Published in final edited form as:
Technometrics. 2015 September 1; 57(4): 524–534. doi:10.1080/00401706.2014.965347.

MATRIX DISCRIMINANT ANALYSIS WITH APPLICATION TO COLORIMETRIC SENSOR ARRAY DATA

Wenxuan Zhong[1] and **Kenneth S. Suslick**[2]

Wenxuan Zhong: wenxuan@uga.edu; Kenneth S. Suslick: ksuslick@illinois.edu

[1]UNIVERSITY OF GEORGIA, ATHENS, GA 30602

[2]UNIVERSITY OF ILLINOIS, CHAMPAIGN, IL 61820

Abstract

With the rapid development of nano-technology, a "colorimetric sensor array" (CSA) which is referred to as an optical electronic nose has been developed for the identification of toxicants. Unlike traditional sensors which rely on a single chemical interaction, CSA can measure multiple chemical interactions by using chemo-responsive dyes. The color changes of the chemo-responsive dyes are recorded before and after exposure to toxicants and serve as a template for classification. The color changes are digitalized in the form of a matrix with rows representing dye effects and columns representing the spectrum of colors. Thus, matrix-classification methods are highly desirable. In this article, we develop a novel classification method, matrix discriminant analysis (MDA), which is a generalization of linear discriminant analysis (LDA) for the data in matrix form. By incorporating the intrinsic matrix-structure of the data in discriminant analysis, the proposed method can improve CSA's sensitivity and more importantly, specificity. A penalized MDA method, PMDA, is also introduced to further incorporate sparsity structure in discriminant function. Numerical studies suggest that the proposed MDA and PMDA methods outperform LDA and other competing discriminant methods for matrix predictors. The asymptotic consistency of MDA is also established. R code and data are available online as supplementary material.

Keywords

Classification; Matrix predictors; Feature selection; Linear discriminant analysis; Regularization; Sensor array

1. Introduction

The development and refinement of sensors for a rapid identification of volatile chemical toxicants (VCTs) is very important. Integrated into a security system, a sensor can be used to automatically trigger an instantaneous response, such as shutting down and isolating

Correspondence to: Wenxuan Zhong, wenxuan@uga.edu.

Supplementary Materials
R code and datasets: The R codes and data files can be downloaded as a .zip file.

ventilation systems when there is an accidental release of VCTs. A powerful sensor is crucial to curtail the spread of chemical spills and to limit the areas of contamination.

The traditional sensor systems that have been widely used in detecting VCTs are vapor sensors. Vapor sensors rely either on absorption into a set of polymers or on oxidations at heated metal oxides. While such systems generally allow for discriminating VCTs in different chemical classes, the discrimination of similar VCTs within one chemical class remains a challenging goal. To surmount the challenge, a low cost yet highly sensitive sensor called "colorimetric sensor array" (CSA) has been developed (Suslick et al. 2007, Lim et al. 2008 and Feng et al. 2010). Analogous to the mammalian olfaction system, which recognizes smells by the composite electronic signals generated by different epithelium olfactory cells in response to the smells, CSA uses large amount of chemical dyes to turn a smell into optical composite signals. Thus, CSA sensor is also referred to as "optical electronic nose". As shown in Figure 1A, CSA is simply a digitally-imaged, two-dimensional extension of litmus paper (Rakow and Suslick 2000; Rakow et al. 2005; Zhang and Suslick 2005). Thirty six chemo-responsive dyes are randomly assigned to 36 spots scattered as a 6 × 6 array on a chip. The 36 dyes can measure multiple chemical interactions, e.g., ligand-metal coordination, Lewis acid-base interactions, and strong dipolar interactions. For any odorant, a response is generated by digital subtraction, pixel by pixel, of the color of 36 pre-print chemo-responsive dyes before and after exposure: red value after exposure (R_{after}) minus red value before (R_{before}), green minus green, blue minus blue. Averaging the centers of the spots (~ 300 pixels) for each dye, the result is simply a 36×3 matrix, where each row represents the color change of a dye and each column represents one of the three spectrum coordinates (Red, Green, Blue) of a color cube. As shown in Figure 1B, the matrix is a color fingerprint of a VCT and can be used to classify VCTs. By measuring a much broader range of chemical interactions, CSA provides dramatic improvements over traditional sensor systems in both sensitivity and, even more importantly, specificity in VCTs detection.

The simplest and most popular classification approach is linear discriminant analysis (LDA). Classical LDA, introduced in Fisher (1936), can be formulated in the following way. Given a training sample, Fisher's LDA aims to find linear combinations of all predictors that maximize the ratio of between-class variance to within-class variance; see Anderson (2003) for a review of Fisher's LDA. It has been shown that the performance of Fisher's LDA and its variants is comparable to that of many advanced classification methods in a variety of settings; see chapter 4 of Hastie, Tibshirani, and Friedman (2009) for a review. Though LDA has succeeded in many real applications, it is not directly applicable to analyze our CSA data due to the following two challenges. First, applying LDA to CSA data requires stacking a 36 × 3 matrix to form a 108 dimensional vector, which renders Fisher's LDA inapplicable for small (say ≤ 100) sample application. Moreover, the rows and columns of CSA output have different interpretations and should be treated differently in classification analysis. Second, for some VCTs, only a small number of dyes are discriminant relevant dyes. See Figure 1C. Thus, using all dyes in classification models, like LDA, may bring in noise and result in high misclassification errors.

To overcome the first challenge, we propose a matrix discriminant analysis method (MDA), in which we project the 36 × 3 matrix into row (dye) space and column (color) space separately. The two projections are estimated iteratively and integrated together for classification. Thus, rows and columns of the data are treated differently. By retaining the matrix structure of the data, MDA provides natural interpretations of the discriminant directions and alleviates the curse of dimensionality. To surmount the second challenge and improve MDA's classification specificity, we impose a sparsity structure on dyes by developing a penalized MDA (PMDA). PMDA reduces not only the number of parameters in the discriminant analysis but also misclassification errors in many applications. More importantly, it provides further insight on which dyes are discriminant relevant for certain VCTs, and can serve as a guidance for designing the next-generation CSA.

It is worth noting that 2D classification methods have been developed in the image processing community. The primary usage of these methods is the classification of 2D images where the 2D refers to the two pixel coordinates. The 2D-LDA that was proposed in Li and Yuan (2005) is one of the popular works in the 2D classification literature. To use the matrix structure, 2D-LDA seeks *d* linear discriminant directions that can maximize the trace of the between-group variation over the trace of the within-group variance. Though 2D-LDA incorporates the matrix structure into the estimation of the discriminant directions, as pointed out by Zheng et al. (2008), 2D-LDA ignores the between-row correlations in the matrix observations. Ignoring between-row correlations leads to substantially higher misclassification errors than Fisher's LDA, when the rows are correlated (Zheng et al. 2008). In our CSA data, some of the chemical responsive dyes have similar chemical structures, such as the PH indicators that respond to Brønsted acidity/basicity. Thus, the rows of the CSA observations are correlated with each other which renders the 2D-LDA approach inapplicable. Moreover, similar to many other 2D-discriminant methods developed in the image processing community, 2D-LDA assumes that results produced by using the original images and the rotated images are the same. This assumption is referred to as the the rotation-invariant property (section 3.7.2 in Gonzales and Woods 2002). This rotation-invariant property does not hold for our CSA data, in which rows represent the chemical responsive dyes and columns represent the spectral components on the color space. Thus, we need a 2D classification method that can treat the rows and columns differently.

Our MDA method is also related to the dimension folding sliced inverse regression method (Li et al. 2009) that is proposed for effectively reducing the dimensionality of the matrix predictors in regression. However, different from the dimension folding method, our primary goal is to reduce the misclassification error when the sample size is small. Moreover, effective methods for imposing sparsity on the estimates in the dimensional folding are still lacking.

The rest of the article is organized as follows. In Section 2, we briefly review the LDA method. We develop the matrix discriminant analysis method (MDA) and present its asymptotic properties in Section 3. In Section 4, we develop the penalized matrix discriminant analysis method (PMDA). Simulations and real data analysis are presented in Sections 5 and 6. A few remarks in Section 7 conclude the article.

2. Fisher’s Linear Discriminant Analysis for *p*-dimensional Vectors

In this section, we briefly review Fisher’s LDA method to motivate our MDA and PMDA methods. To make Fisher’s LDA applicable to our CSA data, we can stack the columns of the random matrix **X** to create a *p*-dimensional random vector **x**. Let $Y \in \{1, \ldots, K\}$ be the class label. We assume that $\mathbf{x}|Y = k$ has a normal distribution with equal covariance matrix for all *k*. Fisher’s LDA seeks a $d \le K - 1$ dimensional projection of **x** with the largest between-group variation relative to within-group variation. Given *d*, the linear discriminant directions, $\beta_1, \ldots, \beta_d$, in Fisher’s LDA can be obtained by progressively maximizing

$$L(\boldsymbol{\eta})=\frac{\boldsymbol{\eta}'\mathrm{Var}(E[\mathbf{x}|Y])\boldsymbol{\eta}}{\boldsymbol{\eta}'\mathrm{Var}[\mathbf{x}]\boldsymbol{\eta}}, \quad (1)$$

with respect to η under the constraints that β_i and β_j are orthogonal with respect to $Var[\mathbf{x}]$; see Fisher (1936). In general, *d* is unknown unless you have extra structure information of the data. A χ^2 test that is described in Section 12.5 of Mardia et al. (1979) can be used to estimate *d*.

