Available online at www.sciencedirect.com
ScienceDirect
Statistics & Probability Letters 76 (2006) 1800–1807
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Masking effects on linear regression in multi-class classification

Chunming Zhang*, Haoda Fu

Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706, USA

Received 10 June 2005; received in revised form 22 December 2005; accepted 24 April 2006
Available online 6 June 2006

Abstract

The linear regression method belongs to the important class of linear methods for multi-class classification. Empirical evidences suggest that a masking problem occurs with the linear regression approach and it is especially prevalent when the number of classes is large. This paper provides an analytical study of this issue and explicitly explains why the linear discriminant analysis procedure removes this problem.

Keywords: Classifier; Decision boundary; Linear discriminant analysis; Linear regression

1. Introduction

Linear methods have been widely used in classification. They partition the input space into a collection of disjoint regions, separated by linear decision boundaries. Notable examples of linear classifiers include those by linear regression method (LRM), linear discriminant analysis (LDA), logistic regression, and separating hyperplanes. A detailed account of these methods can be found in Hastie et al. (2001, Chapter 4). As an illustration, Fig. 1 displays the linear decision boundaries, formed by LRM in the left plot and LDA in the right plot, between three classes in a two-dimensional input space. These two procedures seem to have very similar performances in classification. Unlike the one-versus-the-rest strategy, which makes use of a series of two-class classifiers, both LRM and LDA solve the multi-class classification in a direct fashion by optimizing certain discriminant functions. This feature makes LRM and LDA simple and attractive. See descriptions in the subsequent section.

Curiously, the LRM may result in a masking problem, namely, some classes fail to be separated. Henceforth, the observed number of classes, based on the decision boundaries produced by LRM, is fewer than the actual number of classes. This masking effect is illustrated in Fig. 2, where the left plot reveals that the center class is completely masked by the outside two. In contrast, the right plot indicates that all three classes are perfectly separated by LDA. Hastie et al. (1994, p. 1267) also illustrate a similar masking problem with the *softmax* procedure. When the number of classes increases, the masking effect gets more serious. Fig. 3 makes it evident that for the number of classes equal to 4, only two classes can be separated by LRM, whereas the LDA rule does not suffer from this problem.

*Corresponding author. Tel.: +1 608 262 0084; fax: +1 608 262 0032.
E-mail addresses: cmzhang@stat.wisc.edu (C. Zhang), fuhaoda@stat.wisc.edu (H. Fu).

0167-7152/$ - see front matter
doi:10.1016/j.spl.2006.04.047

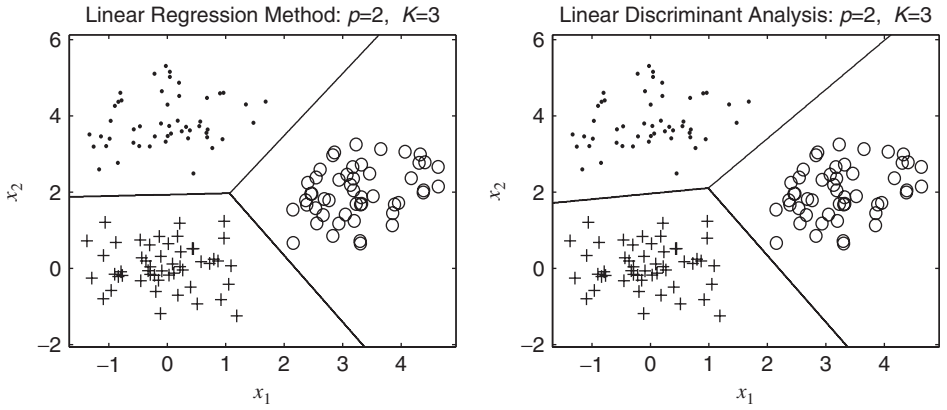

Fig. 1. The data come from three classes in $\mathbb{R}^2$ and are easily separated by linear decision boundaries. The left plot shows the boundaries found by linear regression method. The right plot shows the boundaries found by linear discriminant analysis. The sample size of each class is 50.

