Proceedings of the 28th West Indies Agricultural Economics Conference /2009 Barbados National Agricultural Conference

In collaboration with

The Ministry of Agriculture, Barbados

The University of the West Indies (UWI)

"Food Security, Investment Flows and Agricultural Development in the Caribbean"

Bridgetown, Barbados
6th-10th July, 2009

Neela Badrie
Editor

Application of nonparametric discriminant analysis for assessing food safety issues of Caribbean imports

Asheber Abebe
Department of Mathematics & Statistics, 221 Parker Hall, Auburn University, AL 36849, USA. E-mail: abebeas@auburn.edu. Tel: (334) 844-6567
Corresponding author

Shuxin Yin
Department of Mathematics & Statistics, 221 Parker Hall, Auburn University, AL 36849, USA. E-mail: yinshux@auburn.edu.

Abstract

Caribbean food imports often face detentions and refusals by the US resulting in a major loss of income. In this paper, we consider a classification procedure based on transvariation probabilities to correctly identify cases that lead to food detention. This is based upon several background variables on fourteen Latin American and Caribbean countries. A method for selecting variables according to their contribution towards predicting detention is given. For our particular sample, the selection method chose foreign direct investment as the variable that carried the most information in determining food detention. After removing variables that were non-informative about food detention status, a leave-one-out cross-validation shows that methods based on transvariation probabilities were superior to classical methods in predicting food detention.

Keywords: transvariation; projection pursuit; misclassification error rate; food detention.

1.0 Introduction

Illnesses caused by consuming contaminated foods or beverages have garnered increased attention of late mainly due to certain high profile cases in the United States. Moreover, increasing food demand has led to increasing food imports by the US. These have resulted in an increase in the number of food detentions and refusals at US ports of entry. To export foods and beverages to the US, Caribbean countries must adhere to stringent standards set by the World Trade Organization (WTO) and the US Food and Drug Administration (FDA). In this paper, the focus will be on the two-group discrimination problem of determining, with as much accuracy as possible, whether exports to the US will face detention using a set of variables measured yearly. We are also interested in *ranking* the variables according to their ability to determine food detention. In this paper, the interest is in the development of the methods and reporting of the results. We leave open the discussion of policy decision and further economic implications of the results.

Discriminant analysis is a procedure for assigning an individual data point into one of K ($K > 1$) known groups based on previously known information related to the

K groups. The available information is usually in the form of *training data* whose correct classification is known. A discriminant analysis procedure uses the correct classification information contained in the training data to create a rule for assigning new observations to one of the *K* groups. Although classification decisions have been made for millennia, Fisher (1936) gave what is considered to be the first scientific approach to discriminant analysis. Fisher projected the multivariate data onto a one-dimensional space, where he chose as the projection direction the one that maximizes the variance in the projected space. Allocation is then done in this one-dimensional space using simple Euclidean distances from the group means of the projected data. This method is optimal under multivariate normality and homoscedasticity of the groups. This method is not robust especially to heteroscedasticity.

In this paper, we consider nonparametric classifiers that do not depend on many of the restrictive assumptions required by classical methods. These classifiers are well suited for the data set under consideration since the variation in sizes of countries' economies is asymmetric and contains some outliers. A much more detailed discussion of the procedure used here is found in Nudurupati and Abebe (2009). We will use transvariation probabilities that are a version of *data depth* (Tukey, 1974) to construct nonparametric classifiers and to provide ranking of variables according to their ability to discriminate among the *K* groups. Other classifiers based on depth can be found in Jörnsten (2004); Ghosh and Chaudhuri (2005); Billor et al. (2008); Abebe and Nudurupati (2009).

2.0 Background

Consider two *d*-dimensional populations $\prod_x$ and $\prod_y$ with underlying distributions *F* and *G*, respectively, each defined on $\mathbf{R}^d$ for *d* ? 1. Suppose we have independent random samples from $\prod_x$ and $\prod_y$ given by $\mathbb{X} = \{X_1, \ldots, X_{n_x}\}$ and $\mathbb{Y} = \{Y_1, \ldots, Y_{n_y}\}$. Let F_{n_x} and G_{n_y} represent the empirical distribution functions of $\mathbb{X}$ and $\mathbb{Y}$ respectively.

