

# Discriminant Analysis through a Semiparametric Model

By Y. LIN and Y. JEON

Department of Statistics, University of Wisconsin,

Madison, Wisconsin, 53706, U.S.A.

yilin@stat.wisc.edu

yjeon@stat.wisc.edu

## SUMMARY

We consider a semiparametric generalisation of normal-theory discriminant analysis. The semiparametric model assumes that, after unspecified univariate monotone transformations, the class distributions are multivariate normal. We introduce an estimation procedure based on the distribution quantiles, in which the parameters of the semiparametric model are estimated directly without estimating the nonparametric transformations. The procedure is computationally fast and the estimation accuracy is shown to have the usual parametric rate. The relationship between the method and more general nonparametric discriminant analysis is discussed. The semiparametric specification of the class densities is a submodel of the nonparametric log density functional analysis of variance model in which the main effects are completely nonparametric but the interaction terms are specified semiparametrically. Simulations and real examples are used to illustrate the procedure.

*Some key words:* Distribution quantile; Linear discriminant analysis; Monotone transformation; Naive Bayes method; Nonparametric functional analysis of variance model; semiparametric discriminant analysis.

# 1. INTRODUCTION

Extensions to normal-theory-based discriminant analysis include the use of Box-Cox transformations (McLachlan, 1992, §6.3; Riani & Atkinson, 2001), mixture discriminant analysis (Hastie & Tibshirani, 1996) and general nonparametric discriminant analysis through nonparametric density estimation.

We consider a semiparametric approach in which it is assumed, that after unspecified univariate monotone transformations, the class distributions are multivariate normal. The model is closely related to nonparametric discriminant analysis. The semiparametric specification of the class densities represents a submodel of the nonparametric log density functional analysis of variance model in which the main effects are completely nonparametric but the interaction terms are specified semiparametrically. The model specification and its relationship with other discriminant analysis methods are introduced in §2. An algorithm based on distribution quantiles can be used for model estimation and classification. The procedure is computationally fast and achieves the usual parametric rate for estimation accuracy. The procedure is introduced in §3. In §4, we use simulations to illustrate the efficiency of the classification procedure. Some real examples are shown in §5 and §6 contains a discussion.

## 2. THE SEMIPARAMETRIC MODEL

### 2.1. Notation

Suppose we are given a training set of observations  $\{(x^i, y^i), i = 1, \dots, n\}$ , assumed to be independent realisations of a random pair  $(X, Y)$ . Here  $X = (X_1, \dots, X_d) \in R^d$  is the

observable predictor vector and  $Y \in \{-1, +1\}$  is the class label to be predicted. The classes are to be called the positive class and the negative class. Let  $n_+$  and  $n_-$  be the numbers of positive and negative observations respectively. The goal of classification is to find a classification rule  $\psi : R^d \rightarrow \{-1, +1\}$  with a low expected misclassification rate  $\text{pr}\{\psi(X) \neq Y\}$ . It is well known that the decision-theoretic optimal rule, the Bayes rule, can be expressed as  $\text{sign}\{p_+(x) - 1/2\}$ , where  $p_+(x) = \text{pr}(Y = +1|X = x)$  is the posterior probability of the positive class given the predictor vector  $x$ .

Denote the prior probability of the positive class  $\text{pr}(Y = +1)$  by  $w_+$ , and the density of the positive class by  $f_+$ . The corresponding quantities for the negative class are  $w_-$  and  $f_-$ . By the Bayes formula, we have

$$p_+(x) = \frac{w_+ f_+(x)}{w_+ f_+(x) + w_- f_-(x)}.$$

From this it is easy to see that the Bayes rule is equivalent to

$$\psi_B(x) = \begin{cases} +1 & \text{if } \log f_+(x) - \log f_-(x) + \log \frac{w_+}{w_-} > 0; \\ -1 & \text{otherwise.} \end{cases} \quad (1)$$

## 2.2. The model

**Definition 1.** Let  $V = (V_1, \dots, V_d)$  be a random vector. If there exists a set of univariate strictly monotone functions  $g = (g_1, \dots, g_d)$  such that  $g(V) = (g_1(V_1), \dots, g_d(V_d))$  follows a joint normal distribution with mean  $\mu = (\mu_1, \dots, \mu_d)$ , correlation matrix  $Q = (\rho_{kl})$ , and  $\text{var}\{g_j(V_j)\} = \sigma_j^2$ ,  $k, l = 1, \dots, d$ , then we say  $V$  has a transnormal distribution  $TN(g, \mu, \sigma, Q)$ .

In the above definition  $g_j$ 's are unique only up to scale and shift. The assumption that the  $g_j$ 's are strictly monotone is made to ensure identifiability and easy interpretation. Of course, any continuous univariate random variable has a transnormal representation, since, if  $W$  has cumulative distribution function  $F$ , then  $\Phi^{-1}\{F(W)\} \sim N(0, 1)$ , where  $\Phi$  is the cumulative distribution function of a standard normal random variable. In higher-dimensional spaces, the transnormal distribution imposes a dependence structure among the multiple variates. It can be seen that the transnormal specification is equivalent to the use of normal copulas to describe the relations among random variates; a copula is a continuous joint distribution whose marginals are all uniform on  $(0, 1)$ , and is often used to characterise the dependence structure among multiple random variables (Joe, 1997; Klaassen & Wellner, 1997).

Our semiparametric model assumes that the positive class and the negative class have transnormal distributions  $TN(g, \mu_+, \sigma_+, Q_+)$  and  $TN(g, \mu_-, \sigma_-, Q_-)$ , respectively.

### 2.3. Relationship with nonparametric models

General nonparametric discriminant analysis methods are based on nonparametric density estimation. However, nonparametric density estimation can be very problematic in multivariate problems, because of the curse of dimensionality caused by the sparseness of high-dimensional data. The nonparametric functional analysis of variance model is a popular way of taming the curse of dimensionality. It assumes that the  $d$ -dimensional function to be estimated can be written as

$$\text{constant} + \sum_{j=1}^d \eta_j(x_j) + \sum_{k < l} \eta_{kl}(x_k, x_l) + \dots \quad (2)$$

where the components satisfy side conditions that guarantee uniqueness, and the series is

truncated in some manner. The functional analysis of variance model is easily interpreted when the order of interactions is low. Several authors have applied the functional analysis of variance model in various nonparametric estimation settings; see for example Wahba et al. (1995) and Stone et al. (1997). In the context of nonparametric discriminant analysis, we assume that the log densities of the classes have the decomposition (2).

The nonparametric additive model (Hastie & Tibshirani, 1990) is a special case of (2) with no interaction term. The additive log density model for the class densities implies that, conditional on the classes, the predictor variates are independent, which leads to the so-called nonparametric Naive Bayes method. This assumption of conditional independence is usually unrealistic, but the Naive Bayes method is still very popular among practitioners because of its simplicity and its frequently good performance in practice. However, the Bayes rule (1) depends on the log densities only through the sum of the differences in the log marginal densities, and this simple pooling exercise gives equal weight to the information from each predictor variable. In situations where the predictor variables are highly correlated within classes, this approach is likely to fail. Thus it is desirable to take into account the dependence structure in the class distributions when constructing a classification rule, for example by considering functional analysis of variance models with interaction terms.