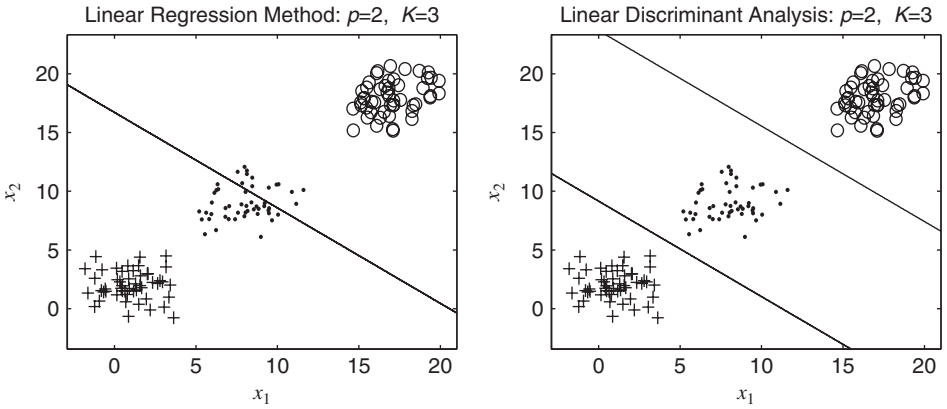

Fig. 2. The data come from three classes in $\mathbb{R}^2$ and are easily separated by linear decision boundaries. The left plot shows the boundaries found by linear regression method; the middle class is completely masked. The right plot shows the boundaries found by linear discriminant analysis. The sample size of each class is 50.

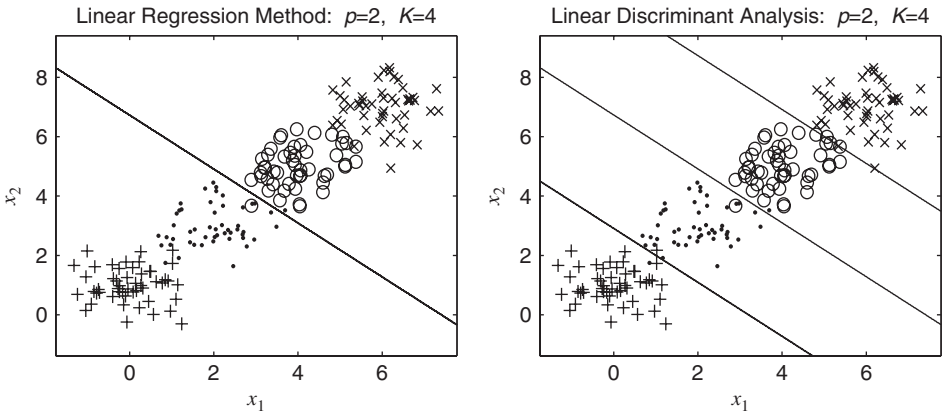

Fig. 3. The caption is similar to that of Fig. 2, except that the number of classes is 4.

However, the reason for the masking effects is not convincingly clear. Some projection plots that help better view the masking effects in the case of three classes is given in Hastie et al. (2001, p. 84). While the graphical plots are helpful for illustrative purposes, they are limited to input spaces of dimension one. The challenge of visualization considerably grows with the higher dimensions of input spaces. This leads to difficulty in understanding the cause and likely impact of masking problems on classification for high-dimensional data. As far as we are aware, no published information exists to explicitly explain this empirical result, and hence a more careful study is needed to yield insight into the masking effects.

The aim of this article is to provide an analytical study, which is applicable to more general cases with no restriction on either the dimension of the input space or the number of outcome classes. Moreover, the study explicitly explains why the classification procedure based on LDA removes the masking problem.

2. Linear regression method and linear discriminant analysis

We first briefly describe LRM and LDA. Suppose there are K classes, p predictor variables, and n records in the training sample, $\{(\mathtt{x}_i, g_i),\ i = 1, \ldots, n\}$, where $\mathtt{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the ith feature vector in a p-dimensional input space, and g_i is the class label taking values in the discrete set $\mathscr{G} = \{1, \ldots, K\}$. The multi-class classification task refers to assigning an observation with input value $\mathtt{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ into one of K classes.