Consider the problem of classifying a new observation $Z \in \prod_x \cup \prod_y$ in either $\prod_x$ or $\prod_y$. Suppose we have a function $D: \mathbf{R}^d \to \mathbf{R}$ such that **Z** is classified in $\prod_x$ if *D* (**Z**; *F*, *G*) > 0. The function *D* is known as a discriminant function. The probabilities of misclassification of an observation from $\prod_x$ in $\prod_y$ and $\prod_y$ in $\prod_x$ are $P_1(y|x)^D = P\{D(Z; F, G) < 0 \mid Z \sim F\}$ and $P_1(x|y)^D = P\{D(Z; F, G) > 0 \mid Z \sim G\}$ respectively. Assuming no prior preference for either population, the total probability of misclassification (TPM) is then $P^D = \frac{1}{2} P^D_{(y|x)} + \frac{1}{2} P^D_{(x|y)}$.

Fisher (1936) looked at a linear combination of the *d*-covariates that maximizes the separation between the two populations $\prod_x$ and $\prod_y$. This gives rise to the Linear Discriminant Function (LDF)

$$L(z; F, G) = (\mu_x - \mu_y)' \Sigma^{-1} \left[z - \frac{1}{2}(\mu_x + \mu_y) \right]$$

A new observation **Z**=**z** is now assigned to

$\prod_x$ if $L(\mathbf{z};F,G) > 0$ and to $\prod_y$ otherwise. This method is optimal (minimizing TPM) in classifying the new observation **Z** under the assumption that *F* and *G* have the distributions $N_d(\mu_x, \Sigma)$ and $N_d(\mu_y, \Sigma)$, respectively.

Given the random samples *X* and *Y*, the sample version of LDF is

$$L\left(\mathbf{z}; F_{n_x}, G_{n_y}\right) = (\bar{\mathbf{x}} - \bar{\mathbf{y}})' S_p^{-1}\left[\mathbf{z} - \frac{1}{2}(\bar{\mathbf{x}} + \bar{\mathbf{y}})\right]$$

where S_p is the pooled estimator of Σ .

The LDF is sensitive to deviations from normality and equal covariance. Lachenbruch et al. (1973), Lachenbruch (1975), Hills (1967), McLachlan (1992), Anderson (1984), Dillon (1979), Johnson et al. (1979) and Seber (1984), among others, have investigated the robustness of LDF. Their work found that the LDF is greatly affected by certain types of non-normality.

3.0 Group Separation

3.1 LDF

Fisher's idea of picking a linear combination that maximizes the separation between the two samples could be reframed as finding $\mathbf{u} \in R^d$, say $\hat{\mathbf{u}}_0$, the projection direction that maximizes the square of the two-sample *t* statistic; that is

$$\hat{\mathbf{u}}_0 = \underset{\|\mathbf{u}\|=1}{\mathrm{Argmax}} \frac{[\mathbf{u}'(\bar{\mathbf{X}} - \bar{\mathbf{Y}})]^2}{\mathbf{u}_p'S\,\mathbf{u}\left(\frac{n_x + n_y}{n_x n_y}\right)} .$$

The data are then reduced to one dimension by projecting them in the direction given by $\hat{\mathbf{u}}_0$ and one would classify a new observation **Z**=**z** into $\prod_x$ if $|Z_0 - \bar{X}_0| < |Z_0 - \bar{Y}_0|$, where $X_{0i} = \hat{\mathbf{u}}_0'\mathbf{X}_i$, $Y_{0i} = \hat{\mathbf{u}}_0'\mathbf{Y}_i$, and $Z_0 = \hat{\mathbf{u}}_0'\mathbf{Z}$, $i = 1, \dots, n_x$ and $j = 1, \dots, n_y$. Otherwise, one classifies **Z** into $\prod_y$.

Montanari (2004) and Chen and Muirhead (1994) used a two-sample Mann-Whitney type nonparametric statistic as a projection index to measure group separation in place of the two-sample *t*-statistic. They showed that their projection pursuit methods are not sensitive to deviations from the homoscedasticity and normality assumptions. Their method is related to the idea of transvariation probability (Gini, 1916).

