centered version of Z is

$$\begin{aligned}
Z - \bar{Z} &= [X, Y] - (p\bar{X} + q\bar{Y}) \\
&= [X, Y] - [\bar{X}, \bar{Y}] + [q(\bar{\mathbf{x}} - \bar{\mathbf{y}})\mathbf{1}_{n_1}^{\mathrm{T}}, -p(\bar{\mathbf{x}} - \bar{\mathbf{y}})\mathbf{1}_{n_2}^{\mathrm{T}}] \\
&= [X - \bar{X}, Y - \bar{Y}] + [q(\bar{\mathbf{x}} - \bar{\mathbf{y}})\mathbf{1}_{n_1}^{\mathrm{T}}, -p(\bar{\mathbf{x}} - \bar{\mathbf{y}})\mathbf{1}_{n_2}^{\mathrm{T}}].
\end{aligned}$$

Thus,

$$\begin{aligned}(n-1)\widetilde{\Sigma} &= (Z-\bar{Z})(Z-\bar{Z})^{\mathrm{T}} \\ &= (X-\bar{X})(X-\bar{X})^{\mathrm{T}} + (Y-\bar{Y})(Y-\bar{Y})^{\mathrm{T}} \\ &\qquad + n_1 q^2(\bar{\mathbf{x}}-\bar{\mathbf{y}})(\bar{\mathbf{x}}-\bar{\mathbf{y}})^{\mathrm{T}} + n_2 p^2(\bar{\mathbf{x}}-\bar{\mathbf{y}})(\bar{\mathbf{x}}-\bar{\mathbf{y}})^{\mathrm{T}} \\ &= (n-2)\widehat{\Sigma} + (n_1 q^2 + n_2 p^2)(\bar{\mathbf{x}}-\bar{\mathbf{y}})(\bar{\mathbf{x}}-\bar{\mathbf{y}})^{\mathrm{T}},\end{aligned}$$

since the cross-product term is zero.

Therefore,

$$\widetilde{\Sigma} = \frac{n-2}{n-1}\widehat{\Sigma} + \frac{(n_1 q^2 + n_2 p^2)(\bar{\mathbf{x}}-\bar{\mathbf{y}})(\bar{\mathbf{x}}-\bar{\mathbf{y}})^{\mathrm{T}}}{n-1}.$$

Now if we apply Lemma 7 with $B = \frac{n-2}{n-1}\widehat{\Sigma}$, $A = \bar{\mathbf{x}} - \bar{\mathbf{y}}$, and $c = (n_1 q^2 + n_2 p^2)/(n-1)$, then

$$\begin{aligned}\widetilde{\Sigma}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}}) &= \left(\frac{n-2}{n-1}\widehat{\Sigma}\right)^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}})\left[I + \frac{n_1 q^2 + n_2 p^2}{n-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}})^{\mathrm{T}}\left(\frac{n-2}{n-1}\widehat{\Sigma}\right)^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}})\right]^{-1} \\ &= \frac{\widehat{\Sigma}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}})}{\frac{n-2}{n-1} + \frac{n_1 q^2 + n_2 p^2}{n-1}(\bar{\mathbf{y}}-\bar{\mathbf{y}})^{\mathrm{T}}\widehat{\Sigma}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}})} \\ &= \frac{\widehat{\Sigma}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}})}{\text{constant}}.\end{aligned}$$

Thus $\mathbf{v}_{\mathrm{MDP}} = \mathbf{v}_{\mathrm{FLD}}$ after standardization. □

It is not obvious that we get the same classifier whether we use the global sample covariance matrix or the pooled sample covariance matrix if $d < n-1$. Let us consider a simple example in $\mathbb{R}^2$ to get some geometrical understanding of this, and how it can be seen in terms of the underlying distributions.

Let X and Y be random variables from two bivariate normal distributions with different means, but with the same covariance matrix, respectively, i.e.

$$X \sim \mathcal{N}_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right), \tag{11}$$

$$Y \sim \mathcal{N}_2\left(\begin{pmatrix}-\mu_1\\ -\mu_2\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right), \tag{12}$$

where μ_1 and μ_2 are real numbers and $-1 < \rho < 1$. Note that by a shift of the data, this mean structure is quite general.

The common underlying covariance structure of X and Y, whose estimator is the pooled sample covariance matrix $\widehat{\Sigma}$, is

$$\Sigma_p := \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Let Z be the random variable whose distribution is the mixture of the two distributions (11) and (12) with equal probabilities. We can write

$$\begin{aligned} Z &= \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \\ &= B \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + U, \end{aligned}$$

where

$$B = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2, \end{cases}$$

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

and B and U are independent.

