Warsaw School of Economics
Institute of Econometrics
Department of Applied Econometrics

Department of Applied Econometrics Working Papers
Warsaw School of Economics
Al. Niepodleglosci 164
02-554 Warszawa, Poland

Working Paper No. 6-07

On modified discriminant analysis

Marcin Owczarczuk
Warsaw School of Economics

This paper is available at the Warsaw School of Economics
Department of Applied Econometrics website at: http://www.sgh.waw.pl/instytuty/zes/wp/

On modified discriminant analysis

Marcin Owczarczuk

mo23628@sgh.waw.pl

22 May 2007

Abstract

Discriminant analysis is mostly used to predict the value of a discrete dependent variable of an observation on the basis of a set of predictors. The commonly used criterion of the predictive power is the fraction of incorrectly predicted cases in the sample. In this article we construct a model for a modified discriminant problem. Namely to find a subpopulation of a given size having the highest percentage of observations of a chosen class. Our model maximizes the following criterion of the predictive power: the fraction of observations from chosen class in the found subpopulation.

Keywords: discriminant analysis, semiparametric estimation, smoothing, binary response .
JEL codes: C14, C35

1 Introduction

The aim of discriminant analysis is to predict the value of discrete explained variable Y on the basis of explanatory variables $\mathbf{X} = (X_1, \ldots, X_k)$, . We may also treat discriminant analysis as a model the aim of which is to split the space of observations into two regions: the first characterized by dominance of observations from class $Y = 0$ and second characterized by [1] $Y = 1$.

In this article we formulate the task in a different manner:

Suppose we are given the parameter τ - the fraction of population. We want to split the population with respect to explanatory variables into two groups: the first one with τ of all observations and the second one with $(1-\tau)$ of all observations. The group of size τ should have as high percentage of observations characterized by $Y = 1$ as possible. In other words we want to find population of size τ which has the highest number of observations from class $Y = 1$ in the family of all subsets of size τ.

Our model is modification of *maximum score estimator* of Manski (1975) and *smoothed maximum score estimator* of Horowitz (1992). In case of their estimators the aim is to split the population into two groups characterized by $Y = 1$ and $Y = 0$ respectively, in order to minimize the fraction of incorrectly classified observations in the sample [2]. We also want to minimize number of incorrectly classified observations but only in the group classified by the model as having $Y = 1$ with additional condition on the size of this group.

This problem arises in many areas, for example in credit scoring and marketing campaigns. In these applications the aim is to separate a small, fixed size group of clients with relative high probability of a positive value of a response variable. In case of marketing campaigns one wants to find a group of clients, on the basis of their features, who are most likely to respond to campaign - the target group. In case of credit scoring the policy of the bank may be based on assumption that a certain fraction of worst clients, for example 5% should not be granted a loan. In that case one should find 5% of customers with highest probability of default.

This problem can be solved using $\hat{p}$ - an estimator of the probability $P(Y = 1)$. It can be calculated for example using logistic regression. We may use the following construction:

1. sort observations in ascending sequence of $\hat{p}$
2. choose τ observations with highest $\hat{p}$

[1] In this paper, for simplicity, we restrict ourselves only to binary explained variable and denote its levels by $Y = 1$ and $Y = 0$.

[2] Their estimators are parameterized by $\alpha \in (0, 1)$. For $\alpha = \frac{1}{2}$ the fraction of incorrectly classified observations in the sample is minimized.

This method has intuitive grounds: observations with higher $\hat{p}$ are more likely to come from class $Y = 1$, so in the group with high $\hat{p}$ will be more observations from class $Y = 1$. Unfortunately we do not have guarantee that the chosen group will be optimal.

2 Problem formulation

We assume, as in case of classical discriminant analysis for binary responses[3], that $\mathbf{X} \in \mathbf{R}^k$ and $Y \in \{0, 1\}$. $\mathbf{X}$ and Y are random variables.