One drawback of the functional analysis of variance model with general interaction terms is that the computational load for model fitting is quite heavy, usually of the order of  $n^3$ . Rather than model interaction terms as general functions, one can sometimes assume special forms for the interaction terms. In the regression context, Hastie and Tibshirani (1990, Chapter 9) suggested a regression model of the form

$$\mu + \beta_1 \eta_1(X_1) + \beta_2 \eta_2(X_2) + \gamma \cdot \eta_1(X_1) \cdot \eta_2(X_2),$$

with  $\eta_j$  monotone,  $E\{\eta_j(X_j)\} = 0$ , and  $\text{var}\{\eta_j(X_j)\} = 1$ . Clearly there are many other possibilities.

The transnormal model is a submodel of the general log density functional analysis of variance model in which the interaction is specified semiparametrically. Let  $V$  be a transnormal random variable  $TN(g, \mu, \sigma, Q)$ . Straightforward calculation shows that the log density function of  $V$  is

$$\frac{\log \det(A)}{2} - \frac{d \log(2\pi)}{2} - \frac{1}{2} \sum_{i,j} a_{ij} \frac{g_i(v_i) - \mu_i}{\sigma_i} \frac{g_j(v_j) - \mu_j}{\sigma_j} + \sum_i \log |g'_i(v_i)| - \sum_i \log \sigma_i,$$

where  $Q^{-1} = A = (a_{ij})$ . Later in the paper, the summation notation  $\Sigma$  in front of a vector will mean the sum of all the elements in the vector. Operations between vectors are componentwise, and functions such as log before a vector are also taken to operate componentwise. Therefore, the above log density of  $V$  can be written as

$$\frac{\log \det(A)}{2} - \frac{d \log(2\pi)}{2} - \frac{1}{2} [\{g(v) - \mu\}/\sigma] A [\{g(v) - \mu\}/\sigma]^T + \sum \log |g'(v)| - \sum \log \sigma. \quad (3)$$

This is a submodel of the functional analysis of variance log density model with

$$\eta_j(v_j) = -\frac{1}{2} a_{jj} \{g_j(v_j) - \mu_j\}^2 / (\sigma_j)^2 + \log |g'_j(v_j)|, \quad (4)$$

$$\eta_{kl}(v_k, v_l) = -a_{kl} \{g_k(v_k) - \mu_k\} \{g_l(v_l) - \mu_l\} / (\sigma_k \sigma_l), \quad (5)$$

and there is no interaction term with order higher than two. The constant in the decomposition guarantees that the density integrates to unity. Note that the main effects  $\eta_j$  are general: the log density function of any one-dimensional continuous random variable can be expressed in the form on the right-hand side of (4), essentially because any one dimensional continuous random variable is transnormal.

Since  $g_j(V_j) \sim N(\mu_j, \sigma_j^2)$ , the log of the marginal density function of  $V_j$  is

$$\log f_j(v_j) = -\frac{1}{2} \log(2\pi) - \log(\sigma_j) + \log |g'_j(v_j)| - \frac{\{g_j(v_j) - \mu_j\}^2}{2\sigma_j^2}.$$

From this we get an equivalent form for (3):

$$\sum_j \log f_j(v_j) + \frac{1}{2} \{(g(v) - \mu)/\sigma\} (I - A) \{(g(v) - \mu)/\sigma\}^T + \frac{\log \det(A)}{2}, \quad (6)$$

where  $f_j(v_j)$  is the density function of  $V_j$ . From this it is clear that the transnormal model generalises the additive log density model, which corresponds to only the first term in the above expression, by incorporating two-way interactions. The presence of any two-way interaction is controlled by one number  $a_{ij}$ . The resulting model permits clear interpretations:  $a_{ij} = 0$  if and only if  $V_i$  and  $V_j$  are independent conditional on all the other random variates. This is seen from (5).

### 3. THE CLASSIFICATION PROCEDURE

#### 3.1. Notation

Let  $X_+ = (X_{+1}, \dots, X_{+d}) = \{X|Y = +1\}$  and  $X_- = \{X|Y = -1\}$  denote the generic random vectors corresponding to the positive class distribution and the negative class distribution, respectively. We assume the  $X_+ \sim TN(g, \mu_+, \sigma_+, Q_+)$  and  $X_- \sim TN(g, \mu_-, \sigma_-, Q_-)$ . Since the transformations  $g_j$ ,  $j = 1, \dots, d$ , are unique only up to scale and shift, without loss of generality we can set  $\mu_+ = (0, 0, \dots, 0)$ ,  $\sigma_+ = (1, 1, \dots, 1)$ , and assume that all  $g_j$  are strictly increasing functions. We will consider three different cases, each corresponding to a set of assumptions about the correlation matrices  $Q$  and the variances  $\sigma$ .

Case 1.  $Q_+ = Q_-$  and  $\sigma_+ = \sigma_-$ .

Case 2.  $Q_+ = Q_-$  but  $\sigma_+$  and  $\sigma_-$  may be different.

Case 3. Both the correlation matrices and the variances may be different.

In Case 1, linear discriminant analysis is applicable after the transformation. In Case 3, quadratic discriminant analysis is applicable after the transformation. Case 2 in intermediate represents a compromise between Cases 1 and 3.

For any  $j \in \{1, \dots, d\}$ , denote the cumulative distribution function of  $X_{+j}$  by  $F_{+j}$  and the cumulative distribution function of  $X_{-j}$  by  $F_{-j}$ . Since  $g_j(X_{+j}) \sim N(0, 1)$ , and  $g_j(X_{-j}) \sim N(\mu_{-j}, \sigma_{-j}^2)$ , we have that

$$g_j = \Phi^{-1} \circ F_{+j} = (\Phi^{-1} \circ F_{-j})\sigma_{-j} + \mu_{-j}, \quad (7)$$

where  $\circ$  denotes the composition of functions. The basic idea of the estimation procedure is to replace  $F_{+j}$  and  $F_{-j}$  by their empirical versions.

Let  $x_+^i$ ,  $i = 1, \dots, n_+$ , be the positive observations, let the empirical cumulative distribution of  $X_{+j}$  based on  $x_+^i$ ,  $i = 1, \dots, n_+$ , be denoted by  $\bar{F}_{+j}$ , and define  $\tilde{F}_{+j} = \{n_+/(n_+ + 1)\}\bar{F}_{+j}$ . Let  $x_-^i$ ,  $\bar{F}_{-j}$  and  $\tilde{F}_{-j}$  be defined similarly for the negative class.