2.1. LRM: linear regression of an indicator response matrix

As described in LeBlanc and Tibshirani (1996, Section 7) and Hastie et al. (2001, Section 4.2), for each observation, LRM codes the response by a response vector $\mathbf{y} = (Y_1, \ldots, Y_K)^{\mathrm{T}}$ of 0–1 entries. That is, $Y_k = 1$ if the observation falls in class k and $Y_k = 0$ otherwise. For the ith training response g_i, $i = 1, \ldots, n$, the response vector $\mathbf{y}_i$ has the value $\mathbf{y}_i = \boldsymbol{e}_k$ if $g_i = k$, in which $\boldsymbol{e}_k$ is the kth column of a $K \times K$ identity matrix. The training sample of size n forms an $n \times K$ indicator response matrix,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{y}_n^{\mathrm{T}} \end{bmatrix}$$

and an $n \times (p+1)$ design matrix,

$$\mathbf{X} = \begin{bmatrix} 1 & \mathtt{x}_1^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & \mathtt{x}_n^{\mathrm{T}} \end{bmatrix}.$$

LRM fits a linear regression model to columns of $\mathbf{Y}$ simultaneously, leading to the fit, $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}$, where $\widehat{\mathbf{B}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is the coefficient matrix. A new observation with input $\mathtt{x}$ is classified as follows. Compute the fitted output, $\widehat{\mathbf{m}}(\mathtt{x}) = (\widehat{m}_1(\mathtt{x}), \ldots, \widehat{m}_K(\mathtt{x}))^{\mathrm{T}}$, where

$$\widehat{\mathbf{m}}(\mathtt{x}) = \{(1, \mathtt{x}^{\mathrm{T}})\widehat{\mathbf{B}}\}^{\mathrm{T}}. \tag{2.1}$$

The class membership is predicted by

$$\widehat{G}_{\mathrm{LRM}}(\mathtt{x}) = \arg\max_{k \in \mathscr{G}} \widehat{m}_k(\mathtt{x}). \tag{2.2}$$

We wish to emphasize here the distinction between LRM for "classification problem" and the traditional linear regression method for "regression problem". In classification, each response data is a K-variate vector, whereas in regression, the response is a scalar. Hence, some well-known results on least-squares regression estimates for conventional linear models (see Rao, 1973, among others) are not straightforwardly applicable for explaining the masking effects on linear regression method in K-class classification, especially when $K \geqslant 2$ is large.

2.2. Linear discriminant analysis (LDA)

Assume that for each class $k \in \mathcal{G}$, the class-conditional density of x is multivariate Gaussian with mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ. The LDA predicts a class label based on a linear discriminant function,

$$\delta_k(\mathrm{x}) = \mathrm{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_k - 2^{-1}\boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_k + \ln(\pi_k), \quad k = 1, \ldots, K,$$

where π_k is the prior probability of class k.

In practice, we do not know the true values of parameters $\{\pi_k\}$, $\{\boldsymbol{\mu}_k\}$, and Σ, and will need to derive estimates from the training data. The estimates are given by

- $\widehat{\pi}_k = n_k/n$, where n_k is the number of class-k observations and is fixed,
- $\widehat{\boldsymbol{\mu}}_k = \sum_{g_i=k} \mathrm{x}_i/n_k$,
- $\widehat{\Sigma} = \sum_{k=1}^{K}\sum_{g_i=k}(\mathrm{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathrm{x}_i - \widehat{\boldsymbol{\mu}}_k)^{\mathrm{T}}/(n-K)$.

Correspondingly, the linear discriminant function is estimated by

$$\widehat{\delta}_k(\mathrm{x}) = \mathrm{x}^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\boldsymbol{\mu}}_k - 2^{-1}\widehat{\boldsymbol{\mu}}_k^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\boldsymbol{\mu}}_k + \ln(\widehat{\pi}_k). \tag{2.3}$$

For a new observation with input x, the predicted class label is given by

$$\widehat{G}_{\mathrm{LDA}}(\mathrm{x}) = \arg\max_{k\in\mathcal{G}} \widehat{\delta}_k(\mathrm{x}). \tag{2.4}$$

In stark contrast to LDA, the normality assumption is not required in LRM. Fuller details on LDA can be found in many statistics textbooks, e.g., Johnson and Wichern (1992).