We denote

- $\mathbb{A}$ - the family of all measurable subsets of $\mathbf{R}^k$.
- f_1 - the distribution of random variable $\mathbf{X}|(Y = 1)$
- f_0 - the distribution of random variable $\mathbf{X}|(Y = 0)$
- $Bin(\pi_1)$ - the distribution of random variable Y

Formally our goal is to find subset A that maximizes:

$$\max_{A \in \mathbb{A}:\ P(A)=\tau} P(Y = 1|\mathbf{X} \in A) \tag{1}$$

We also assume that we have a random sample

$$\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1} \text{ for } y = 1$$

$$\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n_0} \text{ for } y = 0$$

$$n = n_1 + n_0$$

3 Model construction

In this section we show solution to problem described in the previous section.

3.1 Case 1. Predictors are normally distributed with common covariance matrix

This is probably most regular case. We assume that predictors have normal distribution with a common covariance matrix and probably different means among classes. These assumptions are the same as in case of Fisher linear discriminant analysis.

[3] see for example Hastie, Tibshirani, Friedman (2001)

- $\mathbf{X}|(Y=1) \sim N(\mathbf{m}_1, \mathbf{\Sigma})$[4]
- $\mathbf{X}|(Y=0) \sim N(\mathbf{m}_0, \mathbf{\Sigma})$
- $Y \sim Bin(\pi_1)$

Theorem 3.1. *Under above assumptions the optimal solution to problem (1) is*

$$A = \{\mathbf{X} : (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{X} \geq b\} \tag{2}$$

where b is quantile of order $1-\tau$ of the random variable $(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{X}$

Proof.

$$\max_{A \in \mathbb{A}:\ P(A)=\tau} P(Y=1|\mathbf{X} \in A) =$$

$$\max_{A \in \mathbb{A}:\ P(A)=\tau} \frac{P(Y=1 \wedge \mathbf{X} \in A)}{P(\mathbf{X} \in A)} = \max_{A \in \mathbb{A}:\ P(A)=\tau} \frac{\pi_1 \int_A f_1}{\pi_1 \int_A f_1 + (1-\pi_1) \int_A f_0} =$$

$$= \max_{A \in \mathbb{A}:\ P(A)=\tau} \frac{1}{1 + \frac{(1-\pi_1) \int_A f_0}{\pi_1 \int_A f_1}} \tag{3}$$

Note that

$$\operatorname*{argmax}_{A \in \mathbb{A}:\ P(A)=\tau} \frac{1}{1 + \frac{(1-\pi_1) \int_A f_0}{\pi_1 \int_A f_1}} = \operatorname*{argmax}_{A \in \mathbb{A}:\ P(A)=\tau} \frac{\pi_1 \int_A f_1}{(1-\pi_1) \int_A f_0} \tag{4}$$

Then observe that fraction $\frac{\pi_1 f_1(\mathbf{x})}{(1-\pi_1) f_0(\mathbf{x})}$ has a constant value on the line

$$c = \frac{\pi_1 f_1(\mathbf{x})}{(1-\pi_1) f_0(\mathbf{x})} = \frac{\pi_1 \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m}_1)}}{(1-\pi_1) \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m}_0)}} =$$

$$= \frac{\pi_1 e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m}_1)}}{(1-\pi_1) e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m}_0)}} \tag{5}$$

$$\ln(c) = \ln(\frac{\pi_1}{1-\pi_1}) - (\mathbf{x}-\mathbf{m}_1)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m}_1) + (\mathbf{x}-\mathbf{m}_0)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m}_0) =$$

$$= \ln(\frac{\pi_1}{1-\pi_1}) + (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} (\mathbf{m}_1 + \mathbf{m}_0) \tag{6}$$

$$\ln(c) - \ln(\frac{\pi_1}{1-\pi_1}) + \frac{1}{2} (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} (\mathbf{m}_1 + \mathbf{m}_0) = (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{x} \tag{7}$$

[4] $W \sim g$ denotes that random variable W has distribution g

which defines linear equation in $\mathbf{x}$.