### 3.2. Estimation of the correlation matrices

We consider the estimation of  $Q_+ = (\rho_{+jk})$ . The estimation of  $Q_-$  is similar. It is clear that  $\rho_{+jj} = 1$ ,  $j = 1, \dots, d$ . For  $j \neq k$ ,

$$\rho_{+jk} = \text{corr}\{g_j(X_{+j}), g_k(X_{+k})\} = \text{corr}\{\Phi^{-1} \circ F_{+j}(X_{+j}), \Phi^{-1} \circ F_{+k}(X_{+k})\}.$$

is the correlation between two standard normal random variables. If we knew  $F_{+j}$  and  $F_{+k}$ , then  $\rho_{+jk}$  could be estimated by

$$\frac{\frac{1}{n_+} \sum_{i=1}^{n_+} \{\Phi^{-1} \circ F_{+j}(x_+^i)\} \{\Phi^{-1} \circ F_{+k}(x_+^i)\}}{[\frac{1}{n_+} \sum_{i=1}^{n_+} \{\Phi^{-1} \circ F_{+j}(x_+^i)\}^2]^{1/2} [\frac{1}{n_+} \sum_{i=1}^{n_+} \{\Phi^{-1} \circ F_{+k}(x_+^i)\}^2]^{1/2}}.$$

Replacing  $F_{+j}$  and  $F_{+k}$  in the above formula by  $\tilde{F}_{+j}$  and  $\tilde{F}_{+k}$ , we obtain the estimator  $\hat{\rho}_{+jk}$  of  $\rho_{+jk}$ . This estimator is exactly the Van der Waerden normal score rank correlation coefficient. Klaassen & Wellner (1997, §3) showed that  $\hat{\rho}_{+jk}$  is a semiparametrically efficient estimator of  $\rho_{+jk}$  in the transnormal specification, i.e. the normal copula model:

$$n^{1/2}(\hat{\rho}_{+jk} - \rho_{+jk}) \rightarrow N\{0, (1 - \rho_{+jk}^2)^2\},$$

in distribution as  $n \rightarrow \infty$ . The asymptotic variance  $(1 - \rho_{+jk}^2)^2$  is the same as the asymptotic variance of the usual sample correlation coefficient in the case of normal distributions.

In situations where we assume  $Q_+ = Q_- = Q$ , we can simply pool the estimators for  $Q_+$  and  $Q_-$ , giving  $\hat{Q} = (n_+ \hat{Q}_+ + n_- \hat{Q}_-)/n$ , where  $\hat{Q}_+ = (\hat{\rho}_{+jk})$  and  $\hat{Q}_- = (\hat{\rho}_{-jk})$ .

### 3.3. Estimation of the mean and variance of the transformed negative class

We first consider estimation in Case 1. In such a situation  $\sigma_+ = \sigma_- = (1, 1, \dots, 1)$ , and we only need to estimate  $\mu_{-j}$ ,  $j = 1, \dots, d$ . In this subsection, each dimension  $j \in \{1, \dots, d\}$  is treated separately. Therefore, we can suppress the subscript  $j$  in the discussion to ease the notation. Thus in this subsection, for example,  $F_+$  refers to  $F_{+j}$ ,  $\tilde{F}_+$  refers to  $\tilde{F}_{+j}$ ,  $g$  refers to  $g_j$ ,  $\mu_-$  refers to  $\mu_{-j}$  and  $X_+$  refers to  $X_{+j}$ . All quantities in this subsection are one-dimensional.

Fix any two real numbers  $a < b$ . The notation  $h^{a,b}$  stands for the truncation of a function  $h$  at  $a$  and  $b$ . That is,

$$h^{a,b}(x) = \begin{cases} h(x) & \text{if } h(x) \in (a, b); \\ a & \text{if } h(x) \leq a; \\ b & \text{if } h(x) \geq b. \end{cases}$$

Since  $g = \Phi^{-1} \circ F_+$ , we have  $g^{a,b} = \Phi^{-1} \circ F_+^{\alpha,\beta}$ , where  $\alpha = \Phi(a)$  and  $\beta = \Phi(b)$ . Let  $g_{n_+} = \Phi^{-1} \circ \tilde{F}_+$ . Then  $g_{n_+}^{a,b} = \Phi^{-1} \circ \tilde{F}_+^{\alpha,\beta}$  is expected to be close to  $g^{a,b}$ . In fact, it is well known (Dvoretzky et al., 1956) that

$$\sup_{x \in R} |\tilde{F}_+(x) - F_+(x)| = O_p(n_+^{-1/2}).$$

Therefore,

$$\begin{aligned} \sup_{x \in R} |\tilde{F}_+^{\alpha,\beta}(x) - F_+^{\alpha,\beta}(x)| &= O_p(n_+^{-1/2}), \\ \sup_{x \in R} |g_{n_+}^{a,b}(x) - g^{a,b}(x)| &= O_p(n_+^{-1/2}). \end{aligned} \quad (8)$$

Recall that  $g(X_-) \sim N(\mu_-, 1)$ . Therefore the distribution of  $g^{a,b}(X_-)$  is  $N(\mu_-, 1)$  truncated to the interval  $[a, b]$ . If we knew the function  $g^{a,b}$ , we could apply the method of moments or maximum likelihood estimation to the data  $\{g^{a,b}(x_-^i)\}$ ,  $i = 1, 2, \dots, n_-$ , to estimate  $\mu_-$ . We do not know  $g^{a,b}$ , but we can approximate  $g^{a,b}$  by  $g_{n_+}^{a,b}$  in the estimation method of choice. This is the basic idea behind our estimation.

Let  $\tilde{F}_+^{-1}(\alpha) = \inf\{x : \tilde{F}_+ \geq \alpha\}$ . Denote the probability density function of the standard normal distribution by  $\phi$ . Write  $q = 1/n_- \sum_{i=1}^{n_-} 1_{g_{n_+}(X_-^i) \in (a,b)}$ . Our estimator of  $\mu_-$  is

$$\hat{\mu}_- = q^{-1} \left[ \frac{1}{n_-} \sum_{i=1}^{n_-} g_{n_+}(X_-^i) 1_{g_{n_+}(X_-^i) \in (a,b)} + \phi\{\Phi^{-1} \circ \tilde{F}_- \circ \tilde{F}_+^{-1}(\beta)\} - \phi\{\Phi^{-1} \circ \tilde{F}_- \circ \tilde{F}_+^{-1}(\alpha)\} \right]. \quad (9)$$

The proof of the following theorem is given in Appendix 1.

**Theorem 1.** *Assume that  $|g'| < M$  for some positive number  $M$ . Then, for any two real numbers  $a < b$ ,  $\hat{\mu}_- = \mu_- + O_p(n_+^{-1/2} + n_-^{-1/2})$ .*

Asymptotically, the above estimators work for any given  $a < b$ . In practice, we need to select a reasonable pair  $a$  and  $b$ , or equivalently,  $\alpha$  and  $\beta$ . For robustness considerations, we

choose  $\alpha$  and  $\beta$  so that  $\alpha, \beta, \tilde{F}_-^{-1}\{\tilde{F}_+^{-1}(\alpha)\}$  and  $\tilde{F}_-^{-1}\{\tilde{F}_+^{-1}(\beta)\}$  are all between 2.5% and 97.5%. In our implementations, we always take the class with larger sample size as the positive class. Intuitively, this should enhance the estimation accuracy when the two sample sizes are very different.