Observing $(\mathbf{x}_i, Y_i)$, $i = 1, \ldots, n$, we can estimate $E[\mathbf{x}]$ by its sample version $\bar{\mathbf{x}}$, where $\overline{\mathbf{x}}=\frac{1}{n}\sum_i \mathbf{x}_i$, and estimate $E[\mathbf{x}|Y]$ by its sample version $\hat{E}[\mathbf{x}|Y]$, where $\hat{E}[\mathbf{x}|Y]=\frac{1}{n_k}\sum_{\{i:Y_i=k\}}\mathbf{x}_i$ with n_k being the number of observations in group *k*. Similarly, $\mathrm{Var}[\mathbf{x}]$ can be estimated by $\widehat{\mathrm{Var}}[\mathbf{x}]=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})'$ and $\mathrm{Var}[E(\mathbf{x}|Y)]$ can be estimated by $\widehat{\mathrm{Var}}[E(\mathbf{x}|Y)]=\sum_{k=1}^{K}\frac{n_k}{n}(\hat{E}[\mathbf{x}|Y=k]-\overline{\mathbf{x}})(\hat{E}[\mathbf{x}|Y=k]-\overline{\mathbf{x}})'$. Replacing $\mathrm{Var}[E(\mathbf{x}|Y)]$ and $\mathrm{Var}[\mathbf{x}]$ in (1) by $\widehat{\mathrm{Var}}[E(\mathbf{x}|Y)]$ and $\widehat{\mathrm{Var}}[\mathbf{x}]$, we can easily obtain the *d* estimated Fisher’s LDA directions, $\hat{\beta}_1, \ldots, \hat{\beta}_d$, by solving the following linear system:

$$\begin{array}{c}\widehat{\mathrm{Var}}[E(\mathbf{x}|Y)]\boldsymbol{\eta}_i=\lambda_i\widehat{\mathrm{Var}}[\mathbf{x}]\boldsymbol{\eta}_i, \quad i=1,\ldots,d \text{ and } \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d>0,\\ \boldsymbol{\eta}_i'\widehat{\mathrm{Var}}[\mathbf{x}]\boldsymbol{\eta}_j=I_{\{i=j\}},\end{array} \quad (2)$$

with respect to η_i, where $I_{\{\}}$ is the indicator function taking value 1 if $i = j$ and value 0 if $i \ne j$.

With small sample size, stacking the matrix observation of our CSA data may lead to a severe curse of dimensionality, which refers to various difficulties a large number of variables (or dimensions) can cause to function approximation, model fitting, information extraction as well as to computation (Fan and Li 2006). For example, with a typical 36-dye CSA data that can be modeled by 36 parameters for dyes effect and 3 parameters for Red, Green and Blue color spectrum effect, simple vectorization generates 36 × 3 = 108 parameters for a classical vector-based LDA. Thus the simple vectorization renders many vector-based approaches infeasible for a sample with less than 108 observations. Moreover, even if we have a fairly large sample, both the computational efficiency and the estimation accuracy of the classical vector-based LDA will be compromised by simple vectorization (Donoho and Elad 2003; Fan and Li 2006). To alleviate the curse of dimensionality, penalization approaches have been proposed. The regularized discriminant analysis (RDA) method proposed by Friedman (1989) is one of the early proposals. Instead of directly using the sample within-group covariance matrix, a ridge penalty is employed in RDA to stabilize the estimate. Following the same trend, Clemmensen et al. (2011) developed sparse discriminant analysis (SDA). In SDA, an L_1 penalty is employed to obtain a sparse estimate of the discriminant directions. Though penalization can alleviate the curse of dimensionality to some degree, it may also generate some bias and computational complexity in estimating the discriminant directions. A method that can eliminate the bias and achieve sparsity simultaneously, such as our MDA method, is more attractive.

3. Matrix Discriminant Analysis for *r* × *q* dimensional Matrices

To retain the matrix structure of **X** in the CSA data, we shall develop a matrix discriminant analysis method. To discriminate the objects in the form of matrices, we aim to find d orthogonal low dimensional representations of **X**, $\boldsymbol{\beta}_j^{'}\mathbf{X}\boldsymbol{\xi}$, $j = 1, \ldots, d$, that exhibit the maximum ratio of between-class variance to within-class variance. Here, we require the d low dimensional representations to be orthogonal to make the representations be identifiable and easy to interpret. In the CSA application, each entry in β_j specifies the discrimination power of a chemo-responsive dye and each entry in ξ specifies the effect of spectral components coded by the RGB triplets. To discriminate among similar compounds within one chemical class, the dyes used in CSA are the nanoporous pigments, which measure the subtle difference between VCTs in one chemical class. As similar compounds display color differences along one spectral direction on the nanoporous pigments (Feng et al. 2010), we assume ξ resides in a one-dimensional space. As illustrated in the case study that is presented in Section 6, the assumption is valid in most CSA applications. When the effect of spectral components cannot be summarized by a single ξ, we can generalize our method to accommodate multiple ξs. The generalization will be discussed in Section 3.3. Without loss of generality, we assume that **X** is an $r \times q$ matrix. Correspondingly, the β_js, $j = 1, \ldots, d$ are r dimensional vectors and the ξ is a q dimensional vector.

3.1. Matrix discriminant analysis method

Motivated by Fisher's LDA, we aim to find d orthogonal vectors $\beta_1, \ldots, \beta_d$ and ξ that maximize the ratio of between-class variance to within-class variance along the $\boldsymbol{\beta}_j^{'}\mathbf{X}\boldsymbol{\xi}$, $j = 1, \ldots, d$ directions. Here, we assume that the within-group covariance matrices are the same. Thus, the d orthogonal vectors $\beta_1, \ldots, \beta_d$ and the vector ξ can be obtained by maximizing

$$L(\boldsymbol{\eta}, \boldsymbol{\theta}) = \frac{\mathrm{Var}(E[\boldsymbol{\eta}^{'}\mathbf{X}\boldsymbol{\theta}|Y])}{\mathrm{Var}[\boldsymbol{\eta}^{'}\mathbf{X}\boldsymbol{\theta}]} \quad (3)$$

with respect to η and θ. Because (3) is a bivariate quadratic function, we can find $\beta_1, \ldots, \beta_d$ and ξ by iteratively maximizing $L(\eta, \theta)$.

Observing $(\mathbf{X}_i, Y_i)$, $1 \le i \le n$, for any given δ, we use $\widehat{\mathrm{Var}}(E[\mathbf{X}\boldsymbol{\delta}|Y])$ to denote the sample version of $\mathrm{Var}(E[\mathbf{X}\delta|Y])$, where

$\widehat{\mathrm{Var}}(E[\mathbf{X}\boldsymbol{\delta}|Y])=\sum_{k=1}^{K} n_k/n(\hat{E}[\mathbf{X}\boldsymbol{\delta}|Y]-\overline{\mathbf{X}}\boldsymbol{\delta})(\hat{E}[\mathbf{X}\boldsymbol{\delta}|Y]-\overline{\mathbf{X}}\boldsymbol{\delta})^{'}$, and $\widehat{\mathbf{Var}}[\mathbf{X}\boldsymbol{\delta}]$ to denote the sample version of Var[$\mathbf{X}\delta$], where $\widehat{\mathrm{Var}}[\mathbf{X}\boldsymbol{\delta}]=\sum_{i=1}^{n} 1/n(\mathbf{X}_i\boldsymbol{\delta}-\overline{\mathbf{X}}\boldsymbol{\delta})(\mathbf{X}_i\boldsymbol{\delta}-\overline{\mathbf{X}}\boldsymbol{\delta})^{'}$.

Given an initial estimate of ξ, denoted by $\hat{\xi}$, analogous to Fisher's LDA, we can obtain an estimate of $\beta_1, \ldots, \beta_d$, say $\hat{\beta}_1, \ldots, \hat{\beta}_d$, by solving the linear system,

$$\widehat{\mathrm{Var}}(E[\mathbf{X}\hat{\boldsymbol{\xi}}|Y])\boldsymbol{\eta}_i=\lambda_i\widehat{\mathrm{Var}}(\mathbf{X}\hat{\boldsymbol{\xi}})\boldsymbol{\eta}_i, \quad i=1,\ldots,d \text{ and } \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d>0\,\boldsymbol{\eta}_i^{'}\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}]\boldsymbol{\eta}_j=I_{\{i=j\}}, \quad (4)$$

with respect to η_i.

Meanwhile, fixing β at $\hat{\beta}_1$, we can obtain an estimate of ξ, say $\hat{\xi}$, by maximizing,

$$\frac{\boldsymbol{\theta}^{'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{'}\mathbf{X}|Y])\boldsymbol{\theta}}{\boldsymbol{\theta}^{'}\widehat{\mathrm{Var}}[\hat{\boldsymbol{\beta}}_1^{'}\mathbf{X}]\boldsymbol{\theta}}, \quad (5)$$

with respect to θ.

Since the $L(\eta, \theta)$ is bounded from above and nondecreasing at each iteration (detailed proof is given below), we employ the following iterative algorithm for estimating $\beta_1, \ldots, \beta_d$ and ξ. We, first, give an initial estimate of ξ. For the fixed ξ, we can estimate the $\beta_1, \ldots, \beta_d$ by solving (4). Then fixing β at $\hat{\beta}_1$, we can estimate the ξ by maximizing (5). The two steps iterate until convergence.