Then, Z has mean $\mathbf{0}$ and the variance of Z_i is equal to the variance of $\mu_i B + U_i$, which is $1 + \mu_i^2,\ i = 1, 2$. The covariance of Z_1 and Z_2 is equal to the covariance of $\mu_1 B + U_1$ and $\mu_2 B + U_2$, which is $\rho + \mu_1 \mu_2$. Thus, the covariance matrix of Z, whose estimator is the global sample covariance matrix $\widetilde{\Sigma}$, is

$$\Sigma_g := \begin{pmatrix} 1 + \mu_1^2 & \rho + \mu_1 \mu_2 \\ \rho + \mu_1 \mu_2 & 1 + \mu_2^2 \end{pmatrix}.$$

Note that Σ_g can be expressed as a sum of Σ_p and the outer product of the mean difference vector $(2\mu_1, 2\mu_2)^{\mathrm{T}}$ multiplied by a constant:

$$\Sigma_g = \Sigma_p + \frac{1}{4} \begin{pmatrix} 2\mu_1 \\ 2\mu_2 \end{pmatrix} \begin{pmatrix} 2\mu_1 & 2\mu_2 \end{pmatrix}.$$

The true version of the MDP direction vector can be defined as the product of Σ_g^{-1} and the mean difference vector, $(2\mu_1, 2\mu_2)^{\mathrm{T}}$:

$$\begin{aligned}
\text{True } \mathbf{v}_{\text{MDP}} &= \Sigma_g^{-1} \times \begin{pmatrix} 2\mu_1 \\ 2\mu_2 \end{pmatrix} \\
&\propto \begin{pmatrix} 1+\mu_2^2 & -\rho-\mu_1\mu_2 \\ -\rho-\mu_1\mu_2 & 1+\mu_1^2 \end{pmatrix} \times \begin{pmatrix} 2\mu_1 \\ 2\mu_2 \end{pmatrix} \qquad (13) \\
&\propto \begin{pmatrix} \mu_1-\rho\mu_2 \\ \mu_2-\rho\mu_1 \end{pmatrix}. \qquad (14)
\end{aligned}$$

Similarly the true version of the FLD direction vector is the product of Σ_p^{-1} and $(2\mu_1, 2\mu_2)^{\mathrm{T}}$:

$$\begin{aligned}
\text{True } \mathbf{v}_{\text{FLD}} &= \Sigma_p^{-1} \times \begin{pmatrix} 2\mu_1 \\ 2\mu_2 \end{pmatrix} \\
&\propto \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \times \begin{pmatrix} 2\mu_1 \\ 2\mu_2 \end{pmatrix} \\
&\propto \begin{pmatrix} \mu_1-\rho\mu_2 \\ \mu_2-\rho\mu_1 \end{pmatrix}. \qquad (15)
\end{aligned}$$

From (14) and (15), we can see the resulting direction vectors are actually the same in both cases.

One possible interpretation is that the effect of the mean difference vector on the global covariance matrix Σ_g is negated when we take the inverse (13), which cancels out when we multiply the mean difference vector by Σ_g^{-1} to obtain the direction vector. Note that this is only true for the non-HDLSS case.

4 Simulation Study and Open Problems

For a comparison of the MDP direction vector with other linear methods in terms of classification performance, a simulation study in a very simple setting is done here. Each class has a d-dimensional multivariate Gaussian distribution with unit variance. The two classes only differ in their means: Class $+1$ has mean $\mu\mathbf{1}_d$ and Class -1 has mean $-\mu\mathbf{1}_d$. We generate $n = 100$ training data vectors from this distribution, 50 for each class, from Euclidean spaces of dimensions $d = 1, 3, 10, 30, 100$, and 300. We also took μ to have values $.1, .2, .5$, and 1, and for each combination of (d, μ), $1,000$ repetitions are done.

The test sets of size 200 are used to evaluate the misclassification error rates. The linear classification methods considered were the MD, FLD, MDP, Naive Bayes, as well as the theoretically optimal Bayes rule.

