International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN (P): 2319-3972; ISSN (E): 2319-3980
Vol. 5, Issue 2, Feb - Mar 2016; 87-98
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A NONPARAMETRIC DISCRIMINANT VARIABLE-SELECTION ALGORITHM FOR CLASSIFICATION TO TWO POPULATIONS

S. PADMANABAN[1] & MARTIN L. WILLIAM[2]

[1]NIRRH Field Unit, I.C.M.R, KMC Hospital, Chennai, India

[2]Department of Statistics, Loyola College, Chennai, India

ABSTRACT

This paper proposes a nonparametric discriminant variable-selection algorithm to discriminate two multivariate populations and an associated optimal decision rule for membership-prediction. The present work relaxes the 'equal variance-covariance matrices' condition traditionally imposed and develops a discrimination-classification procedure by including variables that best contribute to the 'discrimination', one-by-one in a forward-stepwise manner. The inclusion of variables in the discriminant is determined on the basis of best 'discriminating ability' as reflected in 'maximal difference' between the distributions of the discriminant in the two populations. A new decision-rule for classification or membership-prediction with a view to maximize correct predictions is provided. The proposed algorithm is applied to develop an optimal discriminant for predicting preterm labour among expecting mothers in the city of Chennai, India, and its performance is compared with logistic regression.

KEYWORDS: Classification, Discriminant, Kolmogorov-Smirnov Statistic

1. INTRODUCTION

Discriminant Analysis for discriminating two multivariate populations and classification of members to the two has been in existence for many decades now. Traditionally, for application of the technique to a distribution free (non-parametric) context, the condition of 'equal variance-covariance matrices of the populations' is imposed, although this condition is not needed for multivariate normal populations. The inclusion of any variable in the discriminant is based on a comparison of its means in the two populations. Further, membership-prediction or classification of members to the two populations is based on the 'distances' with the means of the discriminant in the two populations. The objective of the present work is to develop a simple algorithm to build an efficient discriminant model for two populations, with wider scope of application.

It is well known that in many real life situations involving observation of multiple variables in two populations, multivariate normality or equality of variance-covariance matrices is not guaranteed. In case of multivariate normality, test for equality of the variance-covariance matrices can be carried out and, if equality is verified, one can proceed with the traditional linear discriminant function; in the unequal case, quadratic discriminant function can be applied. In non-multivariate-normal situations with equal variance-covariance matrices, the distribution free Fisher's linear discriminant function is applicable but there is no easy procedure for testing the equality of the variance-covariance matrices in such situations. In most applications, practitioners tend to 'assume' equality and proceed. The need to fill this 'gap' and provide a discrimination-classification procedure in a distribution-free setting without the condition of equal variance-covariance matrices is the motivation for the present work. We aim to develop a theoretical framework and a simple tool leading to an efficient procedure which can be applied without the conditions that restrict the existing methods.

Modifications and advancements to the classical theory of discriminant analysis have been the topic of research of a number of authors over the past many years. Innovative approaches to develop discriminant models have been provided by a good number of authors with focus on identifying the important variables for discrimination of the populations. Among the notable early works in this context are the approach given by Chang (1983) using principal components in the context of separating a mixture of two multivariate normal distributions and that of Bensmail and Celeux (1996) who considered Gaussian discriminant analysis through eigen-value decomposition. A stepwise algorithm involving the use of 'Bayesian Information Criterion' was developed by Murphy *et al.* (2010) following the ideas of Raftery and Dean (2006) who proposed a similar approach for model-based clustering. The above-referred approaches are parametric and restricted in their applicability.

Other scholarly works on the topic extends discriminant analysis to non-parametric settings in different directions. Nonparametric versions of discriminant analysis with nonlinear classification schemes were obtained by Hastie *et al.* (1994) in the presence of a large number of predictor variables. Nonlinear discriminant analysis using a kernel approach which is theoretically close to support vector machines was presented by Baudat and Anouar (2000). Nonparametric discriminant analysis with adaptation to nearest-neighbour classification was presented by Bressan and Vitria (2003). Chiang and Pell (2004) combined genetic algorithms with discriminant analysis for identifying key variables. A primary concern in most of the above-mentioned works was on identifying the variables that would enable effective discrimination between the populations.