Algorithm 1 (MDA algorithm)

- *Initialize* $\hat{\xi}^{(0)}$ *such that* $\hat{\xi}^{(0)\prime}\hat{\xi}^{(0)} = 1$.
- *Fixing* $\hat{\xi}$ *at* $\hat{\xi}^{(0)}$, *solve the linear system* (4) *to find* $\hat{\boldsymbol{\beta}}_1^{(0)},\ldots,\hat{\boldsymbol{\beta}}_d^{(0)}$.
- *Iterate until* $L(\hat{\boldsymbol{\beta}}_1^{(1)},\hat{\boldsymbol{\xi}}^{(1)})$ *converges.*
    - *Fixing* $\hat{\beta}$ *a* $\hat{\boldsymbol{\beta}}_1^{(0)}$, *maximize* (5) *to find* $\hat{\xi}^{(1)}$.
    - *Fixing* $\hat{\xi}$ *at* $\hat{\xi}^{(1)}$, *solve the linear system* (4) *to find* $\hat{\boldsymbol{\beta}}_1^{(1)},\ldots,\hat{\boldsymbol{\beta}}_d^{(1)}$.
    - *Update* $\hat{\xi}^{(0)}$ by $\hat{\xi}^{(1)}$ *and* $\hat{\boldsymbol{\beta}}_1^{(0)}$ *by* $\hat{\boldsymbol{\beta}}_1^{(1)}$ *and calculate* $L(\hat{\boldsymbol{\beta}}_1^{(1)},\hat{\boldsymbol{\xi}}^{(1)})$.
- *Output final estimates* $\hat{\boldsymbol{\beta}}_j^{(1)}$ *for j* = 1, …, *d and* $\hat{\xi}^{(1)}$.

In the following, we show that Algorithm 1 converges. Given $\hat{\xi}^{(0)}$, since $\hat{\boldsymbol{\beta}}_1^{(0)}$ is the solution of (4), it follows that for any η,

$$\frac{\boldsymbol{\eta}'\widehat{\mathrm{Var}}(E[\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}|Y])\boldsymbol{\eta}}{\boldsymbol{\eta}'\widehat{\mathrm{Var}}(\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)})\boldsymbol{\eta}} \le \frac{\hat{\boldsymbol{\beta}}_1^{(0)'}\widehat{\mathrm{Var}}(E[\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}|Y])\hat{\boldsymbol{\beta}}_1^{(0)}}{\hat{\boldsymbol{\beta}}_1^{(0)'}\widehat{\mathrm{Var}}(\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)})\hat{\boldsymbol{\beta}}_1^{(0)}} = \frac{\hat{\boldsymbol{\xi}}^{(0)'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\hat{\boldsymbol{\xi}}^{(0)}}{\hat{\boldsymbol{\xi}}^{(0)'}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\hat{\boldsymbol{\xi}}^{(0)}}. \qquad (6)$$

On the other hand, given $\hat{\boldsymbol{\beta}}_1^{(0)}$, since $\xi^{(\hat{1})}$ is the solution of (5), it follows that for any θ,

$$\frac{\boldsymbol{\theta}'\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\boldsymbol{\theta}}{\boldsymbol{\theta}'\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\boldsymbol{\theta}} \le \frac{\boldsymbol{\xi}^{(1)'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\hat{\boldsymbol{\xi}}^{(1)}}{\hat{\boldsymbol{\xi}}^{(1)'}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\hat{\boldsymbol{\xi}}^{(1)}}.$$

Thus, we have

$$\frac{\hat{\boldsymbol{\xi}}^{(0)'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\hat{\boldsymbol{\xi}}^{(0)}}{\hat{\boldsymbol{\xi}}^{(0)'}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\hat{\boldsymbol{\xi}}^{(0)}} \le \frac{\hat{\boldsymbol{\xi}}^{(1)'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\hat{\boldsymbol{\xi}}^{(1)}}{\hat{\boldsymbol{\xi}}^{(1)'}p\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\hat{\boldsymbol{\xi}}^{(1)}}. \qquad (7)$$

Now given $\xi^{(\hat{1})}$, similar to the derivation of (6), we have

$$\frac{\hat{\boldsymbol{\xi}}^{(1)'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\hat{\boldsymbol{\xi}}_1^{(1)}}{\hat{\boldsymbol{\xi}}^{(1)'}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\hat{\boldsymbol{\xi}}^{(1)}} \le \frac{\hat{\boldsymbol{\beta}}_1^{(1)'}\widehat{\mathrm{Var}}(E[\mathbf{X}\hat{\boldsymbol{\xi}}^{(1)}|Y])\hat{\boldsymbol{\beta}}_1^{(1)}}{\hat{\boldsymbol{\beta}}_1^{(1)'}\widehat{\mathrm{Var}}(\mathbf{X}\hat{\boldsymbol{\xi}}^{(1)})\hat{\boldsymbol{\beta}}_1^{(1)}}, \qquad (8)$$

since $\hat{\boldsymbol{\beta}}_1^{(1)}$ is the solution of (4) at $\hat{\boldsymbol{\xi}}=\hat{\boldsymbol{\xi}}_1^{(1)}$. Combining (7) and (8), we have

$$\frac{\hat{\boldsymbol{\xi}}^{(0)'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\hat{\boldsymbol{\xi}}^{(0)}}{\hat{\boldsymbol{\xi}}^{(0)'}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\hat{\boldsymbol{\xi}}^{(0)}} \le \frac{\hat{\boldsymbol{\xi}}_1^{(1)'}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X}|Y])\hat{\boldsymbol{\xi}}^{(1)}}{\hat{\boldsymbol{\xi}}^{(1)'}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1^{(0)'}\mathbf{X})\hat{\boldsymbol{\xi}}^{(1)}} \le \frac{\hat{\boldsymbol{\beta}}_1^{(1)'}\widehat{\mathrm{Var}}(E[\mathbf{X}\hat{\boldsymbol{\xi}}^{(1)}|Y])\hat{\boldsymbol{\beta}}_1^{(1)}}{\hat{\boldsymbol{\beta}}_1^{(1)'}\widehat{\mathrm{Var}}(\mathbf{X}\hat{\boldsymbol{\xi}}^{(1)})\hat{\boldsymbol{\beta}}_1^{(1)}}. \qquad (9)$$

Meanwhile, $L(\eta, \theta)$ is bounded above since Var[$\eta'\mathbf{X}\theta$] = Var[$E[\eta'\mathbf{X}\theta|Y]$] + E[Var[$\eta'\mathbf{X}\theta|Y$]]. Ensuring that the $L(\hat{\boldsymbol{\beta}}_1^{(1)}, \hat{\boldsymbol{\xi}}^{(1)})$ is increasing in each iteration, convergence of $\xi^{(\hat{1})}$ is guaranteed. Then, convergence of $\hat{\boldsymbol{\beta}}_i^{(1)}$ is also guaranteed.

Although Algorithm 1 is a powerful tool for high dimensional classification, like other iterative algorithms, the iteration may reach a local rather than the global maximum of target function $L(\eta, \theta)$ in practice. To avoid being stuck at a local optimum, we uniformly sample multiple $\xi^{(\hat{0})}$s on $\mathbb{R}$ as the initial value and choose the discriminant rule that gives the smallest misclassification errors.

3.2. Theoretical properties

We now show that MDA yields consistent estimators of $\beta_1, \ldots, \beta_d$ if $\xi^{(\hat{0})}$ is reasonably close to the true ξ.

Theorem 1—Given $Y_1, \ldots, Y_n$, we assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{r\times q}$ are independent and identically distributed random matrices with each entry having finite mean, variance and fourth moment. Let $\beta_1, \ldots, \beta_d$ and ξ be the maximizer of (3) and $\hat{\boldsymbol{\beta}}_j^{(1)}$ be the output of Algorithm 1. If $\xi^{(\hat{0})}$ in Algorithm 1 is a $\sqrt{n}$ consistent estimator of ξ, we have that

$$\hat{\boldsymbol{\beta}}_j^{(1)} \rightarrow \boldsymbol{\beta}_j$$

in probability as $n_k \rightarrow \infty$ for $k = 1, \ldots, K$.

The proof of Theorem 1 is sketched as follows. Since $\xi^{(\hat{0})}$ is a $\sqrt{n}$ consistent estimator of ξ, we have that $\mathrm{Var}[E(\mathbf{X}\xi^{(\hat{0})}|Y)] = \mathrm{Var}[E(\mathbf{X}\xi|Y)] + O(1/n)$ and $\mathrm{Var}[\mathbf{X}\xi^{(\hat{0})}] = \mathrm{Var}[\mathbf{X}\xi] + O(1/n)$. Meanwhile, using the Tchebycheff inequality, we can show that $\widehat{\mathrm{Var}}[E(\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}|Y)]=\mathrm{Var}[E(\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}|Y)]+O(1/n)$ and $\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}]=\mathrm{Var}[\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}]+O(1/n)$. Thus, we have that $\widehat{\mathrm{Var}}[E(\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}|Y)]=\mathrm{Var}[E(\mathbf{X}\boldsymbol{\xi}|Y)]+O(1/n)$ and $\widehat{\mathrm{Var}}^{-1}[\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}]=\mathrm{Var}^{-1}[\mathbf{X}\boldsymbol{\xi}]+O(1/n)$. These facts imply that

$$\frac{\boldsymbol{\eta}'\widehat{\mathrm{Var}}[E(\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}|Y)]\boldsymbol{\eta}}{\boldsymbol{\eta}'\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}^{(0)}]\boldsymbol{\eta}'} \rightarrow \frac{\boldsymbol{\eta}'\mathrm{Var}[E(\mathbf{X}\boldsymbol{\xi}|Y)]\boldsymbol{\eta}}{\boldsymbol{\eta}'\mathrm{Var}[\mathbf{X}\boldsymbol{\xi}]\boldsymbol{\eta}'}.$$

By the perturbation theory of matrix eigenvectors, e.g., Stewart and Sun (1990), the conclusion follows immediately.

Similarly, we can show that $\xi^{(\hat{1})}$ is a consistent estimator of ξ, if $\hat{\boldsymbol{\beta}}_1^{(0)}$ is a consistent estimator of β_1.