So problem (1) has solution

$$A = \{\mathbf{X} : (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{X} \geq b\} \tag{8}$$

where b is chosen so that $P(\{\mathbf{X} : (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{X} \geq b\}) = \tau$. In other words b is quantile of order $1 - \tau$ of the random variable $(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{X}$. ■

Parameters $\mathbf{\Sigma}$, $\mathbf{m}_1$ and $\mathbf{m}_0$ are usually unknown and have to be estimated from data. As in case of Fisher discriminant analysis we estimate

$$\hat{\mathbf{m}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_{1i} \tag{9}$$

$$\hat{\mathbf{m}}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{x}_{0i} \tag{10}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-2} \sum_{k=0}^{1} \Big[\sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \hat{\mathbf{m}}_k)(\mathbf{x}_{ki} - \hat{\mathbf{m}}_k)^T \Big] \tag{11}$$

We may note that the optimal discriminant line is parallel to solution of Fisher linear discriminant analysis. Our model has a form:

- $y^* = (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \mathbf{x} = \mathbf{a}^T \mathbf{x} = a_1 x_1 + \cdots + a_k x_k$
- if $y^* \geq b$ then $y = 1$
- if $y^* < b$ then $y = 0$

3.2 Case 2. Predictors are normally distributed with unequal covariance matrix

We assume that predictors have normal distribution but the covariance matrix differs among classes. These assumptions are the same as in case of quadratic discriminant analysis.

- $\mathbf{X}|(Y = 1) \sim N(\mathbf{m}_1, \mathbf{\Sigma}_1)$
- $\mathbf{X}|(Y = 0) \sim N(\mathbf{m}_0, \mathbf{\Sigma}_0)$
- $Y \sim Bin(\pi_1)$

Theorem 3.2. *Under above assumptions the optimal solution to problem (1) is*

$$A = \{\mathbf{X} : \mathbf{X}^T (\mathbf{\Sigma}_1^{-1} \mathbf{m}_1 - \mathbf{\Sigma}_0^{-1} \mathbf{m}_0) - \frac{1}{2} \mathbf{X}^T (\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1}) \mathbf{X} \geq b\} \tag{12}$$

where b is quantile of order $1-\tau$ of the random variable $\mathbf{X}^T(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_0^{-1}\mathbf{m}_0) - \frac{1}{2}\mathbf{X}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1})\mathbf{X}$

Proof.

The proof is similar to proof of Theorem 3.1. We may observe that the fraction $\frac{\pi_1 f_1(\mathbf{x})}{(1-\pi_1)f_0(\mathbf{x})}$ is constant on the curve

$$c = \frac{\pi_1 f_1(\mathbf{x})}{(1-\pi_1)f_0(\mathbf{x})} = \frac{\pi_1 \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}_1|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mathbf{m}_1)}}{(1-\pi_1)\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}_0|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\mathbf{\Sigma}_0^{-1}(\mathbf{x}-\mathbf{m}_0)}} \tag{13}$$

$$\ln(c) = \ln\frac{\pi_1}{1-\pi_1} + \frac{1}{2}\ln\frac{|\mathbf{\Sigma}_0|}{|\mathbf{\Sigma}_1|} + \mathbf{x}^T(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_0^{-1}\mathbf{m}_0) - \frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1})\mathbf{x} - \frac{1}{2}\mathbf{m}_1^T\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 + \frac{1}{2}\mathbf{m}_0^T\mathbf{\Sigma}_0^{-1}\mathbf{m}_0 \tag{14}$$

$$\ln(c) - \ln\frac{\pi_1}{1-\pi_1} - \frac{1}{2}\ln\frac{|\mathbf{\Sigma}_0|}{|\mathbf{\Sigma}_1|} + \frac{1}{2}\mathbf{m}_1^T\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \frac{1}{2}\mathbf{m}_0^T\mathbf{\Sigma}_0^{-1}\mathbf{m}_0 = \mathbf{x}^T(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_0^{-1}\mathbf{m}_0) - \frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1})\mathbf{x} \tag{15}$$

which defines a quadratic equation in $\mathbf{x}$.