This paper takes an approach different from that of the other approaches present in the literature on two-population discriminant analysis while, at the same time, sticking to the basic spirit and mathematical objective of classical discriminant analysis. A 'model performance' measure for the discriminant model reflecting the ability of the derived 'discriminant' to maximally differentiate the two populations in a non-parametric paradigm is suggested. A variable-selection algorithm is presented to build the discriminant model by selecting variables that best contribute to the discrimination-ability one-by-one in a forward-stepwise manner. A decision rule for identifying the optimal cut-off point for classification or membership-prediction with the objective of maximizing correct classifications is provided.

Hence, the objectives of the present work are:

I. To provide a theoretical framework for handling situations with unequal variance-covariance marices.

II. To suggest a 'model performance' measure for judging discriminant models.

III. To present a variable-selection algorithm for discriminating two populations and an easy-to-apply procedure for classification of objects.

IV. To apply the algorithm to a biomedical phenomenon and compare its classification-performance with that of logistic regression.

This paper is organized as follows: Following this introductory section, the basic theoretical framework required for further development is presented in Section 2. The specific theoretical aspects for the two-population discrimination context and the derivation of the optimal discriminant function are given in Section 3. The motivation and the proposition of the new model performance measure is given in Section 4. The new variable-selection algorithm to build an efficient discriminant model is outlined in Section 5. As an immediate real-life application of the proposed methodology, the prediction of 'pre-term labour' in pregnant women is considered. Section 6 discusses the phenomenon of pre-term labour, a potential risk that pregnant woman face and the possible factors associated with the phenomenon. In Section 7, the proposed algorithm of discriminant model building is applied to predict pre-term labour using a sample of 200 women who delivered babies in Department of Obstetrics and Gynaecology, Government Kilpauk Medical College and Hospital, Chennai, India, during the five-month period of 7th May, 2015 to 7th October, 2015.

2. THEORETICAL FRAMEWORK

Consider two populations denoted as π_1 and π_2. The objects in the two classes are to be classified on the basis of measurements on a random vector, say, $X = (X_1, X_2... X_p)^T$. If the spread of values and the distribution of X were '*not substantially*' different for objects in π_1 and π_2, then there would not be any effective discrimination between π_1 and π_2 and any attempt to carry out classification of objects would not be fruitful. But, when there is a '*significant'* difference between the distributions, classification or membership-prediction becomes relevant and the 'correctness' or 'incorrectness' of classifications turns out to be a material issue.

Denote the mean-vectors of X in the two populations as $\mu_1 = E_1(X)$ and $\mu_2 = E_2(X)$ and the variance-covariance matrices of X in the two populations be Σ_1 and Σ_2. In the sequel, we derive a general expression for the mean vector and variance-covariance matrix of the 'combined' population π_1 U π_2. Towards this, we present a lemma, to derive the general expressions for mean vector and variance-covariance matrix of a random vector X in terms of conditional mean vector and conditional variance-covariance matrix of X given a random object W. We denote the Expectation and Variance-Covariance matrix operators under the unconditional distribution of X as $E_X(\cdot)$ and $V_X(\cdot)$. The operators under the conditional distribution of X given W shall be denoted as $E_{X|W}(\cdot)$ and $V_{X|W}(\cdot)$. The corresponding operators under the distribution of W shall be $E_W(\cdot)$ and $V_W(\cdot)$. The relationship for mean vectors is well known to be $E(X) = E_W[E_{X|W}(X)]$ while that for variance-covariance matrices is not well known. So we present a lemma deriving this relationship.

Lemma 2.1: For a random vector X and another random object W, the relationship between the unconditional and unconditional variance-covariance matrices is given by

$$E(X) = E_W[E_{X|W}(X)] \text{ and } V(X) = E_W\{V_{X|W}(X)\} + V_W\{E_{X|W}(X)\} \quad (2.1)$$