3. Understanding the masking effects

To understand how and why the masking problem occurs with LRM, we first explicitly derive the expression of $\widehat{\mathbf{m}}(\mathrm{x})$, upon which the performance of LRM depends. To facilitate discussion, we assume, without loss of generality, that the training data have been re-arranged, so that the first n_1 data points are from the first class, the next n_2 data points are from the second class, and so on. In this way, $\mathbf{Y}$, $\mathbf{X}$ and $\widehat{\boldsymbol{\mu}}_k$ can be rewritten as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n_1}\boldsymbol{e}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{1}_{n_K}\boldsymbol{e}_K^{\mathrm{T}} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{X}_1 \\ \vdots & \vdots \\ \mathbf{1}_{n_K} & \mathbf{X}_K \end{bmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\mu}}_k = \frac{\mathbf{X}_k^{\mathrm{T}}\mathbf{1}_{n_k}}{n_k}, \ k = 1, \ldots, K, \tag{3.1}$$

where $\mathbf{1}_m$ denotes an $m \times 1$ vector of ones, $\boldsymbol{e}_k$ is the kth column of a $K \times K$ identity matrix, and $\mathbf{X}_k$ is an $n_k \times p$ matrix which consists of row vectors $\mathrm{x}_i^{\mathrm{T}}$ coming from the kth class. Clearly, $n = n_1 + \cdots + n_K$.

Lemma 1. *Assume that* $\mathrm{rank}(\mathbf{X}) = p + 1$. *For LRM, we have that*

$$\widehat{\mathbf{m}}(\mathrm{x}) = \sum_{k=1}^{K}\widehat{\pi}_k\boldsymbol{e}_k + n\sum_{k=1}^{K}\widehat{\pi}_k\boldsymbol{e}_k\left\{\left(\widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K}\widehat{\pi}_j\widehat{\boldsymbol{\mu}}_j\right)^{\mathrm{T}} M^{-1}\left(\mathrm{x} - \sum_{l=1}^{K}\widehat{\pi}_l\widehat{\boldsymbol{\mu}}_l\right)\right\}, \tag{3.2}$$

where $M = \sum_{k=1}^{K}\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k - n(\sum_{j=1}^{K}\widehat{\pi}_j\widehat{\boldsymbol{\mu}}_j)(\sum_{l=1}^{K}\widehat{\pi}_l\widehat{\boldsymbol{\mu}}_l^{\mathrm{T}})$.

Proof. We observe from (3.1) that

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} n & \sum_{l=1}^{K} n_l\widehat{\boldsymbol{\mu}}_l^{\mathrm{T}} \\ \sum_{j=1}^{K} n_j\widehat{\boldsymbol{\mu}}_j & \sum_{k=1}^{K}\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k \end{bmatrix}$$

and

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{bmatrix} \sum_{k=1}^{K} n_k \boldsymbol{e}_k^{\mathrm{T}} \\ \sum_{k=1}^{K} n_k \widehat{\boldsymbol{\mu}}_k \boldsymbol{e}_k^{\mathrm{T}} \end{bmatrix}.$$

Applying the formula (Zhang, 1999, p. 31) below for the inverse of a partitioned matrix,

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & -A_{11}^{-1}A_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -A_{21}A_{11}^{-1} & \mathbf{I} \end{bmatrix},$$

we have that

$$\begin{aligned} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} &= \begin{bmatrix} 1 & -\sum_{l=1}^{K} \widehat{\pi}_l \widehat{\boldsymbol{\mu}}_l^{\mathrm{T}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} 1/n & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & M^{-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ -\sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} 1/n & -(\sum_{l=1}^{K} \widehat{\pi}_l \widehat{\boldsymbol{\mu}}_l^{\mathrm{T}})M^{-1} \\ \mathbf{0} & M^{-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ -\sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j & \mathbf{I} \end{bmatrix}, \end{aligned}$$

in which $\mathbf{I}$ denotes an appropriately dimensioned identity matrix. Therefore,