So problem (1) has solution

$$A = \{\mathbf{X} : \mathbf{X}^T(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_0^{-1}\mathbf{m}_0) - \frac{1}{2}\mathbf{X}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1})\mathbf{X} \geq b\} \tag{16}$$

where b is chosen so that $P(\{\mathbf{X} : \mathbf{X}^T(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_0^{-1}\mathbf{m}_0) - \frac{1}{2}\mathbf{X}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1})\mathbf{X}\}) = \tau$. In other words b is quantile of order $1-\tau$ of the variable $\mathbf{X}^T(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_0^{-1}\mathbf{m}_0) - \frac{1}{2}\mathbf{X}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1})\mathbf{X}$. ■

Parameters $\mathbf{\Sigma_1}$, $\mathbf{m_1}$, $\mathbf{\Sigma_0}$ and $\mathbf{m_0}$ are usually unknown and have to be estimated from data. As in case of quadratic discriminant analysis we estimate

$$\hat{\mathbf{m}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_{1i} \tag{17}$$

$$\hat{\mathbf{m}}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0}\mathbf{x}_{0i} \tag{18}$$

$$\hat{\mathbf{\Sigma}}_1 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(\mathbf{x}_{1i} - \hat{\mathbf{m}}_1)(\mathbf{x}_{1i} - \hat{\mathbf{m}}_1)^T \tag{19}$$

$$\hat{\mathbf{\Sigma}}_0 = \frac{1}{n_0 - 1}\sum_{k=1}^{n_0}(\mathbf{x}_{0i} - \hat{\mathbf{m}}_0)(\mathbf{x}_{0i} - \hat{\mathbf{m}}_0)^T \tag{20}$$

We may note that the optimal discriminant line is parallel to solution of quadratic discriminant analysis.

In this case our model has a form:

- $y^* = \mathbf{x}^T(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_0^{-1}\mathbf{m}_0) - \frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1})\mathbf{x} = \sum_{i=1}^k a_i x_i + \sum_{j=1}^k \sum_{i=1}^k a_{ij} x_i x_j$
- if $y^* \geq b$ then $y = 1$
- if $y^* < b$ then $y = 0$

3.3 Case 3. Semiparametric model

In case we cannot make additional assumptions about the distribution of predictors, the problem (1) becomes NP-hard. So we must restrict ourselves to subsets A of particular form. In this subsection we show semiparametric optimal solution in case subset A is bounded by a hyperplane:

$$A = \{(X_1, \ldots, X_k) : a_1 X_1 + a_2 X_2 + \cdots + a_k X_k \geq b\} = \{\mathbf{X} : \mathbf{a}^T\mathbf{X} \geq b\} \tag{21}$$

for some $a_1, a_2, \ldots, a_k, b \in R$. In other words we want to construct linear discriminant function. Since inequality $\mathbf{a}^T\mathbf{X} \geq b$ holds when multiplied by any positive constant we imply normalization $\|\mathbf{a}\| = 1$.

Comment

Let us recall the definition of *maximum score estimator* of Manski (1975):

$$\max_{\mathbf{a}} S_n^{\alpha} = \frac{1}{n}\sum_{i=1}^{n}[(2Y_i - 1) - (1 - 2\alpha)]\mathbf{1}(\mathbf{a}^T\mathbf{x}_i) \tag{22}$$

$$\text{subject to } \|\mathbf{a}\| = 1 \tag{23}$$

which in case $\alpha = \frac{1}{2}$ takes form

$$\max_{\mathbf{a}} S_n^{\frac{1}{2}} = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)\mathbf{1}(\mathbf{a}^T\mathbf{x}_i) \tag{24}$$

$$\text{subject to } \|\mathbf{a}\| = 1 \tag{25}$$

Our problem can be formulated as

$$\max_{\mathbf{a},b} S_n = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)\mathbf{1}(\mathbf{a}^T\mathbf{x}_i \geq b) \tag{26}$$

$$\text{subject to } \|\mathbf{a}\| = 1 \quad \text{and} \quad b = q_{1-\tau}(\mathbf{a}^T\mathbf{x}_1, \ldots, \mathbf{a}^T\mathbf{x}_n) \tag{27}$$