Proof: Consider $V_X(X) = E_X(X \cdot X^T) - E_X(X) \cdot E_X(X^T)$

$$= E_W[E_{X|W}(X \cdot X^T)] - E_W[E_{X|W}(X)] \cdot E_W[E_{X|W}(X^T)]$$

$$= E_W[E_{X|W}(X \cdot X^T)] - E_W[h(W)] \cdot E_W[h(W)]^T \quad (2.2)$$

Where $h(W) = E_{X|W}(X)$

Now, $V_{X|W}(X) = E_{X|W}(X \cdot X^T) - E_{X|W}(X) \cdot E_{X|W}(X^T) = E_{X|W}(X \cdot X^T) - [h(W)] \cdot [h(W)]^T$ so that

$$E_{X|W}(X \cdot X^T) = V_{X|W}(X) + [h(W)] \cdot [h(W)]^T \quad (2.3)$$

Using (2.3) in (2.2), we get

$$V_X(X) = E_W\{V_{X|W}(X) + [h(W)] \cdot [h(W)]^T\} - E_W[h(W)] \cdot E_W[h(W)]^T$$

$$= E_W\{V_{X|W}(X)\} + E_W\{[h(W)] \cdot [h(W)]^T\} - E_W[h(W)] \cdot E_W[h(W)]^T$$

$= E_W \{V_{X|W}(X)\} + V_W \{h(W)\}$

$= E_W \{V_{X|W}(X)\} + V_W \{E_{X|W}(X)\}$

Hence the Lemma.

Next we present a lemma to derive an expression for the 'overall' variance-covariance matrix of the combined population in the context of two-population discriminant analysis. In this context, the random object W has two possible 'values' π_1 and π_2 with probabilities p_1 and p_2 representing the relative sizes of the two populations. Then, we have $E_{X|W}(X \mid \pi_1) = \mu_1$ and $E_{X|W}(X \mid \pi_2) = \mu_2$ and also, $V_{X|W}(X| \pi_1) = \Sigma_1$ and $V_{X|W}(X| \pi_2) = \Sigma_2$. With these, we have the following lemma.

Lemma 2.2: The 'overall' variance-covariance matrix of the combined population is

$$\Sigma = p1\Sigma1+ p_2\Sigma_2+ p_1(1-p_1)\, \mu_{1.}\, \mu_1^T + p_2(1-p_2)\, \mu_{2.}\, \mu_2^T - p_1\, p_2(\mu_1\, \mu_2^T + \mu_2\, \mu_1^T)\ .. \qquad ..(2.4)$$

Proof: In the two-population discriminant analysis context, the random object W has the following "two-valued" distribution: $\pi_{1\,\text{with}}$ probability p_1 and $\pi2$ with probability p_2.

Now, $h(W) = E_{X|W}(X)$ assumes two (vector) values: $\mu_{1\,\text{with}}$ probability p_1 and μ_2 with probability p_2.

The overall mean vector of X in the combined population is $E(X) = E_W [h(W)] = p_1\, \mu_1 + p_2\, \mu_2$

Also, $E_W \{[h(W)]\cdot[h(W)]^T\}= p_1\, (\mu_1\cdot \mu_1^T) + p_2(\mu_2\cdot\mu_2^T)$ (2.5)

and $E_W[h(W)]\cdot E_W[h(W)]^T = (p_1\, \mu_1 + p_2\, \mu_2)\cdot(p_1\, \mu_1 + p_2\, \mu_2)^T$

$$= p_1^2\, (\mu_1\cdot \mu_1^T) + p_1\, p_2\, (\mu_1\cdot\mu_2^T + \mu_2 \cdot\mu_1^T) + p_2^2\, (\mu_2\cdot \mu_2^T) \qquad (2.6)$$

So, $V_W \{E_{X|W}(X)\} = p_1\, (\mu_1\cdot \mu_1^T) + p_2(\mu_2\cdot\mu_2^T) - (p_1\, \mu_1 + p_2\, \mu_2)(p_1\, \mu_1 + p_2\, \mu_2)^T$

$$= p_1\, (1-p_1)\, \mu_{1.}\, \mu_1^T + p_2\, (1-p_2)\, \mu_{2.}\, \mu_2^T - p_1\, p_2\, (\mu_1\, \mu2T + \mu_2\, \mu1T) \qquad (2.7)$$

Next, conditional on W, $V_{X|W}(X)$ is a (random) matrix assuming two possible 'values' namely Σ_1 with probability p_1 and Σ_2 with probability p_2. So,

$$E_W\{V_{X|W}(X)\} = p_1\, \Sigma_1 + p_2\, \Sigma_2 \qquad (2.8)$$

Referring to (2.1) and using (2.7) and (2.8), we get the expression for Σ in (2.4). Hence the lemma.