Theorem 2—Given $Y_1, \ldots, Y_n$, we assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{r\times q}$ are independent and identically distributed random matrices with each entry having finite mean, variance and fourth moment. Let $\beta_1, \ldots, \beta_d$ and ξ be the maximizer of (3) and $\xi^{(\hat{1})}$ be the output of Algorithm 1. If $\hat{\boldsymbol{\beta}}_1^{(0)}$ in Algorithm 1 is a $\sqrt{n}$ consistent estimator of β_1, we have that

$$\hat{\boldsymbol{\xi}}^{(1)} \rightarrow \boldsymbol{\xi}$$

in probability as $n_k \rightarrow \infty$ for $k = 1, \ldots, K$.

The conclusion of Theorem 2 follows immediately from Theorem 1. Theorem 1 and 2 imply that $\xi^{(\hat{1})}$ and $\hat{\boldsymbol{\beta}}_1^{(1)}, \ldots, \hat{\boldsymbol{\beta}}_d^{(1)}$ are consistent estimators when $\xi^{(\hat{0})}$ and $\hat{\boldsymbol{\beta}}_1^{(0)}, \ldots, \hat{\boldsymbol{\beta}}_d^{(0)}$ iterates to a small neighborhood of the true parameters. It is worth noting that Theorem 2 still holds if we replace $\hat{\boldsymbol{\beta}}_1^{(0)}$ in Theorem 2 by any one of $\hat{\boldsymbol{\beta}}_2^{(0)}$ to $\hat{\boldsymbol{\beta}}_d^{(0)}$. We opt to use $\hat{\boldsymbol{\beta}}_1^{(0)}$ in Algorithm 1 to estimation ξ because $\hat{\boldsymbol{\beta}}_1^{(0)}$ is more efficient than any one of $\hat{\boldsymbol{\beta}}_2^{(0)}$ to $\hat{\boldsymbol{\beta}}_d^{(0)}$ (Mardia et al. 1979).

3.3. Matrix discriminant analysis for multiple ξs

Now, we consider a slightly more general case that there are multiple classification relevant ξs, say ξ_1, …, ξ_c. Algorithm 1 can be generalized to the applications with multiple ξs using a two-step strategy. In the first step, we obtain $\hat{\boldsymbol{\beta}}_1^{(1)}, \cdots, \hat{\boldsymbol{\beta}}_d^{(1)}$ using Algorithm 1. In the second step, we obtain estimates of $\hat{\boldsymbol{\xi}}_1^{(1)}, \cdots, \hat{\boldsymbol{\xi}}_c^{(1)}$ by solving the following linear system

$$\begin{gathered}\widehat{\mathrm{Var}}(E[\hat{\boldsymbol{\beta}}_1^{(1)'}\mathbf{X}|Y])\boldsymbol{\theta}_i=\nu_i\widehat{\mathrm{Var}}[\hat{\boldsymbol{\beta}}_1^{(1)'}\mathbf{X}]\boldsymbol{\theta}_i, \quad \nu_1 \geq \nu_2 \geq \ldots \geq \nu_c>0 \\ \boldsymbol{\theta}_i'\widehat{\mathrm{Var}}[\hat{\boldsymbol{\beta}}_1^{(1)'}\mathbf{X}]\boldsymbol{\theta}_j=I_{\{i=j\}}.\end{gathered} \qquad (10)$$

where i = 1, ···, c.. As with algorithm 1, convergence is guaranteed since no iteration is used in the second step of the generalized Algorithm 1. Theorems 1 and 2 imply that $\hat{\boldsymbol{\beta}}_1^{(1)}$ and $\hat{\boldsymbol{\xi}}_1^{(1)}$ converge asymptotically to β_1 and ξ_1. As $\hat{\boldsymbol{\xi}}_1^{(1)}$ is a consistent estimator of ξ_1, we can further show by Theorem 1 that $\hat{\boldsymbol{\beta}}_j^{(1)}$ is also a consistent estimator of β_j for j = 2, ···, d. Similarly, we can show that $\hat{\boldsymbol{\xi}}_i^{(1)}$ is a consistent estimator of ξ_i, where i = 2, …, c.

4. Penalized Matrix Discriminant Analysis

As shown in Figure 1C, not all dyes in CSA are chemo-responsive to decylamine and sec-Bu$_2$amine. Some dyes do not change color and appear as black in the color difference map. Among the chemo-responsive dyes, some dyes, such as the dyes in circles in Figure 1C, are classification irrelevant. Building classification rules using nonreponsive dyes and discriminant-irrelevant dyes can reduce classification accuracy. To surmount this challenge, we shall shrink the effect of discriminant-irrelevant dyes and keep the discriminant relevant dyes in the classification analysis. This goal can be achieved by penalizing the L_1 norm of the parameters which specify the dye effect in MDA method (Tibshirani 1996). In this section, we develop the penalized matrix discriminant analysis (PMDA) method to serve this purpose.

Let $\hat{\boldsymbol{\beta}}_j^*$ be the penalized estimate of β_j. Given $\hat{\xi}$, an MDA estimate of ξ, $\hat{\boldsymbol{\beta}}_1^*$ can be obtained by maximizing

$$\frac{\boldsymbol{\eta}'\widehat{\mathrm{Var}}[E(\mathbf{X}\hat{\boldsymbol{\xi}}|Y)]\boldsymbol{\eta}}{\boldsymbol{\eta}'\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}]\boldsymbol{\eta}} \text{subject to } \|\boldsymbol{\eta}\|_1 \leq \rho \qquad (11)$$

over η, where $\| \cdot \|_a$ denotes the L_a norm. When $\rho = \infty$, the maximizer of (11) is the same as $\hat{\beta}_1$ in MDA, and when ρ = 0, the discriminant effects of all dyes are shrunken to zero. In reality, we need to choose an appropriate ρ such that only the discriminant-relevant dyes are included in the discriminant analysis.

Maximizing (11) is equivalent to solving a generalized eigenvalue problem (Zou et al. 2006, Qiao et al. 2009), which can be further coverted to a least squares problem. Following Theorems 2 and 3 in Zou et al. (2006) and Theorem 1 in Qiao et al. (2009), we reformulate (11) as a least squares type of problem. Let

$$\boldsymbol{Z}=\left(\sqrt{n_1/n}(\hat{E}[\mathbf{X}\hat{\boldsymbol{\xi}}|Y=1]-\hat{E}[\mathbf{X}\hat{\boldsymbol{\xi}}]),\ldots,\sqrt{n_K/n}(\hat{E}[\mathbf{X}\hat{\boldsymbol{\xi}}|Y=K]-\hat{E}[\mathbf{X}\hat{\boldsymbol{\xi}}])\right)'$$

so that $\widehat{\mathrm{Var}}[E(\mathbf{X}\hat{\boldsymbol{\xi}}|Y)]=\boldsymbol{Z}'\boldsymbol{Z}$. Let $\mathbf{R}'\mathbf{R}$ be the Cholesky decomposition of $\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}]$, where $\mathbf{R} \in \mathbb{R}^{p\times p}$ is an upper triangular matrix. Given two non-negative tuning parameters ω_1, ω_2, we can find $\hat{\boldsymbol{\beta}}_1^*$ by minimizing

$$\|\mathbf{Z}\mathbf{R}^{-1}\boldsymbol{\alpha}_1-\boldsymbol{Z}\boldsymbol{\beta}_1\|_2^2+\omega_1\boldsymbol{\beta}_1'\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}]\boldsymbol{\beta}_1+\omega_2\|\boldsymbol{\beta}_1\|_1. \quad (12)$$

with respect to α_1 and β_1, where α_1 satisfies the condition $\|\alpha_1\|_2 = 1$. Here α_1 was created only for the purpose of computational convenience. See Zou et al. (2006) and Qiao et al. (2009) for a detailed explanation of the reformulation.

When there are d classification directions, we let $\mathbf{B} = (\beta_1, \ldots, \beta_d)$ and $\mathbf{A} = (\alpha_1, \ldots, \alpha_d)$. Let $\hat{\mathbf{B}}^*=(\hat{\boldsymbol{\beta}}_1^*,\ldots,\hat{\boldsymbol{\beta}}_d^*)$ and $\hat{\mathbf{A}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_d)$ be their estimates. Then $\hat{\mathbf{B}}^*$ can be obtained by minimizing

$$\sum_{j=1}^{d}\left\{\|\mathbf{Z}\mathbf{R}^{-1}\boldsymbol{\alpha}_j-\boldsymbol{Z}\boldsymbol{\beta}_j\|^2+\omega_1\boldsymbol{\beta}_j'\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}]\boldsymbol{\beta}_j+\omega_{2j}\|\boldsymbol{\beta}_j\|_1\right\} \quad (13)$$

subject to $\mathbf{A}'\mathbf{A} = \mathrm{I}_d$. Whereas the same ω_1 is used for all d directions, different ω_{2j} are allowed to penalize different discriminant directions.