$$\begin{aligned} \widehat{\mathbf{B}} &= \begin{bmatrix} 1/n & -(\sum_{l=1}^{K} \widehat{\pi}_l \widehat{\boldsymbol{\mu}}_l^{\mathrm{T}})M^{-1} \\ \mathbf{0} & M^{-1} \end{bmatrix} \begin{bmatrix} \sum_{k=1}^{K} n_k \boldsymbol{e}_k^{\mathrm{T}} \\ \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j) \boldsymbol{e}_k^{\mathrm{T}} \end{bmatrix} \\ &= \begin{bmatrix} 1/n(\sum_{k=1}^{K} n_k \boldsymbol{e}_k^{\mathrm{T}}) - (\sum_{l=1}^{K} \widehat{\pi}_l \widehat{\boldsymbol{\mu}}_l^{\mathrm{T}})M^{-1}\{\sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j)\boldsymbol{e}_k^{\mathrm{T}}\} \\ M^{-1}\{\sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j)\boldsymbol{e}_k^{\mathrm{T}}\} \end{bmatrix}. \end{aligned}$$

This implies that

$$\begin{aligned} (1, \mathrm{x}^{\mathrm{T}})\widehat{\mathbf{B}} &= 1/n \left(\sum_{k=1}^{K} n_k \boldsymbol{e}_k^{\mathrm{T}} \right) - \left(\sum_{l=1}^{K} \widehat{\pi}_l \widehat{\boldsymbol{\mu}}_l^{\mathrm{T}} \right) M^{-1} \left\{ \sum_{k=1}^{K} n_k \left(\widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j \right) \boldsymbol{e}_k^{\mathrm{T}} \right\} \\ &\quad + \mathrm{x}^{\mathrm{T}} M^{-1} \left\{ \sum_{k=1}^{K} n_k \left(\widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j \right) \boldsymbol{e}_k^{\mathrm{T}} \right\} \\ &= \sum_{k=1}^{K} \widehat{\pi}_k \boldsymbol{e}_k^{\mathrm{T}} + \left(\mathrm{x} - \sum_{l=1}^{K} \widehat{\pi}_l \widehat{\boldsymbol{\mu}}_l \right)^{\mathrm{T}} M^{-1} \left\{ \sum_{k=1}^{K} n_k \left(\widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K} \widehat{\pi}_j \widehat{\boldsymbol{\mu}}_j \right) \boldsymbol{e}_k^{\mathrm{T}} \right\}, \end{aligned}$$

which along with (2.1) complete the proof. □

With equal sample class sizes, it follows that

$$n_k \equiv n/K, \quad \widehat{\pi}_k \equiv 1/K, \quad \sum_{k=1}^{K} \widehat{\pi}_k \boldsymbol{e}_k = K^{-1}\mathbf{1}_K,$$

and thus expression (3.2) reduces to

$$\widehat{\mathbf{m}}(\mathrm{x}) = K^{-1}\mathbf{1}_K + \frac{n}{K} \sum_{k=1}^{K} \boldsymbol{e}_k \left\{ \left(\widehat{\boldsymbol{\mu}}_k - \frac{\sum_{j=1}^{K} \widehat{\boldsymbol{\mu}}_j}{K} \right)^{\mathrm{T}} M^{-1} \left(\mathrm{x} - \frac{\sum_{l=1}^{K} \widehat{\boldsymbol{\mu}}_l}{K} \right) \right\}. \tag{3.3}$$

For simplicity, the rest of the paper will focus on equal-sized classes. Extensions to classes of unequal size can similarly be made with suitable modifications wherever needed.

3.1. LRM: classes may be masked

Theorem 1 sheds some light on the reason for the masking effects associated with the LRM.