3. TWO-POPULATION DISCRIMINATION AND OPTIMAL DISCRIMINANT FUNCTION

In Discriminant Analysis, the Multivariate observations (X) are transformed to univariate observations (Y) by considering linear combinations of the X_i's. Any linear combination of the X_i's may be expressed as $Y = \ell^T X$ where ℓ is a p x 1 vector of constants. It is easily seen that the means of Y in the two populations are $\mu_{1Y} = \ell^T\mu_1$ and $\mu_{2Y} = \ell^T\mu_2$ and in the combined population it is given by $\mu_Y = p_1\, \ell^T\mu_1 + p_2\, \ell^T\mu_2$. And, the variance of Y in the combined population is given by $V(Y) = \ell^T\, \Sigma\, \ell$.

The linear combination which maximizes the (squared) distance between μ_{1Y} and μ_{2Y} relative to the variability in Y helps in discriminating the two groups in the most 'optimal' manner. In the classical Fisher's Linear Discriminant Analysis, the distance was measured relative to the 'common' variability in Y in the two populations. As the objective of the present work is towards developing the 'optimal' discriminant function in the 'unequal variance-covariance matrices' context, and as distances need to be measured between objects that belong to either of the two populations, the distance measurement will be more meaningful if it is measured relative to the 'overall' variability in the combined population. The 'distance-maximizing' linear combination of the X_i's is the 'optimum discriminant function' based on X. We call it 'X-based optimal discriminant'.

Derivation of the Optimal Discriminant Function

Denote any linear combination of the underlying random vector X to be used for Discriminating the populations as $Y = \ell^T X$. The (squared) distance between the means of Y in the two populations relative to the overall variability in Y in the combined population is given by

$$\frac{Squared\ dis\tan ce\ between\ the\ means\ of\ Y}{Var\ (Y)} = \frac{(\mu_{1Y}-\mu_{2Y})^2}{\ell^T\ \Sigma\ \ell} = \frac{(\ell^T\ \delta)^2}{\ell^T\ \Sigma\ \ell} \quad \text{where } \delta = \mu_1 - \mu_2$$

This ratio is to be maximized to get the optimal Discriminant function. By an application of Cauchy-Schwartz inequality, the maximum of the above ratio is attained when $\ell = c\ \Sigma^{-1}\delta$ (for any choice of non-zero scalar 'c'). Choosing c = 1, we get the X-based optimal discriminant as

$$Y = (\Sigma^{-1}\delta)^T X = \delta^T\Sigma^{-1}X = (\mu_1 - \mu_2)^T\Sigma^{-1}X \tag{3.1}$$

Typically, the true mean vectors and the variance-covariance matrix are unknown and so, they are replaced by the sample estimates. The optimal discriminant function converts the two multivariate populations π_1 and π_2 into univariate populations such that the corresponding univariate population means are separated 'as much as possible' relative to the overall variance in the combined population. In the proposed approach, we impose the criterion that, this 'maximal' differentiation is reflected in a 'significant' difference between the distributions of the discriminant scores in the two populations. as measured by the two-sample Kolmogorov-Smirnov Statistic

3. MODEL-PERFORMANCE MEASURE

In a practical application, typically the investigators measure a number of variables that they view as important for their study. But for the purpose of discrimination between two groups and classification (or membership-prediction), some of the variables may be irrelevant. It would be pertinent to build the 'optimal discriminant model' developed in Section 3, with only a subset of the variables observed. Thus, a need to compare the 'optimal discriminant models' built on different subsets of the variables arises and hence, a measure to compare the models becomes essential. As stated in the previous section, the 'optimal discriminant function' must be capable of maximally differentiating the two populations. This 'differentiation' is reflected in the two-sample Kolmogorov-Smirnov Statistic which measures the 'distance' between the distribution functions of the discriminant in the two populations. In the sequel, the proposed model-performance measure is developed:

Suppose $X_{(s)}$ be a subset of the variables used to build the optimal discriminant. Denote the mean vectors of $X_{(s)}$ in the two populations as $\mu_{1(s)}$ and $\mu_{2(s)}$ and the 'overall' variance-covariance matrix of $X_{(s)}$ as $\Sigma_{(s)}$. Proceeding as in Section 3, we get the $X_{(s)}$-based optimal discriminant as

$$Y_{(s)}= (\mu_{1(s)} - \mu_{2(s)})^T\Sigma_{(s)}^{-1}X_{(s)} \tag{4.1}$$

Typically, these parameters are replaced by the sample estimates in practice. Computing the variable $Y_{(s)}$ for all members in both the samples, the performance of the $X_{(s)}$-based optimal discriminant is measured by the two samples Kolmogorov-Smirnov Statistic based on the $Y_{(s)}$ measurements. Denoting the (empirical) cumulative distribution functions of $Y_{(s)}$ for the two populations as $F_{1(s)}(\cdot)$ and $F_{2(s)}(\cdot)$, the performance measure is given by

$$KS_{(s)} = \max_y \left(| F_{1(s)}(y) - F_{2(s)}(y) | \right) \tag{4.2}$$

Given two subvectors $X_{(s1)}$ and $X_{(s2)}$, the optimal $X_{(s1)}$-based discriminant is said to be 'more efficient' than the optimal $X_{(s2)}$-based discriminant if $KS_{(s1)} > KS_{(s2)}$. If there exists a random subvector $X_{(s*)}$ for which $KS_{(s*)} > KS_{(s)}$ for every other random subvector $X_{(s)}$, then the corresponding optimal discriminant $Y_{(s*)}$ is the 'most efficient' discriminant.

However, obtaining the 'most efficient' discriminant is computationally prohibitive in the presence of a very large number of predictor variables (i.e.) in case of very high dimension of the underlying random vector X. This is true of every model-building situation involving a large number of predictor variables and different algorithms are therefore suggested to 'build' improved models sequentially instead of considering 'all possible' models or identifying the 'most efficient'.

In the same spirit, the next section presents a model building algorithm to build a 'sequence' of models leading to an efficient discriminant model.

5. THE PROPOSED VARIABLE-SELECTION ALGORITHM

The proposed Algorithm evaluates each candidate ‘input’ variable for discriminatory capacity in a sequential manner towards constructing the optimal discriminant function. Variable-selection for discriminating between two populations has been addressed in the past too. Interesting references in this context are the papers of Habbema and Hermans (1977) in which selection of variables for Gaussian discriminant analysis was on the basis of F-Statistics and error rates and that of Pfeiffer (1985) wherein smoothing factors of kernel functions for nonparametric discriminant analysis were considered and different criteria like distances, error rates and density-ratios were used for variable selection.

Here, we propose a different route to variable-selection in a forward-stepwise manner. The algorithm proceeds by bringing one input variable at a time on the basis of maximal differentiation between the distributions of the discriminant scores in the two populations, as measured by the two sample Kolmogrov-Smirnov (KS) statistic used for comparison of two distributions. The exact stepwise process is described below.

Let $X_1, X_2, \ldots, X_p$ be the candidate input variables.

Step 1: With one variable at a time, ‘p’ discriminants $Y_{(1)}, Y_{(2)}, \ldots, Y_{(p)}$, where $Y_{(i)}$ is the discriminant based on single input variable X_i, and their corresponding scores are obtained for each individual record in the data. Let the Kolmogorov-Smirnov Statistic for $Y_{(i)}$ is denoted as $KS_{(i)}$. If

$KS_{(i)} > KS_{(j)}$ for every $j \neq i$

Then among the individual variables considered on a one-at-time basis, X_i is the top discriminator between the two populations. The significance of this $KS_{(i)}$ statistic is evaluated and if found significant at a desired level, X_i first 'enters' the model and model building continues.

Step 2: With X_i having been already selected, we take one additional variable at a time and obtain (p–1) discriminants having input-pairs $(X_1,X_i),....,(X_{i-1}, X_i),(X_{i+1},X_i),....,(X_p, X_i)$. Denote the discriminants as $Y_{(1,i)}$, $Y_{(2.i)}$,..., $Y_{(i-1,i)}$, $Y_{(i+1,i)}$, ... $Y_{(p,i)}$ and the corresponding Kolmogorov-Smirnov statistics as $KS_{(1,i)}$, $KS_{(2,i)}$,...,$KS_{(i-1,i)}$, $KS_{(i+1,i)}$,..., $KS_{(p,i)}$. If for some 'm',

$KS_{(m,i)} > KS_{(j,i)}$ for every $j \neq m$, and $KS_{(m,i)} > KS_{(i)}$,

Then X_m enters the model as the second variable. It is to be noted that the significance of $KS_{(m, i)}$ is guaranteed because of the significance of $KS_{(i)}$ in the first step. In contrast, if

$KS_{(m, i)} > KS_{(j, i)}$ for every $j \neq m$, but $KS_{(m, i)} \leq KS_{(i)}$,

Then X_m does not enter the model, nor any of the remaining X_j's enters, as its entry leads to reduced discriminatory ability and the model building stops with only one input variable. Clearly no other variable can enter.