Numerically, $\hat{\mathbf{B}}^*$ can be obtained by iteratively minimizing (13) with respect to $\mathbf{A}$ and then $\mathbf{B}$. We first estimate $\mathbf{B}$ given $\mathbf{A} = \hat{\mathbf{A}}$. Let $\tilde{\mathbf{Y}}_j = \mathbf{ZR}^{-1}\hat{\alpha}_j$. $\hat{\mathbf{B}}^*$ can be obtained by solved q independent LASSO problems

$$\min_{\boldsymbol{\beta}_j}\|\tilde{\boldsymbol{Y}}_j-\boldsymbol{Z}\boldsymbol{\beta}_j\|_2^2+\omega_1\boldsymbol{\beta}_j'\widehat{\mathrm{Var}}[\mathbf{X}\hat{\boldsymbol{\xi}}]\boldsymbol{\beta}_j+\omega_{2j}\|\boldsymbol{\beta}_j\|_1. \quad (14)$$

for $1 \le j \le d$. In practice, we can use either the least angle regression (Efron et al. 2004) or coordinate descent method (Friedman et al. 2010) to estimate $\mathbf{B}$. We then replace $\mathbf{B}$ by the $\hat{\mathbf{B}}^*$ obtained in previous step and estimate $\mathbf{A}$. Because the two penalty terms are positive and do not involve $\mathbf{A}$ at all, minimizing (13) is equivalent to minimizing

$$\sum_{j=1}^{d} \left\| \boldsymbol{Z}\hat{\boldsymbol{\beta}}_1^{*} - \mathbf{Z}\mathbf{R}^{-1}\boldsymbol{\alpha}_j \right\|^2, \quad (15)$$

subject to $\mathbf{A}'\mathbf{A} = \mathrm{I}_d$. Because of the orthogonality constraint, (15) is not a least squares problem but a Procrustes problem (Gower and Dijksterhuis 2004). The solution can be obtained by computing a singular value decomposition on $\mathbf{R}^{-1}(\mathbf{Z}'\mathbf{Z})\mathbf{B}^{\hat{*}}$ (Zou et al. 2006; Qiao et al. 2009). Let $\mathbf{UDV}'$ be the singular value decomposition of $\mathbf{R}^{-1}(\mathbf{Z}'\mathbf{Z})\mathbf{B}^{\hat{*}}$, we have that $\hat{\mathbf{A}}= \mathbf{UV}'$. The PMDA method is outlined in the following algorithm.

Algorithm 2 (PMDA algorithm)

- *Run algorithm 1 to obtain* $\hat{\xi}$.
- *Initialize* $\mathbf{A}$ *by* $\hat{\mathbf{A}}$ *such that* $\hat{\mathbf{A}}'\hat{\mathbf{A}} = \mathbf{I}_d$ *where* $\mathbf{I}_d$ *is a* $d \times d$ *identity matrix.*
- Iterate until convergence.
  - *Find* $\mathbf{B}^{\hat{*}}$ *by optimizing d independent penalized least squares functions (*14).
  - *Replace* $\mathbf{B}$ *by* $\mathbf{B}^{\hat{*}}$ *and perform the singular value decomposition* $\mathbf{UDV}'$ *for* $\mathbf{R}^{-1}(\mathbf{Z}'\mathbf{Z})\mathbf{B}^{\hat{*}}$.
  - *Update* $\mathbf{A}$ *by* $\mathbf{UV}'$.
- *Output final estimates* $\mathbf{B}^{\hat{*}}$.

In Algorithm 2, we estimate β_j by optimizing d independent penalized least squares functions. An alternative approach is that, instead of d independent L_1 penalties, we can use a group penalty, e.g., $\sum_{i=1}^{K} \|\mathbf{B}_i\|_2$, where $\mathbf{B}_i$ represents the ith row of $\mathbf{B}$. By using the group penalty, we can select a subset of dyes that are discriminant-relevant across all discriminant directions. We opt to choose the d independent L_1 penalties for the CSA data because we expect to use different sets of dyes for different discriminant directions. As we will see in the real data analysis, different discriminant directions with different sets of dyes represent different chemical interactions.

The key to the success of method is the selection of the tuning parameters. There are two tuning parameters ω_1 and ω_{2j} involved in the optimization of (13). It was shown in Zou et al. (2006) and Qiao et al. (2009) that the optimizer is independent of the selection of ω_1 when $\omega_{2j} = 0$ (no penalty at all). When $\omega_{2j} \neq 0$, the optimization of (13) may be affected by the value of ω_1. However, our extensive simulations suggest that the minimizer of (13) is robust as ω_1 varies in a wide range in (0.01, 1000). Thus, to alleviate the computational cost, we set $\omega_1 = 1$ in our numerical studies. In practice, we select the tuning parameter ω_{2j} by minimizing a K-fold cross validation (CV) of the misclassification error. Because our CSA data has less than 10 observations in each group in general, we randomly partition the sample under the constraint that each class has at least one observation in the training sample, and at least one observation in the test sample. The CV misclassification error is calculated as the average of the misclassification errors in those test samples.

5. Simulation Studies

To assess the performance of the proposed methods, we carry out extensive analysis on simulated data sets.

5.1. Multiple-class discrimination

This simulation is designed to demonstrate the empirical performance of MDA and PMDA methods in discriminating multiple-class observations. We generated 100 data sets from the following model. Let Y be the class label simulated from a multinomial distribution with four classes. We assume the probability of occurrence in each class equal to $\pi_0 = 0.25$. Let $\mathbf{X} = (x_{i,j})$ be a 36 × 3 random matrix predictor. The conditional distribution of $\mathbf{X}$ given Y is simulated from a multivariate normal distribution with conditional mean

$$E[\mathbf{X}|Y=1]=\begin{pmatrix} \mu\mathbf{A} & 0 \\ 0 & 0 \end{pmatrix}_{36\times 3}, \quad E[\mathbf{X}|Y=2]=\begin{pmatrix} -\mu\mathbf{A} & 0 \\ 0 & 0 \end{pmatrix}_{36\times 3},$$
$$E[\mathbf{X}|Y=3]=\begin{pmatrix} 0 & 0 \\ \mu\mathbf{A} & 0 \end{pmatrix}_{36\times 3}, \quad E[\mathbf{X}|Y=4]=\begin{pmatrix} 0 & 0 \\ -\mu\mathbf{A} & 0 \end{pmatrix}_{36\times 3},$$

where $\mathbf{A} = \mathbf{11}'$ and $\mathbf{1} = (1, 1)'$, and conditional variance

$$\begin{array}{lccc} \mathrm{Var}(x_{i,j}|Y=k)=\sigma^2 & (i,j)=(1,2) & \text{or} & (2,1), \\ \mathrm{Var}(x_{i,j}|Y=k)=1 & (i,j) \neq (1,2) & \text{and} & (2,1), \end{array}$$

where $k = 1, \ldots, 4$. We also set $\mathrm{Cov}(x_{i,j}, x_{i',j'}) = 0$ if $(i, j) \neq (i', j')$. Each data set consists of a training sample of sample size m_1 and a test sample of sample size m_2. We apply Fisher's LDA, MDA and PMDA to the training sample. We then apply the models learned from the training sample to the test sample and calculated the misclassification error rate. As a benchmark, we also calculate Bayes misclassifiction error, the optimal misclassification error that we can have by assuming all the parameters are known. The misclassification error for each method is summarized in Table 1.

We see that MDA and PMDA have substantially lower misclassification error rates than Fisher's LDA. This suggests that incorporating the matrix structure of the predictors and reducing the number of parameters are very important in the multiple class discriminant analysis. The PMDA misclassification error is very close the the Bayes misclassification error. In Figure 2, we project the observations onto the first two Fisher's LDA, MDA and PMDA directions. It is easy to see that the first two Fisher's LDA directions lead to unsatisfactory classification result for the test sample, though the clusters are tight in the training sample. In contrast, the first two MDA directions result in relatively scattered clusters in the training sample, but a much better classification result in the test sample. Figure 2 also indicates that PMDA has better discrimination than MDA in the test sample. In order to understand why PMDA outperforms MDA, we illustrated in Figure 3 the test sample misclassification error for PMDA as ω_{2j} in (14) varies. We plotted in Figure 3 the

CV of the misclassification error against the change of the number of predictors. The predictors are selected by changing ϖ_{2j} in (14). The misclassification error drops quickly as we increase the number of predictors, reaches the minimum when the number of predictors is 4, which is the number of discriminant-relevant predictors in this simulation (recall that, the four discriminant-relevant predictors are the first, the second, the 35th and the 36th predictors). Misclassification error increases as we include more predictors in PMDA analysis. This suggests that including more discriminant-irrelevant predictors in discriminant analysis increases the misclassification error, and thus, PMDA has the edge over MDA and LDA in terms of misclassification error.

5.2. Comparison with other Methods

This simulation is designed to compare the empirical performance of MDA and PMDA methods with other competing methods. We simulated 100 data sets from the following model. Let Y be the class label generated from a Bernoulli distribution with success probability 0.5. Let $\mathbf{X} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$, where $\mathbf{u}_i \in \mathbb{R}^{36}$ and $\mathbf{u}_i$ given Y are generated from the following process.

$$\begin{array}{rll} \mathbf{u}_1|Y{=}k{\sim}N(\mathbf{0}_{36\times 3},\textstyle\sum_1), & \textstyle\sum_1{=}(\sigma_{ij}), & \sigma_{ij}{=}0.5^{|i-j|}, \\ \mathbf{u}_2|Y{=}k{\sim}N(|\mathbf{u}_1|{-}0.3(k{+}1),\textstyle\sum_2), & \textstyle\sum_2{=}0.5\mathrm{I}_{36\times 36}, & \\ \mathbf{u}_3|Y{=}k{\sim}N(\mathbf{0}_{36\times 3},\textstyle\sum_3), & \textstyle\sum_3{=}\mathrm{I}_{36\times 36}. & \end{array}$$

where $|\mathbf{u}_1|$ is a vector with each entry as the absolute value of corresponding entry of $\mathbf{u}_1$. In this setting, $\mathbf{u}_3$ does not have any discriminant power between two classes, but $\mathbf{u}_1$ and $\mathbf{u}_2$ together have discriminant power. In each data set, we randomly generated a training sample of 120 observations and a test sample of 500 observations.

We apply MDA and PMDA to the training sample. For comparison, we also apply to the data the following competing methods: 2D-LDA (Li and Yuan 2005), regularized discriminant analysis (Friedman 1989), Fisher's LDA, sparse discriminant analysis (Clemmensen et al. 2011). In particular, we use the R implementation lda for Fisher's LDA, klaR for regularized discriminant analysis, sparseLDA for sparse discriminant analysis. We implemented 2D-LDA, because the code is not available from the authors. For klaR, sparseLDA, the regularization parameters are chosen by cross-validation. When more than one regularization parameter is required, such as RDA, a grid search method is used. We then apply the models learned from the training sample to the test sample and calculate the misclassification error rate. The boxplots of misclassification errors are plotted in Figure 4.