3.2 Transvariation

Consider two continuous univariate populations $\prod_x$ and $\prod_y$ with distributions *F* and *G*, respectively, defined on $\mathbf{R}$. Suppose we have two random samples $\mathbf{X}_1, \dots, \mathbf{X}_{n_x}$ from $\prod_x$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_y}$ from $\prod_y$. The two samples are said to *transvariate* with respect to their measures of centers m_x and m_y if there is at least one pair (i, j) such that $(\mathbf{X}_i - \mathbf{Y}_j)(m_x - m_y) < 0$, $i = 1, \dots, n_x$ and $j = 1, \dots, n_y$. Any difference satisfying this condition is called a *two-group transvariation*. Similarly, the sample $\mathbf{X}_1, \dots, \mathbf{X}_{n_x}$ and a given constant $c \in \mathbf{R}$ transvariate with respect to m_x, if there is at least one *i* such that $(\mathbf{X}_i - c)(m_x - c) < 0$, $i = 1, \dots, n_x$. This is known as a *point-group transvariation*.

Define

$$\phi(x) = \begin{cases} 0, & x > 0 \\ 1, & x < 0 \\ .5, & x = 0 \end{cases}$$

The two-group transvariation probability between *F* and *G* is defined as

$$\tau(\mathrm{F},\mathrm{G}) = \iint \phi\{(x-y)(\mu(\mathrm{F})-\mu(\mathrm{G}))\}\, d\mathrm{F}(x)d\mathrm{G}(y)$$

where $\mu(\mathrm{F})$ and $\mu(\mathrm{G})$ are the location functions of *F* and *G*. If F_{n_x} and G_{n_y} are the two empirical distributions of the two random samples, then an estimate of ? is given as

$$T_{xy} := \tau\left(\mathrm{F}_{n_x}, \mathrm{G}_{n_y}\right) = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\phi\{(\mathrm{X}_i-\mathrm{Y}_j)(m_x-m_y)\}.$$

T_{xy} is a nonparametric estimator of the overlap between the distributions F_x and F_y. In particular, $n_x n_y T_{xy}$ gives the number of observations that need to be interchanged so that there will be no overlap between the two samples. For symmetric distributions, $0 \le T_{xy} \le 0.5$, where $T_{xy} = 0$ means complete separation. If $\mu_x < \mu_y$, then $\tau_{xy} = P(\mathrm{X} < \mathrm{Y})$ which is estimated by

$$\mathrm{T}_{xy} = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\phi\left(\mathrm{X}_i - \mathrm{Y}_j\right) = \frac{u_{xy}}{n_x n_y}$$

where u_{xy} is the Mann-Whitney statistic (Hollander and Wolfe, 1999). Rank statistics and u_{xy} are related through

$$u_{xy} = \sum_{j=1}^{n_y} R(\mathrm{Y}_j) + \frac{n_y(n_y+1)}{2},$$

where $R(\mathrm{Y}_j)$ is the rank of Y_j in the joint ranking of $\mathrm{X}_1,\ldots,\mathrm{X}_{n_x}$ and $\mathrm{Y}_1,\ldots,\mathrm{Y}_{n_y}$ for $j = 1,\ldots,n_y$.

The point-group transvariation probability between *F* and a constant $c \in \mathbf{R}$ is given by

$$\tau_x(c) = \int \phi\{(x-c)(\mu(\mathrm{F})-c)\}dF(x) .$$

An estimator of $\tau_x(c)$ is

$$\mathrm{T}_x(c) = \frac{1}{n_x}\sum_{i=1}^{n_x}\phi\{(\mathrm{X}_i - c)(m_x - c)\} \quad (3.1)$$

$T_x(c)$ measures the centrality of the constant c in the sample $\mathrm{X}_1,\ldots,\mathrm{X}_{n_x}$. In a way, $T_x(c)$ measures how deep the point *c* is in the sample $\mathrm{X}_1,\ldots,\mathrm{X}_{n_x}$. The quantity $n_x T_x(c)$ is the fewest number of observations in the first sample that c needs to *skip* so that all the sample points are to one side of it.