Figure 6 shows the error rates with error bars for each method with different d and μ. Because of the large number of repetitions, the error bars here are very small. The top four panels show the error rates in the original scale and the bottom ones show them in the $\log_{10}$ scale which allows a closer look at the small error rates. Note that MD and Naive Bayes show nearly identical error rates in all cases. Also note that since a larger d results in a larger distance between the two classes in this particular setting, all the methods show generally better performances for higher dimensions. This also can be explained by the asymptotics in Hall et al. (2002), who showed that when $d \gg n$, under a mild assumption, the pairwise distances between each pair of data points are approximately identical so that the data points form an n-simplex. Thus in discrimination, we have two simplices for each class so that every reasonable classification method finds the same direction in the end, which leads to data piling for all methods.

The FLD and MDP have exactly the same error rates up to $d = 30$, as expected from Theorem 6. The most interesting feature of Figure 6 is that, however, when d reaches 100, which is equal to the sample size, the error rates of both FLD and MDP jump up significantly. Here the error rate of the MDP is worse than that of FLD, with the larger differences for the larger values of μ. Afterwards, their error rates plummet down when $d = 300$ where MDP is better than FLD. This can be explained by the following: As d gets close to n, the estimation of the covariance structure becomes unreliable due to the lack of the data points, which yields increasing error rates. The effect of this unreliable covariance estimation problem peaks at $d = n$ and remains until d is somewhat higher than n. Meanwhile, as d increases past n, the asymptotics in Hall et al. (2002) begins to take effect as discussed earlier, resulting in decreasing error rates.

For a zoomed-in view of this anomaly, we repeated the same simulation with a finer range of d, $d = 90, \cdots, 110$ and the result is shown in Figure 7. The error rates of the FLD and MDP increase as d grows close to the sample size, having the same error rates up through $d = 99$. They differentiate when $d = 100$ and subsequently both error rates decrease as d increases. The FLD performs better than MDP for dimensions slightly larger than the sample size, however, the gap between them diminishes as d increases and eventually MDP performs much better than FLD as shown in Figure 6.

Explanation of this behavior is an open problem. For more theoretically rigorous explanation, the (d, n)-asymptotics developed by Johnstone (2001) and Fujikoshi et al.

(2003) should be useful.

Acknowledgements

The authors are grateful to C. M. Perou for providing the breast cancer data. This research was supported by MIDAG (Medical Image Display Analysis Group, department of computer science, UNC-CH), and by NSF grant (DMS-0308331).

References

Bickel, P. and Levina, E. (2003). Some theory for Fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations. To appear in *Bernoulli*.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Duda, R. D., Hart, P. E. and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience.

Fujikoshi, Y., Himeno, T. and Wakaki, H. (2003). Asymptotic results in canonical discriminant analysis when the sample size and dimension is large compared to the sample size. Tech. Rep. 03-17, Statistical Research Group, Hiroshima University.

Hall, P., Marron, J. S. and Neeman, A. (2002). Geometric representation of high dimension low sample size data. To appear in *Journal of Royal Statiatical Society*.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: data mining, inference, and prediction*. Springer.

Johnson, B. A., Lin, D., Marron, J. S., Ahn, J., Parker, J. and Perou, C. M. (2004). Distance weighted discrimination with censored outcomes. In preparation.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29** 295–327.

Marron, J. S., Todd, M. and Ahn, J. (2004). Distance weighted discrimination. Under revision.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O. and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley-Sons Inc.

Tukey, J. and Tukey, P. (1990). *Strips Displaying Empirical Distributions: I. Textured Dot Strips*. Bellcore.

Zhang, H., Marron, J. S. and Todd, M. J. (2004). Weighted distance weighted discrimination. In preparation.