■

Our model has the same form as in Case 1:

- $y^* = a_1 x_1 + \cdots + a_k x_k$
- if $y^* \geq b$ then $y = 1$
- if $y^* < b$ then $y = 0$

To derive optimal solution let us recall its formulation

$$\max_{A \in \mathbb{A}:\ P(A)=\tau} P(Y = 1 | (X_1, \ldots, X_n) \in A) \tag{28}$$

We may write

$$\max_{A \in \mathbb{A}:\ P(A)=\tau} P(Y = 1 | \mathbf{X} \in A) = \max_{\|\mathbf{a}\|=1, b:\ P(\mathbf{a}^T \mathbf{X} \geq b)=\tau} P(Y = 1 | \mathbf{a}^T \mathbf{X} \geq b) =$$

$$= \max_{\|\mathbf{a}\|=1, b:\ P(\mathbf{a}^T \mathbf{X} \geq b)=\tau} \frac{P(Y = 1 \wedge \mathbf{a}^T \mathbf{X} \geq b)}{P(\mathbf{a}^T \mathbf{X} \geq b)} =$$

$$= \max_{\|\mathbf{a}\|=1, b:\ P(\mathbf{a}^T \mathbf{X} \geq b)=\tau} \frac{1 - F_{\mathbf{a}^T \mathbf{X} | Y=1}(b)}{\tau} =$$

$$= \max_{\|\mathbf{a}\|=1, b:\ 1 - F_{\mathbf{a}^T \mathbf{X}}(b)=\tau} \frac{1 - F_{\mathbf{a}^T \mathbf{X} | Y=1}(b)}{\tau} \tag{29}$$

In case the distributions f_1, f_0, and $Bin(\pi_1)$ are known, problem (29) is a deterministic optimization problem of the form:

$$\max_{\mathbf{a} \in R^p, b \in R} \frac{1 - F_{\mathbf{a}^T \mathbf{X} | Y=1}(b)}{\tau} \tag{30}$$

subject to

$$1 - F_{\mathbf{a}^T \mathbf{X}}(b) = \tau \tag{31}$$

$$\|\mathbf{a}\| = 1 \tag{32}$$

In case distributions f_1, f_0, and $Bin(\pi_1)$ are not known, they have to be estimated from data.

Cumulative distribution functions $F_{X|Y=0}$ and $F_{X|Y=1}$ can be estimated consistently by either empirical cdf or by kernel cdf. In both cases problem (30)-(32) is replaced by

$$\max_{\mathbf{a} \in R^p, b \in R} \frac{1 - \hat{F}_{\mathbf{a}^T \mathbf{X} | Y=1}(b)}{\tau} \tag{33}$$

subject to

$$1 - \hat{F}_{\mathbf{a}^T\mathbf{X}}(b) = \tau \tag{34}$$

$$\|\mathbf{a}\| = 1 \tag{35}$$

In case of differentiable kernels it is a differentiable optimization problem and can be easily solved numerically by standard gradient techniques.

Let us denote by $\hat{\mathbf{a}}$ and $\hat{b}$ the solution of the problem (33)-(35). The following theorem holds

Theorem 3.3. *Assume that*

- *(33)-(35) has almost sure one unique solution,*
- *(30)-(32) has one unique solution*
- $\hat{F}_{\mathbf{a}^T\mathbf{x}}(b)$ *and* $\hat{F}_{\mathbf{a}^T\mathbf{x}|Y=1}(b)$ *are continuous functions of* $\mathbf{a}$,b *and* $\mathbf{x}$.
- f_1 *and* f_0 *are continuous*

Then $\hat{\mathbf{a}}$ *and* $\hat{b}$ *are consistent estimators of* $\mathbf{a}$ *and* b.