Projection pursuit (Friedman and Tukey, 1974) offers a way to generalize the idea of transvariation probability for dimensions higher than one. To that end, let $\mathrm{F_u}$ and $\mathrm{G_u}$ be the distributions of $\mathrm{u}'\mathrm{X}$ and $\mathrm{u}'\mathrm{Y}$, respectively, where **X** ? F from population $\prod_x$ and **Y** ? G from population $\prod_y$ are *d*-dimensional random variables and $\mathrm{u} \in \mathbf{R}^d$ is a unit vector. The overlap between $\mathrm{F_u}$ and $\mathrm{G_u}$ with respect to the transvariation probability is

$$\iint \phi\{(x-y)(\mu(\mathrm{F_u})-\mu(\mathrm{G_u}))\}\, d\mathrm{F_u}(x)d\mathrm{G_u}(y)$$

We are interested in finding the projection direction that minimizes this overlap between $\mathrm{F_u}$ and $\mathrm{G_u}$; that is

$$\mathrm{u}_{opt} = \underset{\|\mathrm{u}\|=1}{\mathrm{Argmax}}\left\{\iint \phi\{(x-y)(\mu(\mathrm{F_u})-\mu(\mathrm{G_u}))\}\, d\mathrm{F_u}(x)d\mathrm{G_u}(y)\right\}$$

Given two independent random training samples $\mathrm{X}_1,\ldots,\mathrm{X}_{n_x}$ and $\mathrm{Y}_1,\ldots,\mathrm{Y}_{n_y}$ from $\prod_x$ and $\prod_y$, respectively, defined on $\mathbf{R}^d$ (*d* ? 2), the estimator of the direction of minimum overlap is given by

$$\hat{\mathrm{u}}_{opt} = \underset{\|\mathrm{u}\|=1}{\mathrm{Argmax}}\frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\phi\left\{\left(\mathrm{u}'\mathrm{X}_i - \mathrm{u}'\mathrm{Y}_j\right)\left(m_x(\mathrm{u}) - m_y(\mathrm{u})\right)\right\},$$

where $m_x(\mathrm{u})$ and $m_y(\mathrm{u})$ are the locations of the two projected samples $\mathrm{u}'\mathrm{X}$ and $\mathrm{u}'\mathrm{Y}$, respectively. This vector gives the direction of maximum separation as measured by Gini's transvariation probability (Gini, 1916).

3.3 Stepwise Variable Selection

Let $V_1, \dots, V_d$ be the variables on which the classification is to be based. Write $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_d]$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_d]$, where $\mathbf{x}_i \in \mathbf{R}^{n_x}$, $\mathbf{y}_j \in \mathbf{R}^{n_y}$, for $i, j = 1, \dots, d$. Thus, $V_i = [\mathbf{x}_i' \ \ \mathbf{y}_i']'$. Firstly, we would like to rank the variables $V_1, \dots, V_d$ according to the amount of *information* they provide for class determination. The most informative variables are those for which there is maximum separation between the two groups. It is, thus, intuitive to use the univariate two-group transvariation probability to measure the contribution of variables towards discriminating the two groups.

Let

$$\tilde{\mathrm{T}}(\mathbf{V}_s) = 1 - \frac{2}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \phi\left\{(\mathbf{x}_{si} - \mathbf{y}_{sj}) m_{xy}^s\right\}$$

for $s = 1, \dots, d$, where m_{xy}^s is the median of $\{\mathbf{x}_{si} - \mathbf{y}_{sj} : 1 \le i \le n_x, 1 \le j \le n_y\}$. As discussed in Calò (2006), using the median of the differences implies that $0 \le \tilde{\mathrm{T}}(\mathbf{V}_s) \le 1$. In this case, higher values of $\tilde{\mathrm{T}}(\mathbf{V}_s)$ imply less overlap between $\mathbf{x}_s$ and $\mathbf{y}_s$. The most informative variable is the one with the highest $\tilde{\mathrm{T}}(\mathbf{V}_s)$. Denote this variable by $\mathbf{V}_{(1)}$.