The estimator  $\hat{\mu}_-$  is essentially an approximate method-of-moments estimator after some simplification. The method-of-moments estimator of  $\mu_-$  based on  $\{g^{a,b}(x_-^i)\}$ ,  $i = 1, 2, \dots, n_-$ , is approximated by replacing the unobservable  $\{g^{a,b}(x_-^i)\}$ ,  $i = 1, 2, \dots, n_-$ , with  $\{g_{n_+}^{a,b}(x_-^i)\}$ ,  $i = 1, 2, \dots, n_-$ . The estimator has the form of an approximate trimmed-mean estimator of the normal mean  $\mu_-$ . It is easily calculated and has good robustness properties. An alternative approach is to approximate the maximum likelihood estimator based on  $\{g^{a,b}(X_-^i)\}$ ,  $i = 1, 2, \dots, n_-$ . The resulting approximate maximum likelihood estimator does not have a nice explicit formula, but can be computed numerically. The rate of convergence is still  $O_p(n_+^{-1/2} + n_-^{-1/2})$ . A derivation of this is sketched in Appendix 2.

Now we briefly discuss Cases 2 and 3, in which the variances are not assumed to be the same. Equation (8) is still valid, and we have  $g(X_-) \sim N(\mu_-, \sigma_-)$ . Approximate method-of-moments estimators or approximate maximum likelihood estimators can then be used to estimate  $\mu_-$  and  $\sigma_-$  simultaneously.

### 3.4. Classifying a future observation

Once again we consider Case 1. The procedures for Cases 2 and 3 are derived in similar fashion. By (3) we can see that the Bayes rule (1) for Case 1 amounts to the sign of the log odds

$$\log(w_+/w_-) - \frac{1}{2}\{g(x)\}Q^{-1}\{g(x)\}^T + \frac{1}{2}\{g(x) - \mu_-\}Q^{-1}\{g(x) - \mu_-\}^T. \quad (10)$$

The first term can be estimated by  $\log(n_+/n_-)$ . After we obtain the estimates  $\hat{Q}$  and  $\hat{\mu}_-$ , all we need to classify a new observation with predictor vector  $x = (x_1, \dots, x_d)$  is to find  $g(x) = (g_1(x_1), \dots, g_d(x_d))$ . One possibility is to use  $\Phi^{-1} \circ \tilde{F}_{+j}$  to approximate  $g_j = \Phi^{-1} \circ F_{+j}$ . This amounts to using the rank of  $x_j$  among the  $x_{+j}^i$ , or equivalently the rank of  $g_j(x_j)$  among  $g_j(x_{+j}^i)$ ,  $i = 1, \dots, n_+$ , to locate  $g_j(x_j)$ . Our implementation instead attempts to use the rank of  $g_j(x_j)$  among  $g_j(x_j^i)$ ,  $i = 1, \dots, n$ , to locate  $g_j(x_j)$ : we know that the distribution of  $g_j(X_j)$  is a mixture of normals  $H_j(\cdot) = w_+ \Phi(\cdot) + w_- \Phi(\cdot - \mu_-)$ . Let  $x_j^{(k)}$  be the  $k$ th smallest among  $x_j^i$ ,  $i = 1, \dots, n$ . Then  $g_j(x_j^{(k)})$  is the  $k$ th smallest among  $g_j(x_j^i)$ ,  $i = 1, \dots, n$ , and  $g_j(x_j^{(k)})$  should be close to the probability quantile  $H_j^{-1}\{(k - 0.5)/n\}$ . This is the same idea as is used in the construction of the probability quantile plot. We use  $\hat{H}_j = (n_+/n)\Phi(\cdot) + (n_-/n)\Phi(\cdot - \hat{\mu}_-)$  in place of  $H_j$  in the calculation of the probability quantile. If  $x_j$  is between  $x_j^{(k)}$  and  $x_j^{(k+1)}$ , then  $g_j(x_j)$  is between  $g_j(x_j^{(k)})$  and  $g_j(x_j^{(k+1)})$ , and we approximate  $g_j(x_j)$  by  $\hat{H}_j^{-1}\{(k+0.5)/(n+1)\}$  because  $g_j(x_j)$  is the  $(k+1)$ th smallest among  $(n+1)$  observations. It is clear that the approximation is of the order  $O_p(n_+^{-1/2} + n_-^{-1/2})$ . The approximated  $g_j(x_j)$  is then plugged into (10) to create a classification rule.

### 3.5. Further considerations

Any ties between observed values are broken at random in our algorithm. An alternative would be to use tied ranks.

We estimate the  $Q$ 's and the  $\mu_-$ 's separately. This is straightforward if the transnormal model specification is satisfied, but the estimation may be sensitive to deviations from the transnormal specification. For robustness considerations we can instead estimate  $\mu_-$  and  $\sigma_-$  as described in §3.3, approximate  $g_j(x_j^i)$ ,  $i = 1, \dots, n$ , by probability quantiles as described

in §3.4, and estimate  $Q_+$  and  $Q_-$  from the approximated  $g(x_+^i)$ ,  $i = 1, \dots, n_+$ , and  $g(x_-^i)$ ,  $i = 1, \dots, n_-$ , as we know  $g(X_+)$  has normal distribution with correlation matrix  $Q_+$ , and  $g(X_-)$  has normal distribution with correlation matrix  $Q_-$ . This is essentially equivalent to applying linear discriminant analysis or quadratic discriminant analysis to the approximated  $g(x^i)$ ,  $i = 1, \dots, n$ . We use this procedure as the default in our implementation.

In our implementation we used an alternative version of  $\tilde{F}_{+j}$ . Let  $x_{+j}^{(k)}$ ,  $k = 1, 2, \dots, n_+$  be the order statistic of the observations  $x_{+j}^i$ ,  $i = 1, 2, \dots, n_+$ . Also,  $\tilde{F}_{+j}(x)$  is 0 if  $x$  is smaller than  $x_{+j}^{(1)}$ , 1 if  $x$  is larger than  $x_{+j}^{(n_+)}$ , and linearly connects points  $(x_{+j}^{(k)}, (k - 0.5)/n_+)$  and  $(x_{+j}^{(k+1)}, (k + 0.5)/n_+)$ , for  $k = 1, 2, \dots, n_+$ . This is more continuous than the original version and is computationally more stable. Theorem 1 and the similar result for the approximate maximum likelihood estimator are still valid, with proofs unchanged, for this new version of  $\tilde{F}_{+j}$ . Details of the algorithm are given in Appendix 3.