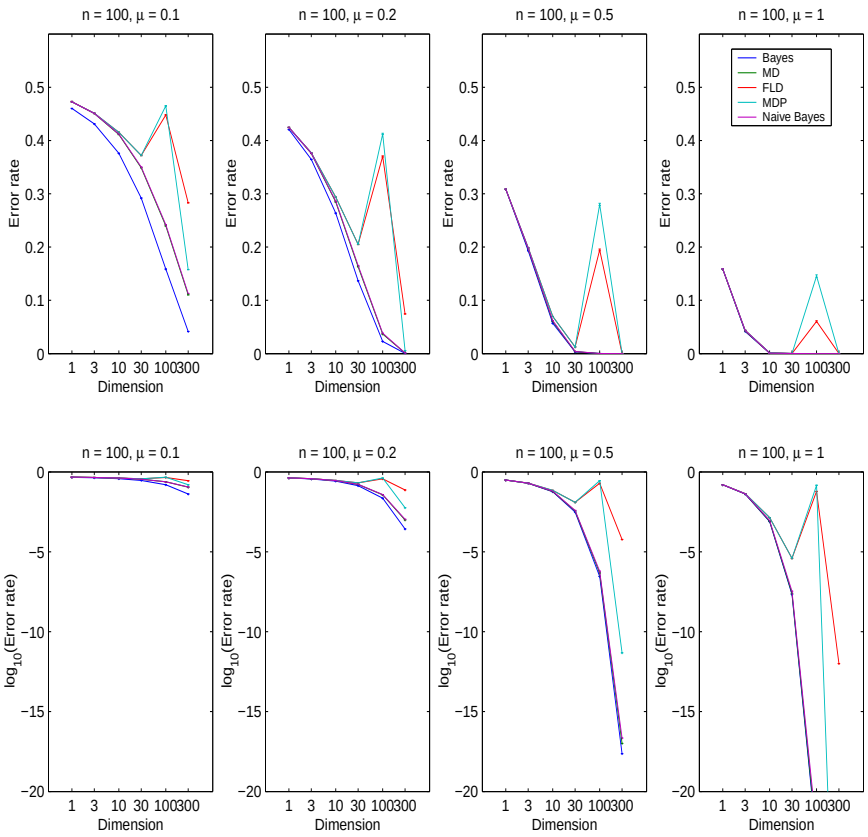

Figure 6: Misclassification error rate for the Bayes, MD, FLD, MDP, and Naive Bayes method from simulation, $n = 100$, $d = (1, 3, 10, 30, 100, 300)$, and $\mu = (0.1, 0.2, 0.5, 1)$, shown in the original scale in the top panels, $\log_{10}$ scale in the bottom panels.

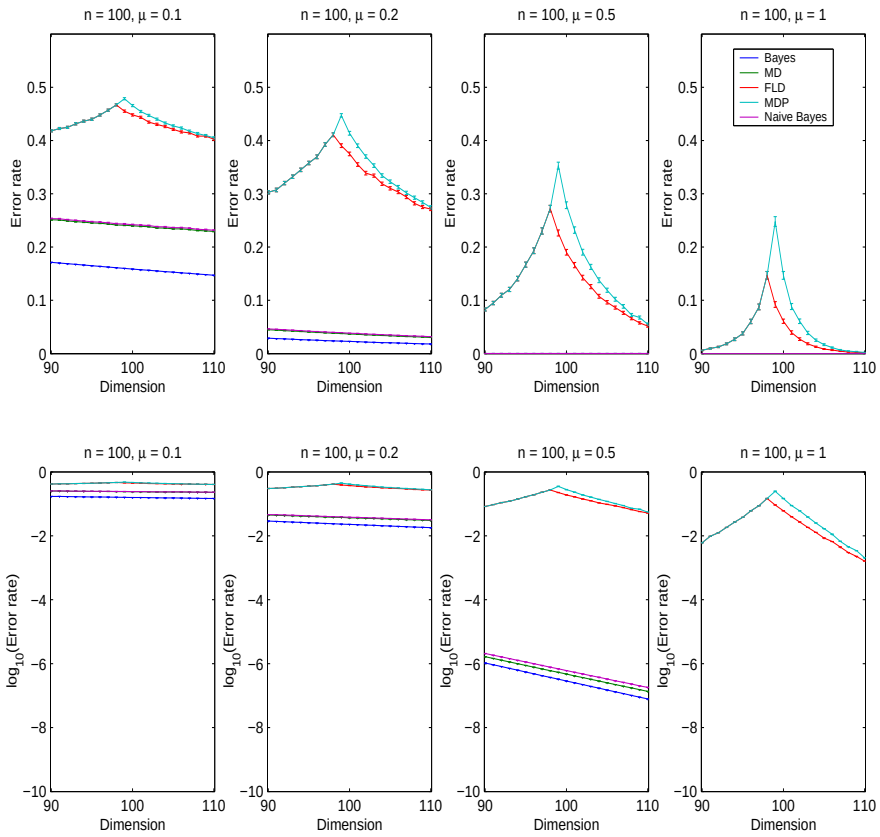

Figure 7: Misclassification error rate for the Bayes, MD, FLD, MDP, and Naive Bayes method from simulation, $n = 100$, $d = (90, 91, \cdots, 110)$, and $\mu = (0.1, 0.2, 0.5, 1)$, shown in the original scale in the top panels, $\log_{10}$ scale in the bottom panels.