As the second most informative variable, it seems reasonable to pick the variable that is the most dissimilar to $\mathbf{V}_{(1)}$ while at the same time giving the highest contribution to distinguishing the two groups. To that end, let

$$\tilde{\mathrm{T}}\left(\mathbf{V}_s \middle| \mathbf{V}_{(1)}\right) = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \phi\left\{\left(\mathbf{x}_{(1)i} - \mathbf{y}_{(1)j}\right) m_{xy}^{(1)} \left(\mathbf{x}_{si} - \mathbf{y}_{sj}\right) m_{xy}^s\right\}$$

Higher values of $\tilde{\mathrm{T}}\left(\mathbf{V}_s \middle| \mathbf{V}_{(1)}\right)$ indicate greater dissimilarity of $\mathbf{V}_s$ to $\mathbf{V}_{(1)}$. Now we take as the second variable

$$\mathbf{V}_{(2)} = \left\{\mathbf{V}_p : \tilde{\mathrm{T}}(\mathbf{V}_p)\,\tilde{\mathrm{T}}\left(\mathbf{V}_p \middle| \mathbf{V}_{(1)}\right) \ge \tilde{\mathrm{T}}(\mathbf{V}_j)\,\tilde{\mathrm{T}}\left(\mathbf{V}_j \middle| \mathbf{V}_{(1)}\right); j \in \{1, \dots, d\} \backslash \{(1)\}\right\}$$

; that is, the one with the largest product of $\tilde{\mathrm{T}}(\cdot)$ and $\tilde{\mathrm{T}}(\cdot \,|\mathbf{V}_{(1)})$.

Then using $\left[\mathbf{V}_{(1)} \ V_{(2)}\right]$, we find the vector $\hat{\mathbf{u}}_{opt}$ that gives us the optimal direction of separation using (3.3). We then replace $\mathbf{V}_{(1)}$ by $\hat{\mathbf{u}}_{opt}'\left[\mathbf{V}_{(1)} \ V_{(2)}\right]$ in (3.4) to find the next variable $\mathbf{V}_{(3)}$, say. This process is continued until the product of $\tilde{\mathrm{T}}(\cdot)$ and $\tilde{\mathrm{T}}\left(\cdot \,|\mathbf{V}_{(j)}\right)$ gets smaller than a threshold value, $0 < \alpha < 1$. In the current paper, we use $\alpha = 0.10$.

4.0 Allocation Schemes

Once the direction of maximum separation is found, the next step is to project all the data (including the new sample point) onto that direction and allocate the new point to one of the two populations.

4.1 Allocation Based on Distance

Given two independent random training samples $\mathbf{X}_1, \dots, \mathbf{X}_{n_x}$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_y}$ from $\prod_x$ and $\prod_y$, respectively, defined on $\mathbf{R}^d$ (*d* ? 2), a new observation $\mathbf{Z}$ is classified in $\prod_x$ if $\left|\hat{\mathbf{u}}_{opt}'\mathbf{Z} - m_x(\hat{\mathbf{u}}_{opt})\right| < \left|\hat{\mathbf{u}}_{opt}'\mathbf{Z} - m_y(\hat{\mathbf{u}}_{opt})\right|$ and in $\prod_y$ otherwise. Here $m_x(\hat{\mathbf{u}}_{opt})$ and $m_y(\hat{\mathbf{u}}_{opt})$ are centers of the two projected groups. One may take either the mean or the median as a measure of location. The LDF uses the mean whereas transvariation based methods use the median (Montanari, 2004). Hereafter the classifier obtained using this allocation method will be referred to as Transvariation-Distance (TD) classifier. This method can be adversely affected by skewness and outliers.