## 4. SIMULATIONS

To illustrate the transformation discriminant analysis procedure we consider a problem with dimensionality seven. The positive sample and the negative sample are generated from different normal distributions with identical covariance structure. Therefore linear discriminant analysis is the optimal method for this example, while quadratic discriminant analysis should also produce reasonable results. The purpose of the simulation is to see how our procedures compare with the optimal procedure. Let  $(Z_1, \dots, Z_7)$  be independent standard normal variables. The distribution of  $X_+$  is defined through  $X_{+1} = Z_1$ ;  $X_{+2} = X_{+1} + Z_2$ ;  $X_{+3} = X_{+2} + Z_3/2$ ;  $X_{+4} = X_{+3} + Z_4/3$ ;  $X_{+5} = Z_5$ ;  $X_{+6} = Z_6$ ;  $X_{+7} = X_{+5} + X_{+6} + Z_7$ . The negative class has a normal distribution with the same covariance

matrix and mean  $(1/7, 2/7, \dots, 7/7)$ . We consider two training sample sizes  $n = 1000$  and  $n = 100$ . For each sample size, we generate two-thirds of the training set from the positive class and the rest from the negative class. We generate the test set in the same way. The test set size is 1000 for both training sample sizes. We applied the transformation discriminant analysis procedures, as well as linear discriminant analysis, quadratic discriminant analysis and the Naive Bayes method. For each transformation discriminant analysis procedure, we report the results for the procedure with both the approximate method-of-moments estimator (9) and the approximate maximum likelihood estimator. These are denoted by TDA.mm and TDA.ml. We ran the simulation 100 times and the results are displayed in Fig 1. The mean misclassification rates based on the 100 simulations are given in Table 1.

The transformation discriminant analysis methods give very similar performance to linear discriminant analysis, which is the optimal procedure in this situation. The Naive Bayes method cannot incorporate the dependence structure of the variates, and gives poor results. In our implementation of the Naive Bayes method, the univariate densities are estimated using the kernel density estimator with automatic bandwidth selection implemented in the R library KernSmooth.

The TDA.mm and TDA.ml procedures perform similarly, although the procedures TDA.mm are computationally faster. In the TDA.ml procedures, the approximate maximum likelihood estimation requires numerical maximisation. We used the R optimisation functions ‘optimize’ for the univariate optimisation in TDA1.ml and ‘nlm’ for the bivariate optimisation in TDA2.ml and TDA3.ml. The function ‘nlm’ sometimes gives warnings; in our experiments these warnings have never caused any problem in terms of classification accuracy.

We can easily construct examples in which our procedures beat linear discriminant analysis, since if we apply monotone transformation to each direction our procedure will have exactly the same performance, but linear discriminant analysis perform poorly when we force the optimal decision boundary to be far from linear.

## 5. EXAMPLES

In a comprehensive study of the accuracy of classification methods, Lim et al. (2000) compared thirty-three classification procedures on thirty-two datasets. There were sixteen original datasets but, to study the effect of uninformative predictor variates, they added noise variables to each of those datasets, yielding a total of thirty-two. The classification methods include twenty-two decision-tree algorithms, nine classical and modern statistical algorithms and two neural networks. The POLYCLASS algorithm (Koopberg, 1997) performed best in term of overall accuracy in the study of Lim et al. (2000). This algorithm is related to functional analysis of variance modelling. Another algorithm considered in the study that is related to functional analysis of variance modelling is Flexible Discriminant Analysis (Hastie et al., 1994), denoted by FM2 in Lim et al. (2000). Fourteen of the datasets are from real-life applications and two are artificially constructed.

Of the thirty-two datasets, six satisfy the condition that there are only two classes and all the predictor variables are continuous. These datasets are available at the Repository of Machine Learning Databases at University of California, Irvine. The URL is [www.ics.uci.edu/~mllearn/MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html).

*Example 1.* **Wisconsin breast cancer** (bcw) and its noisy version (bcw+). This dataset was contributed by W. H. Wolberg. The objective is to predict whether a tissue

sample taken from a patient’s breast is benign or malignant. There are nine continuous predictor variables. Following Lim et al. (2000), we delete the sixteen observations that contain missing values, leaving a sample size of 683. The noisy version `bcw+` is constructed by adding nine independent noise predictor variables, each with a uniform distribution over the integers 1 to 10 inclusive.

*Example 2. **Bupa liver disorder** (bld) and its noisy version (bld+).* This dataset was contributed by R. S. Forsyth. The objective is to predict whether or not a male patient has a liver disorder. The sample size is 345. There are six continuous predictor variables. The dataset `bld+` is created by adding nine independent noise predictor variables, each with a standard normal distribution.

*Example 3. **Pima Indian diabetes** (pid) and its noisy version (pid+).* This dataset was contributed by V. Sigillito. The objective is to predict whether or not a patient would test positive for diabetes given a number of physiological measurements and medical test results. Following Lim et al. (2000), we removed the variable serum insulin, which contains many zero values that are physically impossible, and records that have impossible values in other variables. The resulting dataset has 532 records and 7 predictor variables. The dataset `pid+` is created by adding eight independent noise predictor variables, each with a standard normal distribution.

We apply transformation discriminant analysis as well as linear discriminant analysis, quadratic discriminant analysis and the Naive Bayes method to the six datasets. We repeat ten-fold crossvalidation 100 times and report the average. The results are summarised in Table 2. For comparison, we also include in the table the error rates of Flexible Discriminant Analysis (FDA) and the POLYCLASS algorithm (POL), as well as the minimum and

maximum error rate achieved by the thirty-three classification methods as reported by Lim et al. (2000). The error rates in Lim et al. (2000) were estimated using ten-fold crossvalidation. When the nonparametric Naive Bayes method is used on the datasets `bcw` and `bcw+`, we encountered some numerical problems. Therefore, for these two datasets we have used parametric normal density estimation instead of nonparametric kernel density estimation in the Naive Bayes procedure. Transformation discriminant analysis methods perform quite well. The results reported here are based on the `TDA.mm` procedures. The `TDA.ml` procedures perform similarly, and actually perform slightly better in these examples.

## 6. DISCUSSION

The computation complexity of the algorithm is roughly linear in the sample size  $n$ . In fact, the most significant piece of computation is the sorting of the observations in each dimension, which has a linear expected computation time and worst-case computation time of order  $O(n \log n)$ . This compares favourably with most nonparametric discriminant analysis procedures.

As discussed in §2.3, the assumption of transnormality amounts to a second-order functional analysis of variance log density model in which the main effects are general, but each two-way interaction is specified semiparametrically. This assumption is much more general than the assumptions made in parametric methods or the nonparametric Naive Bayes method. However, it clearly imposes some restriction on the class density which may not be satisfied in practice. Furthermore, the assumption in our classification algorithm is actually a bit stronger than the transnormality of the class distributions: we require that the two classes be transformed to multivariate normal by the same set of transformations. Our

algorithm makes checking the model assumption feasible: under the model assumption, we effectively transformed the two classes into multivariate normals in Step 2 of our algorithm, and therefore the usual tests for multivariate normality (McLachlan, 1992, Ch. 6; Andrews et al., 1973; Fatti et al., 1982; Hawkins, 1981) can be applied to the transformed data.

An alternative to checking the model assumption is to check the quality of the resulting classification rule directly by looking at its classification performance. In classification problems it may happen that the classification rule developed under certain assumptions works reasonably well in situations where the assumptions are violated (McLachlan, 1992, Ch. 6). This is an approach widely used in the machine learning literature, and is employed in this paper.

The transformation discriminant analysis procedure described here applies to binary classification. Similar ideas can be applied to the multi-class problems, but it is probably easier to deal with the multi-class problems by applying the the binary transformation discriminant analysis to all pairs of classes and then combining the results. This is a general approach to solve multi-class problems with binary classifiers; see Hastie & Tibshirani (1998) and an unpublished Stanford University technical report by J. H. Friedman. How to best combine our binary procedure to treat multi-class problems is under study.