Theorem 1. *Assume that* $\text{rank}(\mathbf{X}) = p + 1$. *Suppose that the sample class centroids are located along a straight line and are distinct. Assume that the sample class sizes are equal. For any dimension* $p \geqslant 1$ *and any class number* $K \geqslant 2$, *at least* $K - 2$ *classes will be masked by others when the LRM is used in the classifier.*

Proof. Notice that the sample class centroids agree with $\widehat{\boldsymbol{\mu}}_k$, $k = 1, \ldots, K$. Using Cartesian coordinates, the sample class centroids can be written in the parametric form,

$$\widehat{\boldsymbol{\mu}}_k = \mathbf{c} + t_k \mathbf{d}, \quad k = 1, \ldots, K, \tag{3.4}$$

for scalars t_k, an intercept vector $\mathbf{c}$ and a gradient vector $\mathbf{d}$. In addition, we deduce that $\mathbf{d} \neq \mathbf{0}$. Denote $\bar{t} = \sum_{j=1}^{K} t_j / K$. Then

$$\frac{\sum_{j=1}^{K} \widehat{\boldsymbol{\mu}}_j}{K} = \mathbf{c} + \bar{t}\mathbf{d},$$

$$\widehat{\boldsymbol{\mu}}_k - \frac{\sum_{j=1}^{K} \widehat{\boldsymbol{\mu}}_j}{K} = (t_k - \bar{t})\mathbf{d}.$$

These two equations applied to (3.3) lead to the expression,

$$\widehat{\mathbf{m}}(\mathrm{x}) = \frac{1}{K}\mathbf{1}_K + \frac{n}{K}\left\{\sum_{k=1}^{K}(t_k - \bar{t})\boldsymbol{e}_k\right\}\{\mathbf{d}^{\mathrm{T}} M^{-1}(\mathrm{x} - \mathbf{c} - \bar{t}\mathbf{d})\}.$$

Thus, the kth component of $\widehat{\mathbf{m}}(\mathrm{x})$ becomes

$$\widehat{m}_k(\mathrm{x}) = 1/K + (n/K)(t_k - \bar{t})\{\mathbf{d}^{\mathrm{T}} M^{-1}(\mathrm{x} - \mathbf{c} - \bar{t}\mathbf{d})\},$$

which depends on k only through a linear term of t_k.

Without loss of generality, assume that $t_1 < \cdots < t_K$. For any point $\mathrm{x} \neq \mathbf{c} + \bar{t}\mathbf{d}$, the maximum value of $\{\widehat{m}_k(\mathrm{x}), k = 1, \ldots, K\}$ can only be achieved at either $k = 1$ or $k = K$, depending on whether the sign of the multiplicative constant, $\mathbf{d}^{\mathrm{T}} M^{-1}(\mathrm{x} - \mathbf{c} - \bar{t}\mathbf{d})$, is negative or positive. Namely, the predicted class label, $\widehat{G}_{\mathrm{LRM}}(\mathrm{x})$, can only take values equal to either 1 or K. This is the reason at most two distinct decision boundaries, $\{\mathrm{x} : \widehat{G}_{\mathrm{LRM}}(\mathrm{x}) = 1\}$ and $\{\mathrm{x} : \widehat{G}_{\mathrm{LRM}}(\mathrm{x}) = K\}$, rather than K distinct decision boundaries, can be found by LRM. □

Now let us revisit Figs. 2 and 3, in which the sample centroids are indeed located along a straight line and are distinct. The observed masking effects on LRM agree with the conclusion of Theorem 1.

3.2. LRM: maximum number of classes

A further question is: for a fixed dimension p of input space, is there any maximum number of classes without being masked when the LRM is used? Theorem 2 below guarantees that in certain instances, there is no maximum number of classes without being masked when the LRM is used.

Theorem 2. *Assume that* $\text{rank}(\mathbf{X}) = p + 1$. *Let* $\widetilde{\boldsymbol{\mu}}_k = \widehat{\boldsymbol{\mu}}_k - \sum_{j=1}^{K} \widehat{\boldsymbol{\mu}}_j / K$ *be the centralized sample class centroid. If for any* $k \neq j$, *where* $k, j = 1, \ldots, K$, *the conditions*

$$\widetilde{\boldsymbol{\mu}}_k^{\mathrm{T}} M^{-1} \widetilde{\boldsymbol{\mu}}_k = \widetilde{\boldsymbol{\mu}}_j^{\mathrm{T}} M^{-1} \widetilde{\boldsymbol{\mu}}_j, \tag{3.5}$$

$$\widetilde{\boldsymbol{\mu}}_k \neq \widetilde{\boldsymbol{\mu}}_j \tag{3.6}$$

hold, then the LRM can separate out all of the K classes.