4.2 Allocation Based on Point-Group Transvariation

An allocation method suggested by Montanari (2004) is based on the ranking of the new observation among the two samples. This utilizes the point group transvariation (3.2). Allocate a new observation **Z** into $\prod_x$ if $\mathbf{T}_x(\mathbf{Z}) > \mathbf{T}_y(\mathbf{Z})$; otherwise, it is assigned to $\prod_y$ where

$$\mathrm{T}_x(\mathrm{Z}) = \frac{1}{n_x}\sum_{i=1}^{n_x} \phi\left\{\left(\hat{\mathrm{u}}'_{opt}\mathrm{X}_i - \hat{\mathrm{u}}'_{opt}\mathrm{Z}\right)\left(m_x(\hat{\mathrm{u}}_{opt}) - \hat{\mathrm{u}}'_{opt}\mathrm{Z}\right)\right\}$$

and

$$\mathrm{T}_y(\mathrm{Z}) = \frac{1}{n_y}\sum_{j=1}^{n_y} \phi\left\{\left(\hat{\mathrm{u}}'_{opt}\mathrm{Y}_j - \hat{\mathrm{u}}'_{opt}\mathrm{Z}\right)\left(m_y(\hat{\mathrm{u}}_{opt}) - \hat{\mathrm{u}}'_{opt}\mathrm{Z}\right)\right\}$$

This allocation scheme is robust against skewness and outliers. However, it does not perform as well for data with unequal sample sizes. This is because the vote of each *X* is either 0 or $\frac{1}{n_x}$ whereas the vote of each *Y* is 0 or $\frac{1}{n_y}$. Montanari (2004) abandoned this scheme for this very reason. This allocation scheme has also a problem of ties between $\mathbf{T}_x$ and $\mathbf{T}_y$ given in (4.2). The likelihood of ties is the greatest in the case of equal sample sizes, which happens to be the only situation where this scheme works efficiently. We will use random tie breaking where a *coin* is flipped to decide allocation in the case of a tie. The classifier obtained using this allocation scheme will be referred to as Point-Group Transvariation (PGT) classifier.

4.3 Symmetrized Allocation Based on Group-Group Transvariation

To allocate a new observation **Z**, we define $\mathbb{X}^* = \mathbb{X} \cup \{\mathrm{Z}\}$ and $\mathbb{Y}^* = \mathbb{Y} \cup \{\mathrm{Z}\}$. The idea is to find the transvariation probability between $\mathbb{X}^*$ and $\mathbb{Y}$ given by

$$\mathrm{T}_{x^*y} = \frac{1}{n_x(1+n_x)}\sum_{\mathrm{x}^*\in\mathbb{X}^*}\sum_{\mathrm{y}\in\mathbb{Y}} \phi\left\{\left(\hat{\mathrm{u}}'_{opt}\mathrm{x}^* - \hat{\mathrm{u}}'_{opt}\mathrm{y}\right)\left(m_x(\hat{\mathrm{u}}_{opt}) - m_y(\hat{\mathrm{u}}_{opt})\right)\right\}$$

and the transvariation probability between $\mathbb{X}$ and $\mathbb{Y}^*$ given by

$$\mathrm{T}_{xy^*} = \frac{1}{n_y(1+n_y)}\sum_{\mathrm{x}\in\mathbb{X}}\sum_{\mathrm{y}^*\in\mathbb{Y}^*} \phi\left\{\left(\hat{\mathrm{u}}'_{opt}\mathrm{x} - \hat{\mathrm{u}}'_{opt}\mathrm{y}^*\right)\left(m_x(\hat{\mathrm{u}}_{opt}) - m_y(\hat{\mathrm{u}}_{opt})\right)\right\}$$

and see the effect of the new observation on the quantities $\mathbf{T}_{x^*y}$ and $\mathbf{T}_{xy^*}$. We allocate the new observation to $\prod_x$ if $\mathbf{T}_{x^*y} < \mathbf{T}_{xy^*}$, else we classify it in $\prod_y$. Note that we do not have the unequal voting problem here. The vote of all observations is either 0 or approximately $n_x n_y$. The interested reader may find a discussion of this method is given in Nudurupati and Abebe (2009).