## ACKNOWLEDGEMENT

The research was partly supported by the Wisconsin Alumni Research Foundation. The work of the second author was partly supported by the US National Science Foundation through a grant with Grace Wahba. The authors are grateful to the editor, associate editor and referee for their helpful comments and very constructive suggestions.

## APPENDIX 1

### Proof of Theorem 1

The notation in this appendix follows the same conventions as those in §3.3.

Since  $g(X_-) \sim N(\mu_-, 1)$ , we have

$$E\{g(X_-)1_{g(X_-) \in (a,b)}\} = \mu_- \{\Phi(b - \mu_-) - \Phi(a - \mu_-)\} + \phi(a - \mu_-) - \phi(b - \mu_-).$$

By the Central Limit Theorem we have

$$\begin{aligned} & \left( \frac{1}{n_-} \sum_{i=1}^{n_-} g(X_-^i) 1_{g(X_-^i) \in (a,b)} \right) + \phi(b - \mu_-) - \phi(a - \mu_-) \\ &= \mu_- \{\Phi(b - \mu_-) - \Phi(a - \mu_-)\} + O_p(n_-^{-1/2}), \end{aligned} \quad (\text{A1})$$

$$\frac{1}{n_-} \sum_{i=1}^{n_-} 1_{g(X_-^i) \in (a,b)} = \Phi(b - \mu_-) - \Phi(a - \mu_-) + O_p(n_-^{-1/2}). \quad (\text{A2})$$

On the other hand, since  $g(X_-) \sim N(\mu_-, 1)$ , we have

$$\Phi(a - \mu_-) = \text{pr}\{g(X_-) \leq a\} = \text{pr}\{F_+(X_-) \leq \alpha\} = F_- \circ F_+^{-1}(\alpha). \quad (\text{A3})$$

The distance between  $\tilde{F}_-$  and  $F_-$  is of order  $O_p(n_-^{-1/2})$  in the  $L_\infty$  norm, and  $\tilde{F}_+^{-1}(\alpha) - F_+^{-1}(\alpha) = O_p(n_+^{-1/2})$ ; see Serfling (1980, p. 75). Therefore

$$F_- \{F_+^{-1}(\alpha)\} - \tilde{F}_- \{\tilde{F}_+^{-1}(\alpha)\} \quad (\text{A4})$$

$$= F_- \{F_+^{-1}(\alpha)\} - F_- \{\tilde{F}_+^{-1}(\alpha)\} + F_- \{\tilde{F}_+^{-1}(\alpha)\} - \tilde{F}_- \{\tilde{F}_+^{-1}(\alpha)\} \quad (\text{A5})$$

$$= O_p(n_+^{-1/2} + n_-^{-1/2}) \quad (\text{A6})$$

The last equality (A6) uses the assumption that  $g'$  is bounded from infinity, and therefore  $F'_-$  is bounded from infinity by (7). By (A3) and (A6) we obtain  $\phi(a - \mu_-) = \phi(\Phi^{-1}[F_- \{\tilde{F}_+^{-1}(\alpha)\}]) + O_p(n_+^{-1/2} + n_-^{-1/2})$ . Similarly we can obtain

$\phi(b - \mu_-) = \phi(\Phi^{-1}[\tilde{F}_-^{-1}\{\tilde{F}_+^{-1}(\beta)\}]) + O_p(n_+^{-1/2} + n_-^{-1/2})$ . Combining this with (8), (A1) and (A2), the result follows.

## APPENDIX 2

### Rate of convergence of the approximate maximum likelihood estimator

The notation in this appendix follow the same conventions as those in §3.3.

Now we consider the approximate maximum likelihood estimator. Since  $g(X_-) \sim N(\mu_-, 1)$ , the scaled log likelihood for  $\mu_-$  based on  $g^{a,b}(X_-^i)$  is

$$\begin{aligned} & \frac{1}{n_-} \sum_{i=1}^{n_-} \left[ \log \phi\{g(X_-^i) - \mu_-\} 1_{a < g(X_-^i) < b} + \log \Phi(a - \mu_-) 1_{g(X_-^i) \leq a} + \log\{1 - \Phi(b - \mu_-)\} 1_{g(X_-^i) \geq b} \right] \\ &= C - \Lambda_1 (Z - \mu_-)^2 / 2 + \Lambda_2 \log \Phi(a - \mu_-) + \Lambda_3 \log \Phi(\mu_- - b) \\ &= L(\mu_-; Z, \Lambda_1, \Lambda_2, \Lambda_3), \end{aligned}$$

where  $C$  does not depend on  $\mu_-$ , and  $\Lambda_1 = (1/n_-) \sum_{i=1}^{n_-} 1_{a < g(X_-^i) < b}$ ,  $\Lambda_2 = (1/n_-) \sum_{i=1}^{n_-} 1_{g(X_-^i) \leq a}$ ,  $\Lambda_3 = (1/n_-) \sum_{i=1}^{n_-} 1_{g(X_-^i) \geq b}$  and  $Z = \sum_{i=1}^{n_-} g(X_-^i) 1_{a < g(X_-^i) < b} / (\Lambda_1 n_-)$ . Let  $W = (Z, \Lambda_1, \Lambda_2, \Lambda_3)$ .

Since  $\log \Phi(t)$  is strictly concave, we see that  $L$  is strictly concave in  $\mu_-$ . The maximum likelihood estimator  $\bar{\mu}_- = h(W)$  of  $\mu_-$  based on  $g^{a,b}(X_-^i)$  is the solution of the likelihood equation  $\partial L(\mu_-; W) / \partial \mu_- = 0$ . By standard arguments for maximum likelihood estimation we have that

$$\bar{\mu}_- = \mu_- + O_p(n_-^{-1/2}). \tag{A7}$$

By the definition of  $h$ , we have that

$$\partial h / \partial W = - \frac{\partial^2 L}{\partial \mu_- \partial W} / \frac{\partial^2 L}{\partial \mu_-^2}. \tag{A8}$$

Define  $\tilde{\Lambda}_1, \tilde{\Lambda}_2, \tilde{\Lambda}_3, \tilde{Z}$  and  $\tilde{W}$  analogously to  $\Lambda_1, \Lambda_2, \Lambda_3, Z$  and  $W$ , with  $g_{n_+}$  replacing  $g$ . The approximate maximum likelihood estimator is  $\tilde{\mu}_- = h(\tilde{Z}, \tilde{\Lambda}_1, \tilde{\Lambda}_2, \tilde{\Lambda}_3)$ , the solution of  $\partial L(\mu_-; \tilde{Z}, \tilde{\Lambda}_1, \tilde{\Lambda}_2, \tilde{\Lambda}_3) / \partial \mu_- = 0$ .

From (A8), direct calculation shows that  $\partial h / \partial W$  is bounded in probability. From (8) we have that  $|\tilde{\Lambda}_i - \Lambda_i| = O_p(n_+^{-1/2} + n_-^{-1/2})$ ,  $i = 1, 2, 3$ ;  $|\tilde{Z} - Z| = O_p(n_+^{-1/2} + n_-^{-1/2})$ .