Lemma 3.1. *If* $f : R^n \times R^m \to R$ *is continuous and* $B \in R^n$ *is bounded and closed, then* $\max_{x\in B} f(x, y) = g(y)$ *is continuous.*

Lemma 3.2. *If* $X_n \xrightarrow{P} X$, f *is continuous on* A *and* $P(X \in A) = 1$ *then* $f(X_n) \xrightarrow{P} f(X)$

Proof of Theorem 3.3

We may note that for any fixed $\mathbf{a}^*$ and b^*

$$\frac{1 - \hat{F}_{\mathbf{a}^{*T}\mathbf{X}|Y=1}(b^*)}{\tau} \xrightarrow{P} \frac{1 - F_{\mathbf{a}^{*T}\mathbf{X}|Y=1}(b^*)}{\tau} \tag{36}$$

$$1 - \hat{F}_{a^{*T}X}(b^*) \xrightarrow{P} 1 - F_{a^{*T}X}(b^*) \tag{37}$$

We may use Lemma 3.2 and Lemma 3.1 and write

$$\max_{\mathbf{a}\in R^p, b\in R} \frac{1 - \hat{F}_{\mathbf{a}^T\mathbf{X}|Y=1}(b)}{\tau} \xrightarrow{P} \max_{\mathbf{a}\in R^p, b\in R} \frac{1 - F_{\mathbf{a}^T\mathbf{X}|Y=1}(b)}{\tau} \tag{38}$$

■

Comment

Let us recall the definition of *smoothed maximum score estimator* of Horowitz (1992) for $\alpha = \frac{1}{2}$:

$$\max_a S_n^{\frac{1}{2}} = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)K\left(\frac{a^T\mathbf{x}_i}{h}\right) \tag{39}$$

$$\text{subject to } \|a\| = 1 \tag{40}$$

where $K(\cdot)$ is smooth cdf. Our model can be formulated as

$$\max_{a,b} S_n = \frac{1}{n}\sum_{i=1}^{n} Y_i K\left(\frac{b - a^T \mathbf{x}_i}{h}\right) \tag{41}$$

$$\text{subject to } \|a\| = 1 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{b - a^T \mathbf{x}_i}{h}\right) = 1 - \tau \tag{42}$$

■

4 Examples of model performance

In this section we show some examples of the model performance on artificial data. We compare our semiparametric model described in the previous section to logistic regression, which is also linear. We set $\tau = 10\%$, that is both models should find subpopulation of size 10% with highest fraction of $Y = 1$. Our model finds the best subpopulation directly. In case of logistic regression we find optimal τ% of observations by technique described in introduction, namely we calculate for each obervation $\hat{p}_i$, the estimator of $P(Y_i = 1)$, $i = 1, \ldots, n$ and choose τ% observations with highest $\hat{p}_i$. The smoothing parameter h in semiparametric model is set to 1. There are 5% observations with $Y = 1$ in the sample. The observations come from the mixture of multivariate normal distributions.

4.0.1 Example 1

We generated the following sample:

$$\begin{cases} (X_1, X_2) \sim N\left(\begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}\right), \quad Y = 1, \quad n = 110 \\ (X_1, X_2) \sim N\left(\begin{bmatrix} 9 \\ 9 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}\right), \quad Y = 1, \quad n = 90 \\ (X_1, X_2) \sim N\left(\begin{bmatrix} 10 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}\right), \quad Y = 1, \quad n = 100 \\ (X_1, X_2) \sim N\left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}\right), \quad Y = 0, \quad n = 5700 \end{cases} \tag{43}$$

The graph of this sample is shown on Figure 1.

4.0.2 Example 2

We generated the following sample:

$$\begin{cases} (X_1, X_2) \sim N\left(\begin{bmatrix} 1 \\ 7 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}\right), \quad Y = 1, \quad n = 300 \\ (X_1, X_2) \sim N\left(\begin{bmatrix} 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}\right), \quad Y = 0, \quad n = 5700 \end{cases} \tag{44}$$

The graph of this sample is shown on Figure 2.