5.0 Application to Caribbean Food Detention Data

The data set contains information from FDA and USDA on the number of rejections by country for certain Latin American and Caribbean (LAC) countries for the years 1992 to 2003. This data set was investigated in Jolly et al. (2007) using zero-inflated count data mixed models. Data on exports from 14 selected LAC countries, FDI from US to these countries, and pesticide used were collected from USDA, WDI, and FAO data bases. The variables considered in the current study are:

- *t* = year (1992 - 2003)
- *FDI* = Foreign direct investment, net inflows (Balance of Payments (BoP), current US $)
- *Fert* = Fertilizer consumption (metric tons)

- *USImp* = U.S. Imports by Country, (1985-03; Millions of Dollars)
- *AgImp* = Total Agricultural Import to the US (million $)
- *GNI* = Gross national income per capita, Atlas method (current US $)
- *Y* = Detention Status (*Y*=0 no detention; *Y*=1 detention)

After missing values were removed, we were left with $n_0 = 75$ cases of no detentions and $n_1 = 67$ cases of detentions. We use the procedures mentioned in the Section 3.3 to select first four desirable variables. Table 1- 3 gives the results.

FDI appears to be the most important variable in determining the likelihood of food detention. Considering the interaction of dissimilarity to *FDI* and overlap between the two groups, fertilizer consumption is the second most important variable determining food detention. We stop further variable selection since none of the remaining variables give criteria that exceed our selection threshold of $\alpha = 0.10$. *GNI* is in particular not very useful in this context since $\tilde{T}(\mathrm{GNI})$ is close to 0.

We will consider GGT, TD, LDF, and PGT. We will also consider maximum depth classifier, MaxD, based on L_1 depth given by Ghosh and Chaudhuri (2005) and Jörnsten (2004). First, we will use all 6 variables compute the leave-one-out cross-validation misclassification error rate of all the aforementioned methods of classification. We then *select* the two most informative variables given in Table 1 and perform a leave-one-out cross-validation to compute the rate of misclassification error. Table 3 gives the results.

At 38% and 38.7% misclassification percentages, LDF seems to give inferior performance to all the others considered. MaxD gives the best performance before variable selection and GGT and PGT gives the best performance after variable selection. GGT and PGT improve after variable selection while all the other methods give the same misclassification error rates or perform worse than the 6D case.

6.0 Conclusion

A number of nonparametric discriminant analysis procedures are studied. The development of the nonparametric discriminant analysis procedures given in Montanari (2004) and Nudurupati and Abebe (2009) are reviewed. The procedures use the idea of projection pursuit to measure group separation of multivariate data using the two-group transvariation probability (Gini, 1916). Allocation of new observations using the symmetrized method of Nudurupati and Abebe (2009) provides optimal classification when training samples are drawn from heavy tailed or skewed distributions and when the difference between the training sample sizes is large.

These methods were applied to evaluate determinants of food detention from certain Latin American and Caribbean countries. Variables were selected based on their ability to discriminate between cases of detention and no detention. It was found that foreign direct investment was the most informative variable whereas gross national income was the least informative out of the variables considered in the study. Using leave-one-out cross-validation, the maximum depth classifier of Jörnsten (2004) gave the most optimal results for the original data. However, after removing the non-informative variables, procedures based on transvariation probabilities gave the best results. Linear discriminant analysis gave the worst result of all the methods in both cases.

When there is some stochastic component in a system, no variable is the best variable in explaining another unless the variables are connected via a deterministic relationship. For that reason, we do not contend that the variables considered in this study are the most informative of all variables that could have been taken into account to explain the probability of food detention. The fact that none of the methods used in this study (existing and new) gave a misclassification error rate that is less than 30% is indicative of the inherent overlap that exists in the two samples and not necessarily a failure of all the methods. In other words, high misclassification error rates may be unavoidable for this particular sample regardless of the procedure used. This paper proposed a procedure that gave better performance than some of the better known existing methods given the same, possibly lossy, information. The results should be interpreted rather narrowly in the sense that the procedure only provides a way for scientists to effectively prioritize the variables in their *existing* data. It does not give a way to pick the best variable out of all possible variables.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. DMS-0604726. The authors acknowledge insightful discussions with Prof. Curtis Jolly.

References

Abebe, A. and Nudurupati, S. V. 2009. Rank-based classification using robust discriminant functions. *Comm. Statist.- Sim. Comp.*, **38**(2):199 – 214.