We can see that PMDA has the smallest misclassification error, and MDA has slightly larger misclassification error. Both PMDA and MDA have smaller misclassification error than other competing methods. As expected, LDA gives larger misclassification, and adding regularization, i.e., RDA and sparse LDA, reduces the misclassification error. Since the 2D-LDA does not take into account the between-row correlation, we can see the 2D-LDA has slightly larger misclassification error than LDA. Since RDA is a compromise between linear and quadratic discriminant analysis and provides a nonlinear discriminant boundary, its misclassification error is low relative to other linear discriminant methods.

6. Case Studies

6.1. Classification of CSA data after exposure to TICs at IDLH concentrations

A series of CSA experiments for 147 chemicals were conducted with the aim to classify these chemicals into either non-toxic or one of 20 toxic industrial chemicals (TICs). These 20 TICs are listed as “High Hazard TICs” on the NATO International Task Force 25 and are summarized in Table 2. The experiments consist of seven chemicals in the non-toxic class and in each of the 20 TICs classes. The color changes of all 36 dyes in CSA were measured and recorded as RGB triplets before and two minutes after exposure to TICs at their concentrations that are Immediately Dangerous to Life or Health (IDLH). The primary interest is to assess the prediction accuracy using the CSA and build classification rules that can be used to monitor the chemical exposure in workplaces.

First, we apply MDA to the difference map of 147 chemicals. In order to do that, we need to determine whether the single ξ can summarize the spectral component effect of RGB colors. Following the algorithm in Section 3.3, we calculate the ratios of between-group variation to within-group variation for three ξs, where ξ_1 accounts for 96.35% of all the variation, ξ_2 accounts for 3.23% and ξ_3 for 0.41%. We thus opt to keep a single ξ for the rest of the analysis. The 147 chemicals are projected on the first two MDA directions and the projection is plotted in Figure 5(a). It is easy to see that the two MDA discriminant directions are not adequate to make a clear classification for some closely related chemicals even though some chemicals with different chemical structures can be well discriminated. To determine the number of classification relevant directions, a classic *F* test is employed(Section 6.3, Kshirsagar 1972) to test the null hypothesis that there are *d* classification relevant directions against the alternative hypothesis that there are more than *d* classification relevant directions. Eight classification relevant directions are identified at 0.005 significance level. We then check the coefficients of each dye to interpret the effect of each dye in the classification analysis. The dyes with large coefficients in the first two directions are those that tend to be active on the van der Waals interactions. The van der Waals interaction are commonly used in traditional sensors for simple chemical structure detection. Whereas the dyes with large coefficients in the other six directions are those that respond to intermolecular interactions between the nanoporous pigments and the VCTs. This suggests that those dyes that react to the intermolecular interactions are important to discriminate the toxicants with complicated chemical structures. This also supports the design of the current CSA, which probes a much wider range of chemical interactions than traditional sensors.

To compare the empirical performance of Fisher’s LDA, MDA and PMDA methods on this dataset, we calculate the misclassification error using seven-fold CV. To keep the balance between classes, we randomly select one observation from each class to form the test sample and use the rest of observations as the training sample. Plotted in Figure 5(b) is the CV of misclassification error with its standard error against the number of directions. The plot shows that, for all the three methods, the misclassification errors decrease drastically as we increase the number of discriminant directions up to eight. We also see that MDA method and PMDA method outperform Fisher's LDA method uniformly in reducing the misclassification error. Moreover, there is no significant difference of the misclassification error between MDA and PMDA methods. This is well expected since the design of the CSA is highly optimized to ensure that all dyes respond to TICs at the IDHL concentration. Thus, shrinking the number of dyes in MDA cannot significantly improve the discrimination accuracy. Finally, it is worth pointing out that with 147 observations and 21 classes the standard errors of misclassification error for PMDA method are small.

6.2. Classification of the CSA data after exposure to the TICs at PEL concentrations

A pressing need for the environmental control of industrial chemical workplace and more general epidemiological studies is to accurately monitor low concentrations of TICs because multiple low-level exposures to the TICs may cause extremely serious effects on an individual's health. Thus, 147 difference maps were obtained for the 20 TICs (Table 2) at their Permissible Exposure Level (PEL) before and after five-minute exposure.

Compared to the IDLH concentrations used in the previous example, only a limited number of dyes show significant color change at the PEL concentration. Many dyes do not respond at this concentrations. This gives a significant advantage to PMDA, which assumes the sparseness of underlying classification functions.

We apply Fisher's LDA, MDA and PMDA to the PEL data. For each of the three methods, we calculate misclassification errors using seven-fold CV for different number of discriminant directions. The result is plotted in Figure 6 (a). We can see that PMDA consistently outperforms the other two methods and the optimal classification can be achieved using 13 directions. We further plot in Figure 6 (b) the classification error along the change of the number of classification relevant dyes using the 13 directions. It is easy to see that the misclassification is minimized using 15 classification relevant dyes.

To further understand the performance of PMDA, we choose three TICs in one training sample and project them on the first two directions obtained by MDA and PMDA respectively in Figure 7. We can clearly see that PMDA can better discriminate the three TICs than MDA in both training and test samples. It suggests that imposing sparsity structure can further improve the classification accuracy if the underlying classification rule is indeed sparse.

7. Discussion

In this article, we developed a simple and efficient matrix classification method named MDA to improve the classification sensitivity and specificity of colorimetric sensor array (CSA) data. Our MDA method can be viewed as an extension of Fisher's LDA to the data in the form of matrices. By retaining the matrix structure of the data, MDA can substantially reduce the misclassification error of chemicals that belong to the same chemical class. For general matrix classification with multiple discriminant relevant linear combinations on both rows and columns, we generalized MDA method with a one-step extension. The generalization of MDA algorithm (section 3.3) is fast and easy to implement. However, it may lose classification relevant directions. The method that can pick up all the classification relevant directions is under study and will be introduced in a follow-up publication.

To further reduce the misclassification error, we proposed PMDA method by imposing a sparse structure on discriminating functions using L_1 penalty. Numerical studies suggest that MDA and PMDA outperform many competing linear classification methods. A potential improvement on MDA and PMDA methods is to extend them to nonlinear discriminant analysis by using appropriate kernel functions. In that case, we may need fewer discriminant directions than MDA and PMDA in discriminating matrix data. A potential drawback of using the nonlinear discriminant analysis is that we may lose the model interpretability.

An R code implementing MDA and PMDA methods is provided in the supplementary material.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editor, the associate editor, and two reviewers for insightful comments.

This research is supported by NIH R01 GM113242-01 to WZ, NSF Award 1406843 to WZ, NSF Award 1440038 to WZ, and NIH U01 ES016011 to KS.

References

Anderson, TW. An Introduction to Multivariate Statistical Analysis. 3. New York: John Wiley & Sons, Inc; 2003.

Clemmensen L, Hastie T, Witten D, Ersboll B. Sparse discriminant analysis. Technometrics. 2011; 53:406–413.

Donoho DL, Elad M. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. Proceedings of the National Academy of Sciences. 2003; 100:2197–2202.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Annals of Statistics. 2004; 32:407–499.

Fan, J.; Li, R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. Proceedings of the International Congress of Mathematicians; European Mathematical Society; 2006. p. 595-622.

Feng L, Musto CJ, Kemling JW, Lim SH, Zhong W, Suslick KS. Colorimetric sensor array for determination and identification of toxic industrial chemicals. Analytical Chemistry. 2010; 82:9433–9440. [PubMed: 20954720]

Fisher R. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7:179–188.

Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate desent. Journal of Statistical Software. 2010; 33:1–22. [PubMed: 20808728]

Friedman JH. Regularized discriminant analysis. Journal of American Statistical Association. 1989; 84:165–175.

Gonzales, RC.; Woods, RE. Digital Image Processing. 2. Upper Saddle River, NJ: Prentice-Hall; 2002.

Gower, JC.; Dijksterhuis, GB. Procrustes Problems. Oxford: Oxford University Press; 2004.

Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. New York: Springer-Verlag; 2009.

Kshirsagar, AM. Multivariate Analysis. New York: Marcel Dekker; 1972.

Li B, Kim MK, Altman N. On dimension folding of matrix or array valued statistical objects. Annals of Statistics. 2009; 20:835–39.

Li M, Yuan B. 2D-LDA: A statistical linear discriminant analysis for image matrix. Pattern Recognition Letters. 2005; 26:527–532.

Lim SH, Musto CJ, Park E, Zhong W, Suslick KS. A colorimetric sensor array for detection and identification of sugars. Organic Letters. 2008; 10:4405–4408. [PubMed: 18783231]

Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate Analysis. New York: Academic Press; 1979.

Qiao Z, Zhou L, Huang JZ. Effective linear discriminant analysis for high dimensional low sample size data. IAENG International Journal of Applied Mathematics. 2009; 39:48–60.

Rakow N, Sen A, Janzen M, Ponder J, Suslick K. Molecular recognition and discrimination of amines with a colorimetric array. Angewandte Chemie International Edition. 2005; 44:4528–4532.

Rakow N, Suslick K. A colorimetric sensor array for odour visualization. Nature. 2000; 406:710–714. [PubMed: 10963592]

Stewart, GW.; Sun, J. Matrix Perturbation Theory. New York: Academic Press; 1990.