We may note that in case of predictors having normal distribution both models achieve similar results but in multimodal case such as mixture of normal distributions within class, logistic

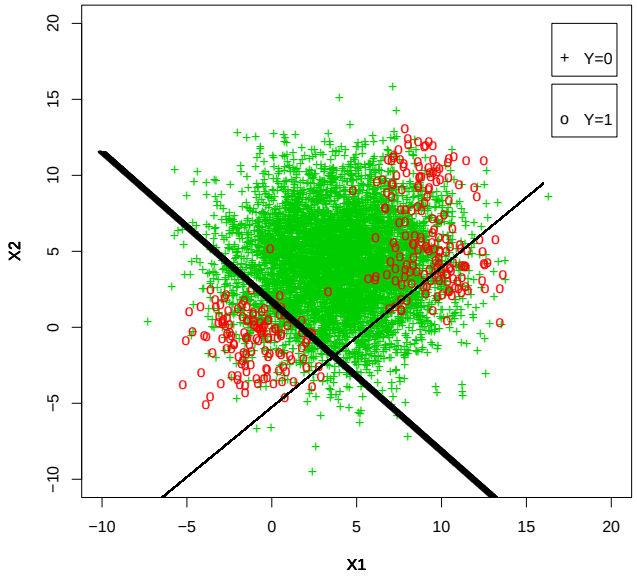

Figure 1: Example 1. The thick line represents the boundary of the separated region for semiparametric model and thin line for logistic regression. Fraction of $Y = 1$ in subpopulation separated by logistic regression: 8.83%. Fraction of $Y = 1$ in subpopulation separated by semiparametric model: 22.3%

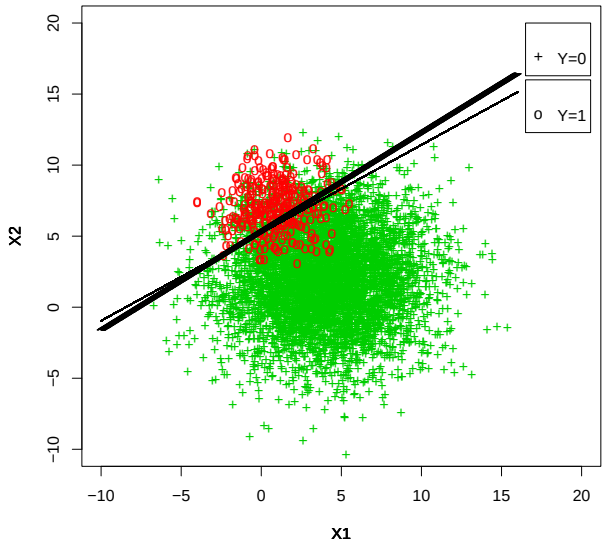

Figure 2: Example 2. The thick line represents the boundary of the separated region for semiparametric model and thin line for logistic regression. Fraction of $Y = 1$ in subpopulation separated by logistic regression: 37.89%. Fraction of $Y = 1$ in subpopulation separated by semiparametric model: 37.19%

regression is outperformed by semiparametric model. In Example 1 linear discriminant function is probably not optimal but still reasonable. These results can be generalized due to Theorem 3.3, that is for large samples, semiparametric model is almost surely better or at least as good as logistic regression.

5 Conclusions

The paper shows the solution to modified discriminant analysis problem. We proved that when predictors have normal distribution, the optimal solution is parallel to Fisher linear and quadratic discriminant analysis in case of equal and unequal covariance matrix, respectively. When predictors are not Gaussian we have shown how to build an optimal semiparametric linear discriminant function . Semiparametric model is superior to logistic regression, especially when distributions are multimodal. In some cases linear functions are not optimal but our approach can be easily generalized to neural networks in order to capture nonlinearities in the data.

References

[1] Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning*, Springer-Verlag 2001

[2] Horowitz, J. L. *A Smoothed Maximum Score Estimator for the Binary Response Model*, Econometrica, 60(3) 1992, p. 505-531

[3] Manski, C. F *Maximum Score Estimation of the Stochastic Utility Model of Choice*, Journal of Econometrics 3 1975, p. 205-228