Anderson, T. W. 1984. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition.

Billor, N., Abebe, A., Turkmen, A., and Nudurupati, S. V. 2008. Classification based on depth transvariations. *J. Classification*, 25:249–260.

Calò, D. G. 2006. On a transvariation based measure of group separability. *J. Classification*, 23(1):143–167.

Chen, Z.-Y. and Muirhead, R. J. 1994. A comparison of robust linear discriminant procedures using projection pursuit methods. In *Multivariate analysis and its applications (Hong Kong, 1992)*, volume 24 of *IMS Lecture Notes Monogr. Ser.*, pages 163–176. Inst. Math. Statist., Hayward, CA.

Dillon, W. R. 1979. The performance of the linear discriminant function in nonoptimal situations and the estimation of classification error rates: A review of recent findings. *Journal of Marketing Research*, **16**:370–381.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, VII(II):179–188.

Friedman, J. H. and Tukey, J. W. 1974. Projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C **23**(9):881–890.

Ghosh, A. K. and Chaudhuri, P. 2005. On maximum depth and related classifiers. *Scand. J. Statist.*, **32**(2):327–350.

Gini, C. 1916. Il concetto di transvariazione e le sue prime applicazioni. *Giornale degli economisti Rivista di statistica*.

Hills, M. 1967. Discrimination and allocation with discrete data. *Applied Statistics*, **16**(3):237–250.

Hollander, M. and Wolfe, D. A. 1999. *Nonparametric statistical methods*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York, second edition. A Wiley-Interscience Publication.

Johnson, M. E., Wang, C., and Ramberg, J. S. 1979. Robustness of fisher's linear discriminant function to departures from normality. In *Los Alamos Techinical Report LA-8068-MS*. Los Alamos, NM 87545.

Jolly, C. M., Namugabo, E., and Abebe, A. 2007. Food safety issues between Latin American and Caribbean countries. *Paper presented at the 27th West Indies Agricultural Economics Conference*.

Jörnsten, R. 2004. Clustering and classification based on the *L*1 data depth. *J. Multivariate Anal.*, **90**(1):67–89.

Lachenbruch, P. A. 1975. *Discriminant analysis*. Hafner Press, New York.

Lachenbruch, P. A., Sneeringer, C., and Revo, L. T. 1973. Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Comm. Statist.*, **1**(1):39–56.

McLachlan, G. J. 1992. *Discriminant analysis and statistical pattern recognition*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Montanari, A. 2004. Linear discriminant analysis and transvariation. *J. Classification*, **21**(1):71–88.

Nudurupati, S. V. and Abebe, A. 2009. A nonparametric allocation scheme for classification based on transvariation probabilities. *J. Statist. Comp. Simul.*, **79**(8):977–987.

Seber, G.A. F. 1984. *Multivariate Observations*. John Wiley, New York.

Tukey, J. 1974. Address to international congress of mathematicians. Vancouver.

Table 1: Choosing first two variables

	t	FDI	Fert	USImp	AgImp	GNI
$\tilde{T}(V_s)$	.166	.552	.400	.521	.382	.063
$\tilde{T}(V_s \mid FDI)$	.548		.256	.177	.259	.361
$\tilde{T}(V_s)\tilde{T}(V_s \mid FDI)$	.091		.103	.092	.099	.023

Source: Compiled by authors

Table 2: Choosing the third variable

	FDI & Fert	*t*	USImp	AgImp	GNI
$\tilde{T}(V_s)$	.556	.166	.521	.382	.063
$\tilde{T}(V_s \mid FDI,Fert)$		.546	.174	.256	.363
$\tilde{T}(V_s)\tilde{T}(V_s \mid FDI,Fert)$		.090	.091	.098	.023

Source: Compiled by authors

Table 3: Percentage of Observations Misclassified
(Leave-one-out Cross-Validation)

	GGT	PGT	TD	MaxD	LDF
6*D*	34.5	35.9	33.1	31.7	38.0
2*D*	30.3	30.3	33.1	31.7	38.7

Source: Compiled by authors