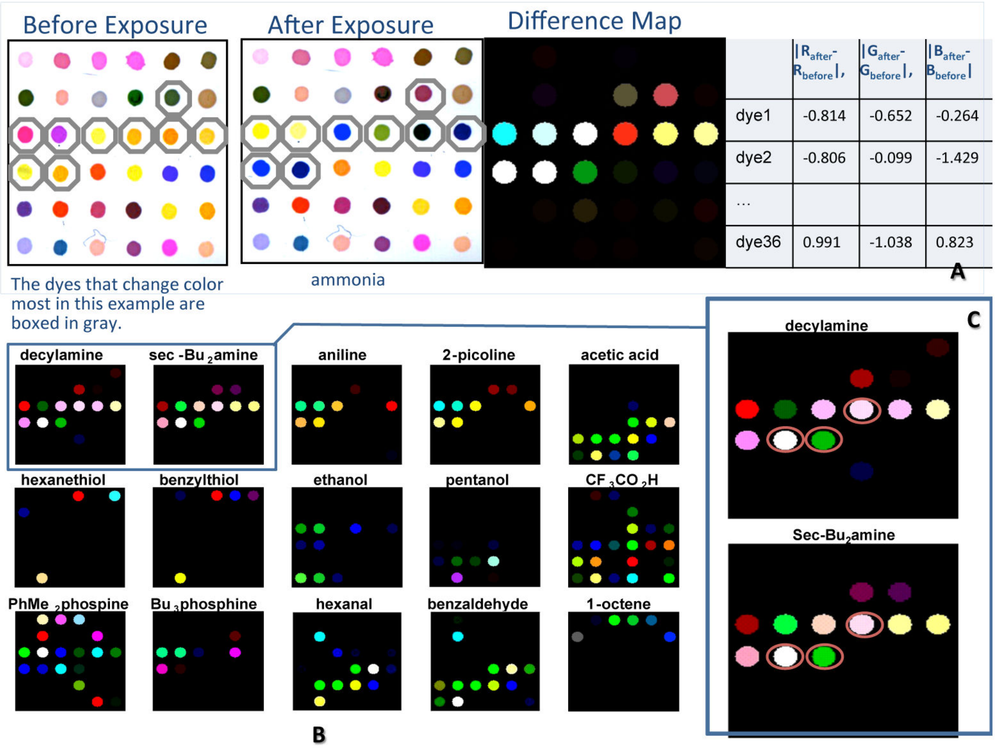
Suslick KBDP, Ingison CK, Janzen M, Kosal ME, McNamara WB III, Rakow NA, Sen A, Weaver JJ, Wilson JB, Zhang C, Nakagaki S. Seeing smells: development of an optoelectronic nose. Quimica Nova. 2007; 30:677–681.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society, Series B. 1996; 58:267–288.

Zhang C, Suslick K. A colorimetric sensor array for organics in water. Journal of the American Chemical Society. 2005; 127:11548–11549. [PubMed: 16104700]

Zheng W, Lai JH, Li SZ. 1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based? Pattern Recognition. 2008; 41:2156– 2172.

Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics. 2006; 15:265–286.



	$|R_{after}-R_{before}|$,	$|G_{after}-G_{before}|$,	$|B_{after}-B_{before}|$
dye1	-0.814	-0.652	-0.264
dye2	-0.806	-0.099	-1.429
…			
dye36	0.991	-1.038	0.823

Figure 1.

Difference maps for colorimetric sensor arrays. Panel A illustrates the calculation of a difference map, generated by digital subtraction at pixel level. The before and after exposure images are subtracted to form the difference map. Differences in each of red, green and blue are averaged over each circle and stored in a 36 × 3 matrix, with each row corresponding to a chemo-responsive dye and each column corresponding to one of the three colors. Plotted in panel B are the color changes of CSA for 15 representative toxic industrial chemicals at their IDLH (immediately dangerous to life or health) concentration after two-minute exposure. Enlargements of color changes for two of the 15 cases are presented in Panel C. Three dyes shown in the red circles are chemo-responsive but discriminant-irrelevant dyes. The black regions, regions show no color changes, are the non-chemo-responsive dyes.

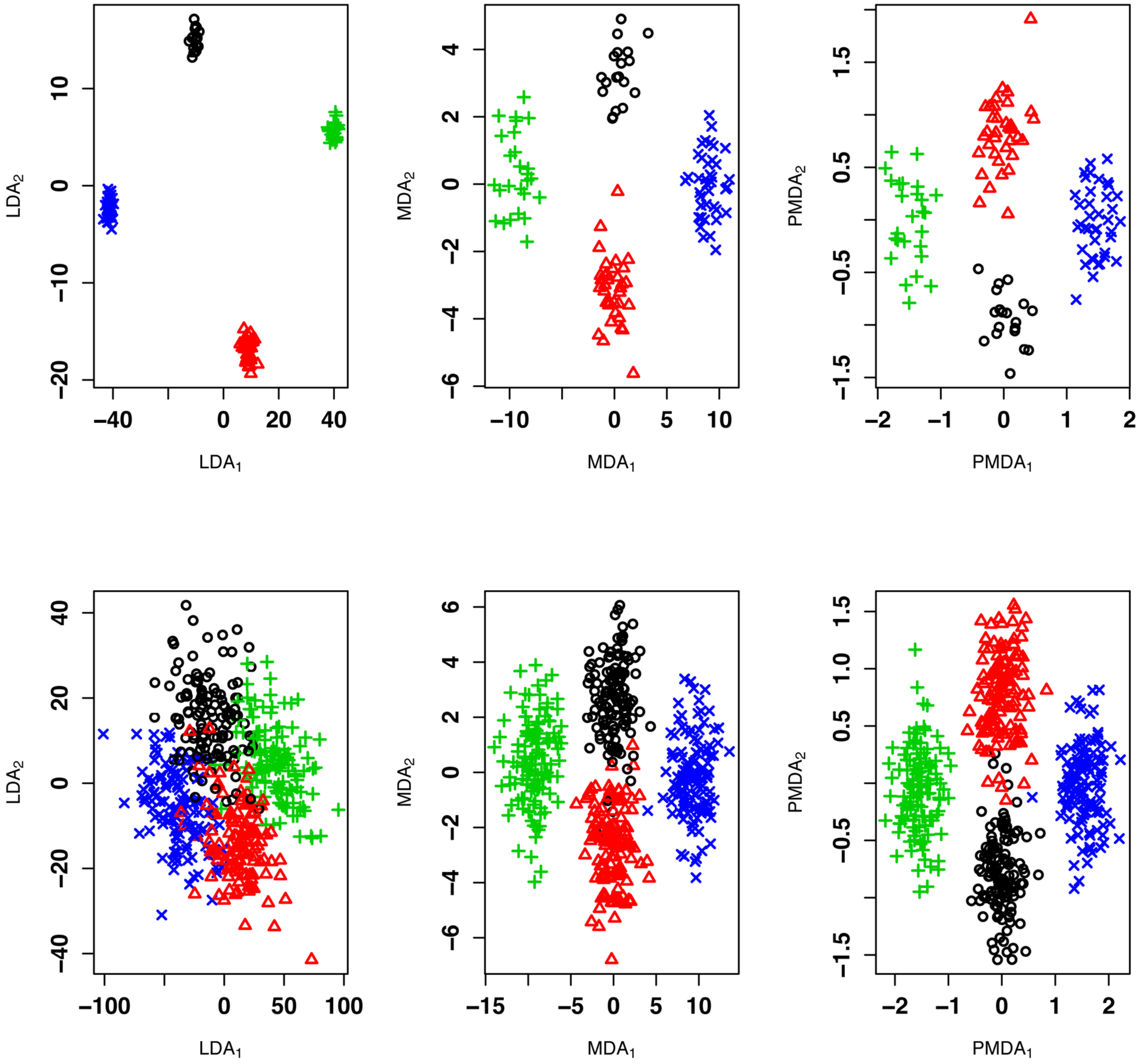


Figure 2.

The upper panels show a training sample projected on the first two Fisher's LDA, MDA and PMDA discriminant directions; The lower panels show the corresponding test sample projected on the first two Fisher's LDA, MDA, and PMDA discriminant directions. The sample was generated from the four-class model in Section 5.1 with $\mu = 3$, $\sigma^2 = 3$, $m_1 = 120$, and $m_2 = 500$. The MDA and PMDA discriminant directions are calculated using $\hat{\boldsymbol{\beta}}_j^{(1)'} X \hat{\boldsymbol{\xi}}^{(1)}$ for $j = 1, 2$ respectively.

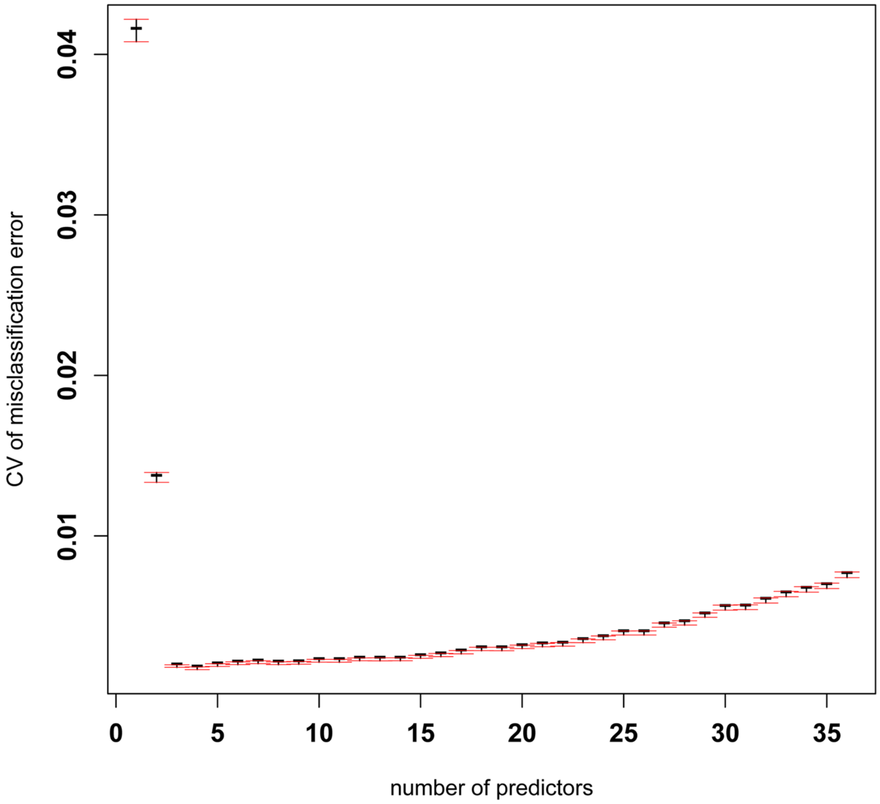


Figure 3.