Proof. To prove that every class can be separated, we only need to prove that for each class we can find at least one input point to be classified to this class. For the jth class, we choose the point to be the sample class

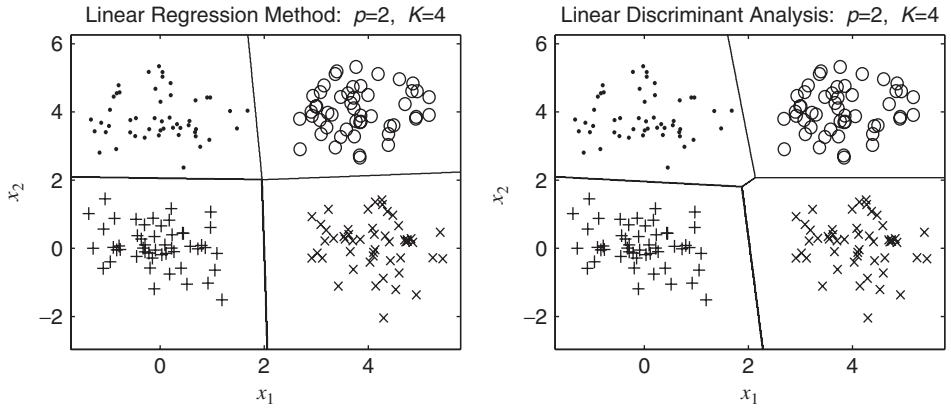

Fig. 4. The data come from four classes in $\mathbb{R}^2$ and are easily separated by linear decision boundaries. The left plot shows the boundaries found by linear regression method. The right plot shows the boundaries found by linear discriminant analysis. The sample size of each class is 50.

centroid $\widehat{\boldsymbol{\mu}}_j$. For the positive definite matrix M, we can define the inner product and norm as

$$\langle \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\mu}}_j \rangle = \widetilde{\boldsymbol{\mu}}_k^{\mathrm{T}} M^{-1} \widetilde{\boldsymbol{\mu}}_j,$$

$$\|\widetilde{\boldsymbol{\mu}}_j\| = \sqrt{\langle \widetilde{\boldsymbol{\mu}}_j, \widetilde{\boldsymbol{\mu}}_j \rangle}.$$

By the Cauchy–Schwarz inequality and (3.5)–(3.6), we know that

$$|\langle \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\mu}}_j \rangle| \leqslant \|\widetilde{\boldsymbol{\mu}}_k\|^2, \tag{3.7}$$

in which the equality is satisfied if and only if $k = j$. For the input $\widehat{\boldsymbol{\mu}}_j$, (3.3) indicates that

$$\widehat{m}_k(\widehat{\boldsymbol{\mu}}_j) = 1/K + (n/K)\langle \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\mu}}_j \rangle.$$

According to (3.7), the maximum value of $\{\widehat{m}_k(\widehat{\boldsymbol{\mu}}_j), k = 1, \ldots, K\}$ is uniquely achieved at $k = j$. Because the index j is arbitrary, this proves that, under conditions (3.5) and (3.6), any sample class centroid can always be classified to its own class, and thus no classes will be masked when LRM is used. □

Conditions (3.5) and (3.6) can be approximately satisfied when the sample class centroids are located on the surface of a spherical ball. In Fig. 4 with $p = 2$ and $K = 4$, the sample class centroids are symmetrically located along a circle, which is a two-dimensional spherical ball. From the left plot, we can see that no classes are masked by others when LRM is used. Analogously, the disappearance of masking effects from LRM in Fig. 1 lends convincing support to Theorem 2.