At every subsequent step that is considered, one more additional variable enters provided the maximum KS value at that step exceeds the maximum KS value of the previous step. If it is equal to or less than the previous maximum, the process stops. When the process stops at the (k+1) th step, the optimal discriminant function is the one obtained in the kth step with the maximum KS value, leading to significant and maximum discrimination between the two populations. We denote the final subset of variables reached in this process as $X_{(S*)}$ and the 'final' efficient discriminant as $Y_{(s*)}$.

Classification or Prediction Rule

The classification or prediction rule to allocate an object to one of the two populations is based on the optimal cut point at which the KS statistic value is attained. Let y_0 be the point such that

$$KS_{(s*)} = \max_{y} \left(| F_{1(s*)}(y) - F_{2(s*)}(y) | \right) = | F_{1(s*)}(y_0) - F_{2(s*)}(y_0) |$$

This point y_0 gives maximum differentiation between the distributions of the $Y_{(s*)}$ scores in the two populations and is the 'efficient cut-point'.

Now, let the means of the final efficient discriminant $Y_{(s*)}$ in the two populations π_1 and π_2 be denoted as $\mu_{1Y(s*)}$ and $\mu_{2Y(s*)}$and, let $\mu_{1Y(s*)} > \mu_{2Y(s*)}$. For membership-prediction, we proceed as follows:

If $y_{(s*)}$ is the value of the final efficient discriminant $Y_{(s*)}$ for an object, then the following classification rule is to be applied:

$$\text{Classify object to: } \begin{cases} \pi_1 \;\; if \;\; y_{(s*)} > y_0 \\ \pi_2 \;\; if \;\; y_{(s*)} \leq y_0 \end{cases}$$

With the above classification criterion, we observe that $F_{1(s*)}(y_0)$ gives the proportion of π_1 objects wrongly classified to π_2 and $F_{2(s*)}(y_0)$ gives the proportion of π_2 objects correctly classified to π_2 using the cut-point y_0. And, the final Kolmogorov-Smirnov statistic $KS_{(s*)}$ can be 'explained' as follows:

$KS_{(s*)}$ = |Proportion of π_1 objects misclassified − Proportion of π_2 objects correctly classified |

= |Proportion of π_2 objects correctly classified−(1− Proportion of π_1 objects correctly classified) |

= |(Proportion of π_1 objects correctly classified + Proportion of π_2 objects misclassified) − 1|

It is easily seen that, higher value of the $KS_{(s*)}$ statistic indicates higher proportion of correct classifications to both π_1 and π_2.

We note that, considering any point other than the efficient cut-point y_0 would lead to an overall reduction in the proportion of correct classifications. It is also interesting to note that, instead of working with the cumulative distribution functions, we can work with the 'reliability functions' and use the absolute difference of these as the Kolmogorov-Smirnov Statistic, leading to equivalent results. We note that evaluating the KS statistic through reliability function would require the descending order arrangement of the discriminant scores in contrast to the usual method of ascending order arrangement.

6. THE PHENOMENON OF PRETERM LABOUR

Preterm Labour

The lifestyle changes brought about by technological revolution, the job nature of the younger generation people and careless food habits have brought in many health-related disorders among youngsters. Women are not free from this problem and in the case of married women this results in pregnancy-related issues and delivery-complications. Giving birth to the baby ahead of the normal delivery deadline is a serious anomaly which can affect the child’s growth milestones and create other physical difficulties for the child for life.

Preterm labour is defined as the presence of uterine contractions of sufficient frequency and intensity to effect progressive effacement and dilatation of the cervix prior to term gestation between 20-37 weeks.

Incidence

Incidence of preterm birth (PTB) has been found to be 12% of all deliveries and accounts for a majority of neonatal deaths and nearly half of all cases of congenital neurological disability, including cerebral palsy.