The misclassification error vs the number of predictors used in PMDA method. The predictors are selected by changing the ω_{2j} in (14)). The middle line is the average misclassification error for each corresponding number of predictors and the upper and lower line are the average misclassification error $\pm 2\times$ standard error. The sample was generated from the four-class model in Section 5.2 with $\mu = 3$, $\sigma^2 = 3$, $m_1 = 120$, and $m_2 = 500$.

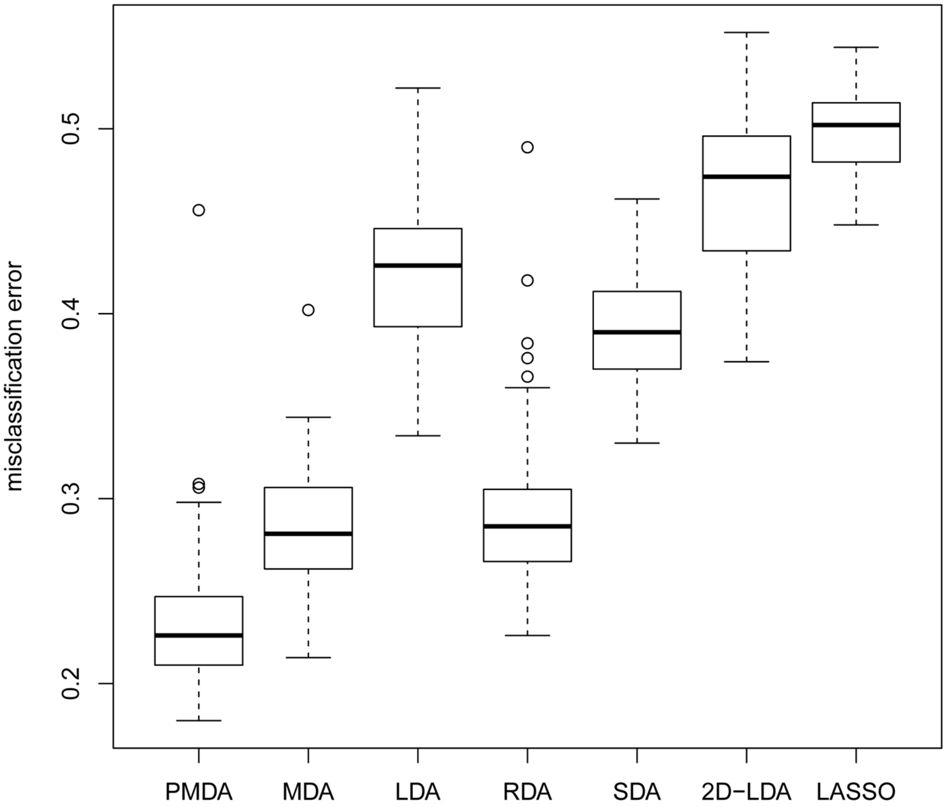
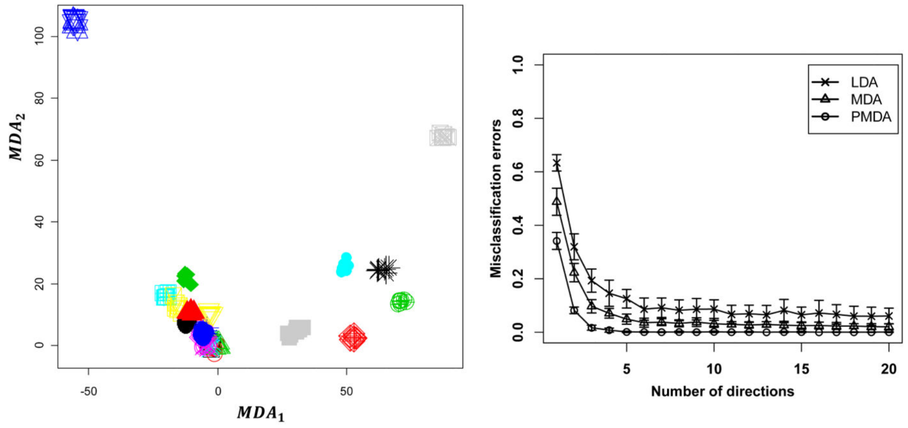


Figure 4.

The boxplots of misclassification error of PMDA, MDA, Fisher's LDA, regularized discriminant analysis (RDA), sparse discriminant analysis (SDA) and 2D-LDA

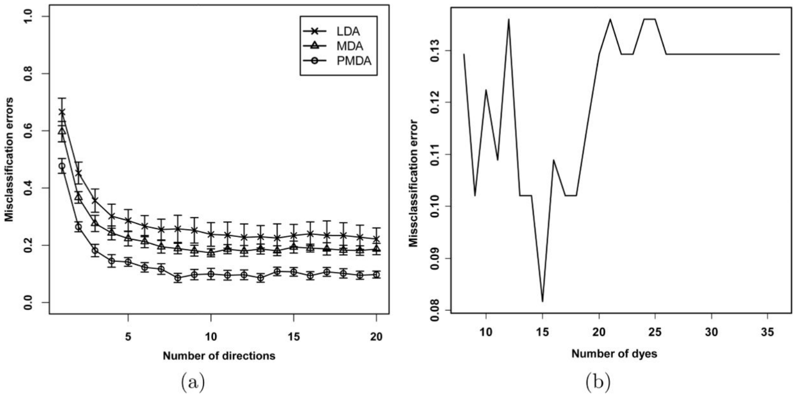


(a)

(b)

Figure 5.

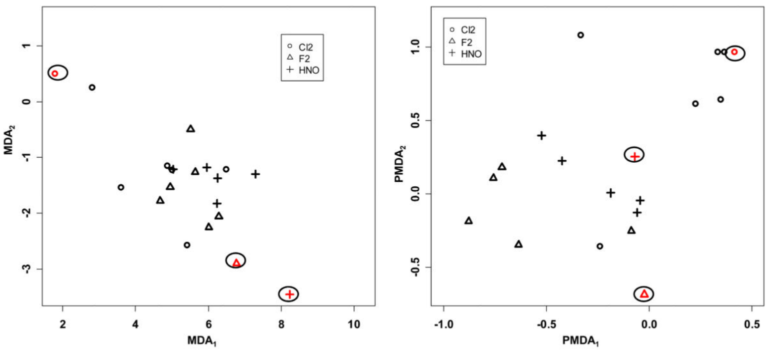
Panel (a) shows the CSA data of the 147 VCTs projected on the first two MDA directions. Panel (b) is the seven-fold CV misclassification error versus the number of discriminant directions used in the classifications.



(a) (b)

Figure 6.

Plotted in (a) is the misclassification error calculated at different number of discriminant directions for CSA data at PEL concentrations. Plotted in (b) is the misclassification error of PMDA for *d* = 13 calculated using different number of dyes.



(a) (b)

Figure 7.

(a)The CSA data of Chlorine (CL2), Formaldehyde (F2) and Nitric Acid (HNO) projected on the first two MDA discriminant directions. (b) The same data projected on the first two PMDA discriminant directions. The test sample is circled.

Table 1

Each entry reports the mean and standard deviation (in parentheses) of the misclassification error calculated based on the 100 test samples.

	$\pi_0 = 0.25$		
Methods	$\mu = 3, \sigma^2 = 3$ $m_1 = 120, m_2 = 500$	$\mu = 1.5, \sigma^2 = 3$ $m_1 = 120, m_2 = 500$	$\mu = 1.5, \sigma^2 = 3$ $m_1 = 200, m_2 = 500$
Bayes	0.000 (0.000)	0.015(0.006)	0.014(0.005)
Fisher’s LDA	0.186 (0.069)	0.468(0.060)	0.126(0.020)
MDA	0.010 (0.018)	0.168 (0.060)	0.110 (0.032)
PMDA	0.003 (0.008)	0.093 (0.053)	0.071(0.023)

Table 2

The list of toxic industrial chemicals (TICs) at their immediately dangerous to life of health (IDLH) and permissible exposure level (PEL) concentrations in ppm.

TIC	IDLH	PEL	Symbol
Ammonia	300	50	⊕
Arsine	3	0.05	○
Chlorine	10	1	⊞
Diborane	15	0.1	+
Dimethylamine	500	10	☒
Fluorine	25	0.1	⊗
Formaldehyde	20	0.75	□
Hydrogen Chloride	50	5	✡
Hydrogen Cyanide	50	10	▲
Hydrogen Fluoride	30	3	◆
Hydrogen Sulfide	100	20	▽
Hydrazine	50	1	•
Methylamine	100	10	*
Methyl Hydrazine	20	0.2	■
Nitric Acid	25	2	●
Nitrogen Dioxide	20	5	⍍
Phosgene	2	0.1	⋄
Phosphine	50	0.3	×
Sulfur Dioxide	100	5	△
Trimethylamine	200	10	⟐