Therefore,

$$\tilde{\mu}_- - \bar{\mu}_- = h(\tilde{W}) - h(W) = O_p(n_+^{-1/2} + n_-^{-1/2}).$$

Combining this and (A7) gives  $\tilde{\mu}_- - \mu_- = O_p(n_+^{-1/2} + n_-^{-1/2})$ .

## APPENDIX 3

### Details of the algorithm

The algorithm takes as inputs the training set  $x^i$ ,  $i = 1, \dots, n$ , the class labels of the training set, and the test set  $x_0^\ell$ ,  $\ell = 1, \dots, n_0$ . The output is the class labels of the test set. The R code of the implementation and sample usage are available from the web pages of the authors, along with a description of the forms of the input and output. We implement three versions of the procedure corresponding to the three different Cases 1-3. The three versions are to be called TDA1, TDA2 and TDA3, where TDA stands for transformation discriminant analysis. Two options are available for each of TDA1, TDA2 and TDA3, corresponding to the two different ways of estimating  $\mu_-$  and  $\sigma_-$ , namely the method of moments or maximum likelihood. We only describe the method of moments option. The maximum likelihood option is similar, but does not allow explicit formulae in Step 1.4 below.

The algorithm in its simplest form has three steps. In Step 1, we estimate  $\mu_{-j}$  and  $\sigma_{-j}$  for each dimension  $j = 1, \dots, d$ . In Step 2, we transform the training set and the test set

according to the estimated  $\mu_{-j}$  and  $\sigma_{-j}$ . In Step 3, we estimate the correlation matrices  $Q$ 's and apply linear discriminant analysis, in version TDA1, or quadratic discriminant analysis, in versions TDA2 and TDA3, to the data transformed in Step 2. In the first two steps, the estimation and transformation in each dimension  $j$  is done separately, and therefore we can suppress the notation  $j$  for dimension. All the quantities in the first two steps are one-dimensional.

*Step 1 : Estimating  $\mu_{-}$  and  $\sigma_{-}$ .* The  $j$ th coordinate of the  $i$ th positive example is denoted by  $x_{+}^i$ , and that of the  $i$ th negative example by  $x_{-}^i$ . Let  $r_{+}^i$  be the rank of  $x_{+}^i$  among all the positive examples, and let  $R^i$  be the overall rank of  $x^i$  among the combined set of both positive and negative examples. We assume that the ranges of the positive examples and that of the negative examples overlap. Otherwise the two classes are completely separable in a single dimension, and we can simply use this dimension for classification.

*Step 1.1.* For each positive example  $x_{+}^i$ ,  $i = 1, \dots, n_{+}$ , compute its normal score  $N_{+}^i$  among the positive examples:

$$N_{+}^i = \Phi^{-1}\left(\frac{r_{+}^i - 0.5}{n_{+}}\right).$$

These are approximations for the transformed positive examples.

*Step 1.2.* For a fixed proportion  $p$ , such as 2.5%, find the  $p$ th quantile and the  $(1-p)$ th quantile of the positive examples  $x_{+}^i$ 's. Denote these by  $l_{+}$  and  $u_{+}$ . Similarly find  $l_{-}$  and  $u_{-}$  for the negative class. Let  $[l, u]$  be the intersection of the two intervals  $[l_{+}, u_{+}]$  and  $[l_{-}, u_{-}]$ . If the intersection is empty, adjust the value of  $p$  so that there is an intersection.

*Step 1.3.* For each negative example  $x_{-}^i \in [l, u]$ , find the positive examples  $x_{+}^L$  and  $x_{+}^R$  closest to it from below and from above. Let  $R_{+}^L$ ,  $R_{+}^R$  and  $R_{-}^i$  be the overall ranks corresponding to  $x_{+}^L$ ,  $x_{+}^R$  and  $x_{-}^i$  respectively among the combined set of both positive and

negative examples. Calculate the transformed value of  $x_-^i$  by linearly interpolating between the transformed values of  $x_+^L$  and  $x_+^R$ , see Step 1.1:

$$g_-^i = N_+^L + (N_+^R - N_+^L)(R_-^i - R_+^L)/(R_+^R - R_+^L).$$

Transform  $l$  and  $u$  in the same way, and denote the transformed values by  $a$  and  $b$ . Let the average of the  $g_-^i$ 's for  $x_-^i \in [l, u]$  be  $\bar{g}_-$  and the variance be  $V_{g_-}$ .

*Step 1.4.* Denote the proportions of the negative examples in  $(-\infty, l)$  and  $(u, \infty)$  by  $\alpha$  and  $1 - \beta$ , respectively. These are the same as the proportions of the negative examples below  $a$  or above  $b$  after transformation. Since the negative class distribution is  $N(\mu_-, \sigma_-^2)$  after transformation, we can now use the method of moments to estimate  $\mu_-$  and  $\sigma_-$  based on  $\alpha$ ,  $\beta$ ,  $\bar{g}_-$ ,  $V_{g_-}$ ,  $a$  and  $b$ : in version TDA1  $\sigma_- = 1$  and  $\hat{\mu}_- = \bar{g}_- + \{\phi(b) - \phi(a)\}/(\beta - \alpha)$ ; and in versions TDA2 and TDA3 the estimators are  $\hat{\sigma}_-^2 = V_{g_-}/[\{\beta - \alpha - b\phi(b) + a\phi(a)\}/(\beta - \alpha)] - \{\phi(b) - \phi(a)\}/(\beta - \alpha)^2$  and  $\hat{\mu}_- = \bar{g}_- + \hat{\sigma}_- \{\phi(b) - \phi(a)\}/(\beta - \alpha)$ .

*Step 2 :* Transform any examples  $x^i$ , positive or negative, to the  $\{(R^i - 0.5)/n\}$ th quantile of the mixture normal distribution  $w_+N(0, 1) + w_-N(\hat{\mu}_-, \hat{\sigma}_-)$ . Here  $R^i$  is the overall rank of  $x^i$  among all the training examples, positive or negative. Let  $R$  be the overall rank of a test observation when it is added to the training set. Compute the corresponding transformed value, which is the  $\{(R - 0.5)/(n + 1)\}$ th quantile of the mixture normal distribution.

*Step 3 :* Apply linear discriminant analysis, in version TDA1, or quadratic discriminant analysis, in version TDA3, to the transformed data from Step 2. Version TDA2 is in between

versions TDA1 and TDA3 in that it pools the estimates of the correlation matrices, but allows different  $\sigma_+$  and  $\sigma_-$ .