Of all preterm births that occur, 40 - 45% result from onset of labour, 25 - 30% result from preterm premature rupture of membranes (PPROM) and 30 - 35% are medical decisions. A PTB resulting from labour or PPROM are referred to as spontaneous PTB (SPTB).

Potential Factors Associated with Preterm Labour

During pregnancy, maternal lipids are important both for steroid genesis of the mother, placenta and fetus as well as for fetal growth. It may reflect an individual’s predisposition to develop metabolic syndrome. Metabolic syndrome is a disorder marked by increased inflammation, a reported pathway for preterm labour. For research relating lipid profile to preterm labour, reference is made to the article of Mudd *et al.* (2012) among others. Dyslipidemia has been suggested to be one pathway that explains why women at risk for preterm labour are also at risk for developing cardiovascular disease later in life. We refer to Catov *et al.* (2007) in this context. In this study, we consider the following factors (Lipid Profiles):

Lipid Profiles

1. Total Cholesterol (TC)

Directly linked to risk of heart and blood vessel disease.

2. High Density Lipoprotein (HDL) “Good cholesterol”

   High levels linked to a reduced risk of heart and blood vessel disease. The higher the HDL, the better.

3. Low Density Lipoprotein (LDL) “Bad cholesterol”

   High levels are linked to an increased risk of heart and blood vessel disease, including coronary artery disease, heart attack and death. Reducing LDL levels is a major treatment target for cholesterol-lowering medications.

4. Triglycerides (TGL)

   This is elevated in obese or diabetic patients. Level increases from eating simple sugars or drinking alcohol. Associated with heart and blood vessel disease.

   In addition to the above factors, we consider the following:

5. Amniotic fluid index (AFI)

   This is an estimate of the amount of amniotic fluid and is an index for the fetal well-being.

6. Prepregnancy Body Mass Index (BMI)

   A general indication of physical health-being for anyone including expecting women.

Objective: This study aims to relate the above factors to preterm labour through the discriminant analysis model developed in the earlier sections of this paper. We wish to identify the significant factors that are associated to the risk of preterm labour.

Study Design: Cross sectional comparative study

Sample Size: Study group (women with spontaneous preterm labour) is 100; Comparative group (women with term labour) is 100.

PREDICTION OF PRETERM LABOUR THROUGH EFFICIENT DISCRIMINANT ANALYSIS

A sample of the data on the six variables listed under 'Potential factors' along with the birth outcome (Term labour = 1, sPTB = 2) is given below:

Table 1

Record #	X_1 (BMI)	X_2 (AFI)	X_3 (TC)	X_4 (TGL)	X_5 (HDL)	X_6 (LDL)	Outcome
1	12.6	14.2	274	168	76	114	1
2	19.3	9.5	276	288	89	186	2
3	12.6	14.2	235	168	76	114	1
4	12.6	14.2	274	168	76	114	1
5	19.7	9.6	310	298	89	186	2

We apply the variable-selection algorithm developed in this paper and get the following results.

Step 1: The KS statistics for models with single variables are found to be

$KS_{(X1)} = 0.210$, $KS_{(X2)} = 0.400$, $KS_{(X3)} = 0.880$, $KS_{(X4)} = \mathbf{0.920}$, $KS_{(X5)} = 0.250$, $KS_{(X6)} = 0.800$

X_4 enters the model in the first step. The KS value of 0.920 is found to be statistically significant.

Step 2: The KS statistics for models with one additional variable with X_4 are found as

$KS_{(X1, X4)} = 0.910$, $KS_{(X2, X4)} = 0.930$, $KS_{(X3,X4)} = \mathbf{0.960}$, $KS_{(X5,X4)} = 0.470$, $KS_{(X6,X4)} = 0.920$

X_3 enters the model in the second step.

Step 3: In this step we get

$KS_{(X1, X3, X4)} = 0.950$, $KS_{(X2, X3, X4)} = \mathbf{0.980}$, $KS_{(X5, X3, X4)} = 0.210$, $KS_{(X6, X3, X4)} = 0.930$

X_2 enters the model in the third step.