3.3. LDA: classes will not be masked

It is natural to ask the question why the masking problem does not arise from the LDA method when the sample class centroids are located along a straight line and are distinct. In this case, it is easy to see from (2.3) and (3.4) that

$$\widehat{\delta}_k(\mathrm{x}) = \mathrm{x}^{\mathrm{T}} \widehat{\Sigma}^{-1}(\mathbf{c} + t_k \mathbf{d}) - 2^{-1}(\mathbf{c} + t_k \mathbf{d})^{\mathrm{T}} \widehat{\Sigma}^{-1}(\mathbf{c} + t_k \mathbf{d}) + \ln(1/K).$$

To prove that no classes will be masked, it suffices to prove that for each class we can find at least one input point to be classified to this class. To this end, for each class $j = 1, \ldots, K$, we consider the sample class centroid $\widehat{\boldsymbol{\mu}}_j$. Then

$$\widehat{\delta}_k(\widehat{\boldsymbol{\mu}}_j) = (\mathbf{c} + t_j \mathbf{d})^{\mathrm{T}} \widehat{\Sigma}^{-1}(\mathbf{c} + t_k \mathbf{d}) - 2^{-1}(\mathbf{c} + t_k \mathbf{d})^{\mathrm{T}} \widehat{\Sigma}^{-1}(\mathbf{c} + t_k \mathbf{d}) + \ln(1/K).$$

To seek the optimal k to maximize $\widehat{\delta}_k(\widehat{\boldsymbol{\mu}}_j)$, we only need to consider the optimal λ to maximize the function,

$$(\mathbf{c}+t_j\mathbf{d})^{\mathrm{T}}\widehat{\Sigma}^{-1}(\mathbf{c}+\lambda\mathbf{d})-2^{-1}(\mathbf{c}+\lambda\mathbf{d})^{\mathrm{T}}\widehat{\Sigma}^{-1}(\mathbf{c}+\lambda\mathbf{d})+\ln(1/K). \tag{3.8}$$

Since (3.8) is quadratic in λ, it can easily be verified that the maximizer of (3.8) is uniquely achieved at $\lambda=t_j$. This in turn demonstrates that the maximum value of $\widehat{\delta}_k(\widehat{\boldsymbol{\mu}}_j)$ is uniquely achieved at $k=j$, that is, $\widehat{G}_{\mathrm{LDA}}(\widehat{\boldsymbol{\mu}}_j)=j$. Therefore, when the sample class centroids are located along a straight line, no classes will be masked when LDA is used.

4. Discussion

There is a diverse and extensive literature addressing classification methods. A comparison of 33 old and new classification methods (including LDA as one of the top performers) on 32 data sets is given in Lim et al. (2000), which indicates on p. 204 that "The StatLog Project finds that no algorithm is uniformly most accurate over the datasets studied. Instead, many algorithms possess comparable accuracy". The current paper stresses a point that has been often ignored: LRM, as a computationally simpler and distributionally more robust classifier, is as good as LDA in some situations (as in Figs. 1 and 4) without suffering from the masking problem. Furthermore, we explicitly examine under what circumstances and to what extent the masking effects (as in Figs. 2 and 3) may occur with LRM but disappear with LDA. In summary, our analytical study contributes to an improved understanding of the masking effects and supplements some empirical comparison between LRM and LDA, other than encouraging the use of LRM (or LDA) in all situations. Indeed, following the combination scheme proposed in LeBlanc and Tibshirani (1996), both LRM and LDA classifiers can be combined with others to obtain one which is better than any of the individual classifier.

Acknowledgment

The research is supported in part by National Science Foundation Grant DMS-03-53941 and Wisconsin Alumni Research Foundation. The authors thank the editor, Professor Richard Johnson, and a referee for their constructive suggestions that greatly improved the presentation of this paper. The authors are also grateful to Yuefeng Lu for helpful comments.

References

Hastie, T., Tibshirani, R., Buja, A., 1994. Flexible discriminant analysis by optimal scoring. J. Amer. Statist. Assoc. 89, 1255–1270.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.
Johnson, R.A., Wichern, D.W., 1992. Applied Multivariate Statistical Analysis, third ed. Prentice-Hall, NJ.
LeBlanc, M., Tibshirani, R., 1996. Combining estimates in regression and classification. J. Amer. Statist. Assoc. 91, 1641–1650.
Lim, T.-S., Loh, W.-Y., Shih, Y.-S., 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning J. 40, 203–228.
Rao, C.R., 1973. Linear Statistical Inference and its Applications, second ed. Wiley, New York.
Zhang, F., 1999. Matrix Theory. Basic Results and Techniques. Springer, New York.