## REFERENCES

- ANDREWS, D. F., GNANADESIKAN, R. , & WARNER, J. L. (1973). Methods for assessing multivariate normality. In *Multivariate Analysis* (Vol. III), Ed. P. R. Krishnaiah, pp. 95–116. New York: Academic Press.
- BOX, G. E. P. & COX, D. R. (1964). An analysis of transformations (with Discussion). *J. R. Statist. Soc. B* **26**, 211–52.
- DVORETZKY, A., KIEFER, J. & WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27**, 642–69.
- FATTI, L. P., HAWKINS, D. M. & RAATH, E. L. (1982). Discriminant analysis. In *Topics in Applied Multivariate Analysis*, Ed. D. M. Hawkins, pp. 1–71. Cambridge: Cambridge University Press.
- HASTIE, T. & TIBSHIRANI, R. (1990). *Generalized Additive Models*. New York : Chapman and Hall.
- HASTIE, T. & TIBSHIRANI, R. (1996). Discriminant analysis by Gaussian mixtures. *J. R. Statist. Soc. B* **58**, 155–76.
- HASTIE, T. & TIBSHIRANI, R. (1998). Classification by pairwise coupling. *Ann. Statist.* **26** 451–71.
- HASTIE, T., TIBSHIRANI, R. & BUJA, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Am. Statist. Assoc.* **89** 1255–70.

- HAWKINS, D. M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics* **23** 105–10.
- KLASSESEN, C. A. J. & WELLNER, J. A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli* **3**, 55–77.
- KOOPERBERG, C., BOSE, S. & STONE, C. J. (1997). Polychotomous regression. *J. Am. Statist. Assoc.* **92** 117–27.
- JOE, H. (1997). *Multivariate Models and Dependence Concepts*. New York : Chapman and Hall.
- LIM, T., LOH, W. & SHIH, Y. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learning* **40**, 203–28.
- MCLACHLAN, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.
- RIANI, M. & ATKINSON, A. C. (2001). A unified approach to outliers, influence, and transformations in discriminant analysis. *J. Comp. Graph. Statist.* **10**, 513–44.
- SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
- STONE, C. J., HANSEN, M. H., KOOPERBERG, C. & TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with Discussion). *Ann. Statist.* **25**, 1371–470.
- WAHBA, G., WANG, Y., GU, C., KLEIN, R. & KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23**, 1865–95.

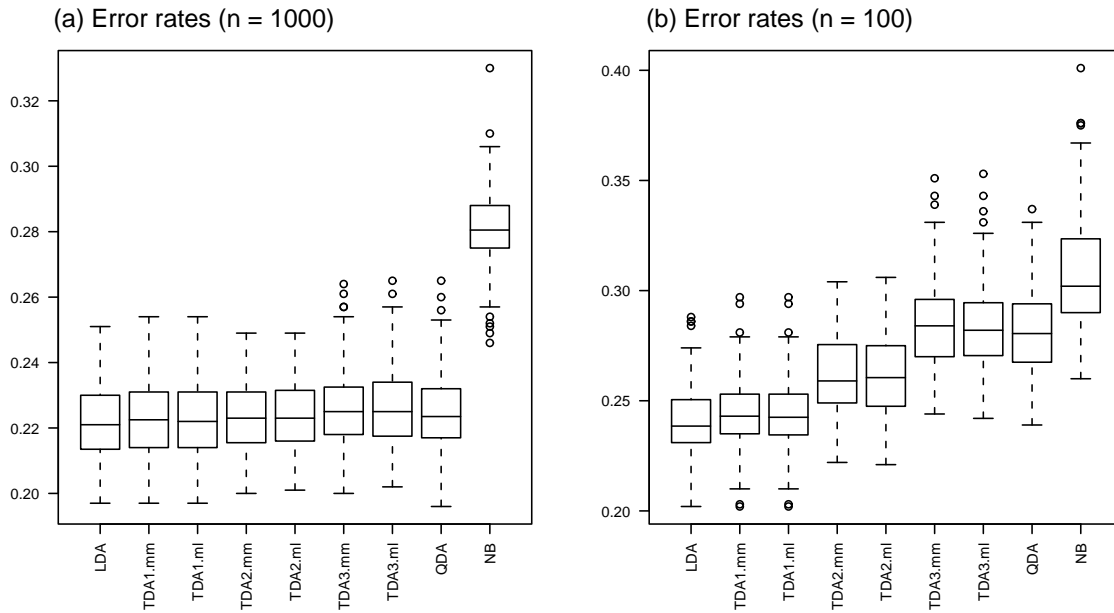


Figure 1: Box-plots of the error rates of different methods in the simulation example of §4, based on 100 simulations, (a) for  $n = 1000$ , (b) for  $n = 100$ . The methods included are linear discriminant analysis, transformation discriminant analysis methods, quadratic discriminant analysis and the Naive Bayes method.

Table 1: *The mean error rates of different methods in the simulation example of §4, based on 100 simulations. The methods included are linear discriminant analysis, transformation discriminant analysis methods, quadratic discriminant analysis, and Naive Bayes method. The standard errors for the entries in the row for  $n = 100$  are about 0.002, and the standard errors for the entries in the row  $n = 1000$  are slightly larger than but close to 0.001.*

$n$	LDA	TDA1.mm	TDA1.ml	TDA2.mm	TDA2.ml	TDA3.mm	TDA3.ml	QDA	NB
100	0.246	0.249	0.249	0.265	0.266	0.282	0.283	0.282	0.312
1000	0.222	0.222	0.222	0.224	0.224	0.226	0.226	0.225	0.281

Methods compared: LDA, linear discriminant analysis; TDA 1-3, transformation discriminant analysis corresponding to Cases 1-3 in §3.1, with mm and ml stand for the method of moments and the maximum likelihood, respectively; QDA, quadratic discriminant analysis; NB, the Naive Bayes method.

Table 2: *Classification error rates, in percentage, in six datasets. The last two columns give the minimum and maximum error rates of the thirty-three classification methods considered in Lim et al. (2000) on the same datasets. The numbers in parentheses are the ranks of the transformation discriminant analysis methods among the thirty-six, i.e. thirty-three plus the three transformation discriminant analysis methods, classification procedures, if the methods of transformation discriminant analysis were included in the comparison study of Lim et al. (2000).*

Dataset	TDA1	TDA2	TDA3	LDA	QDA	NB	FDA	POL	Lmin	Lmax
bcw	2.53(1)	3.01(3)	3.04(4)	3.98	4.90	3.78	3.80	4.24	2.78	8.48
bcw+	2.57(1)	3.04(4)	3.06(5)	3.96	4.89	3.81	3.21	3.83	2.93	7.60
bld	27.4(1.5)	27.8(3)	27.4(1.5)	31.9	40.1	35.4	28.0	28.6	27.9	43.2
bld+	27.7(3)	27.5(2)	27.4(1)	32.2	37.8	37.0	32.0	28.6	28.6	44.1
pid	22.6(8.5)	21.7(1)	24.4(25)	22.1	24.0	22.7	24.8	23.7	22.1	31.0
pid+	21.8(3)	20.9(1)	24.2(18)	21.9	23.6	22.9	22.8	21.7	21.7	31.8

TDA 1-3, transformation discriminant analysis corresponding to Cases 1-3 in §3.1 based on the method of moments; LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; NB, the Naive Bayes method; FDA, flexible discriminant analysis; POL, the POLYCLASS algorithm; Lmin and Lmax, the minimum and maximum error rate achieved by the thirty-three classification methods as reported by Lim et al. (2000).