Step 4: In this step we get

$KS_{(X1, X2, X3, X4)} = 0.980$, $KS_{(X5, X2, X3, X4)} = 0.980$, $KS_{(X6, X2, X3, X4)} = 0.960$

As none of the latest KS statistics exceeds the previous maximum KS value, the variable selection algorithm stops with three variables being selected in the order of X_4, X_3 and X_2.

The 'Efficient Discriminant' obtained at the end of Step 3 of our algorithm is:

$$Y = 0.1853{*}AFI - 0.0343{*}TC - 0.0228{*}TGL \qquad (7.1)$$

The estimated means of Y in the two populations are found to be

$$\mu_{1Y} = -11.3522,\ \mu_{2Y} = -\ 14.6097$$

and the 'efficient cut-point' is $y_0 = -12.578$

Here, '1' denotes 'term labour group' and '2' denotes 'sPTB group'.

<u>Membership-Prediction Rule</u>: If 'y' denotes the measured value of the 'Efficient Discriminant' Y of (7.1) for an individual, then the prediction rule is as follows:

$$\text{Classify individual to: } \begin{cases} Term\ \ Labour\ \ Group & if\ y > -12.578 \\ \Pr eTerm\ Labour\ Group & if\ y \leq -12.578 \end{cases}$$

We observe form (7.1) that, increased AFI, lower TC and lower TGL indicate the likelihood of normal term labour for a woman. Accordingly, we find that lower AFI, higer TC and higher TGL increase the risk for preterm labour for a woman.

Comparison with Logistic Regression Model:

Denoting 'preterm labour outcome' as the outcome of interest, we build a logistic regression model using the stepwise method of model building.

Step 1: TC entered with very high significance and with a positive coefficient.

Step 2: TGL entered with very high significance and with a positive coefficient.

The model building process stops with this and we have the following logit equation from the model:

$$\log\left(\frac{p}{1-p}\right) = -\,59.154 + 0.119{*}\mathrm{TC} + 0.108{*}\mathrm{TGL}$$

where 'p' is the probability of preterm labour. The KS for this model is found to be 0.950 which is less than the KS obtained for the 'Efficient Discriminant Model'. Thus, the new method performs better than binary logistic regression method in predicting preterm labour among pregnant women.

It is also interesting to note that, while logistic regression identifies two factors TC and TGL, our model captures one more important factor AFI. In this context we refer to the article of Weismann-Brenner *et al.* (2009) in which it was stated that the mean AFI differs significantly between PPROM (PTB) cases and the normal cases. Our discriminant model confirms that AFI is an important discriminator between preterm and term labour cases and that lower AFI points to the risk of preterm labour. The finding here supports the discovery of the medical research team of Brenner *et al.*

We emphasize that, though the new approach needs to be applied to many more situations wherein logistic regression is applied to decide its effectiveness in prediction of 'binary' outcomes, the present findings suggest that this approach is a promising alternative to logistic regression model. It is expected that this approach is also capable of performing better than logistic regression approach in some applications and could also discover some important discriminators which the latter fails to identify.

REFERENCES

1. Baudat, G. and Anouar, F. (2000). Generalized Discriminant Analysis using a Kernel Approach. Neural Computation, 12 (10), 2385-2404
2. Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through Eigen value decomposition. J. Amer. Statist. Assoc. 91, 1743-1748
3. Bressan, M. and Vitria, J. (2003). Nonparametric Discriminant Analysis and Nearest Neighbor Classification. Pattern Recognition Letters. 24, 2743-2749
4. Catov, J.M., Bodnar, L.M., Ness, R.B., Barron, S.J. and Roberts, J.M. (2007). Inflammation and Dyslipidemia related risk of Spontaneous Preterm Birth. Am. J. Epidemiol. 166, 1312-1319
5. Chang, W.-C. (1983). On using Principal Components before Separating a Mixture of two Multivariate Normal Distributions. J. Roy. Statist. Soc. Ser C. 32, 267-275
6. Chiang, L.H. and Pell, R.J. (2004). Genetic algorithms combined with discriminant analysis for key variable identification. J. Process Control, 14, 143-155
7. Habbema, J.D.F. and Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics. 19, 487-493
8. Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible Discriminant Analysis by Optimal Scoring. J. Amer. Statist. Assoc. 89, 1255-1270
9. Mudd, L.M., Holzman, C.B., Catov, J.M., Senagore, P.K. and Evans, R.W. (2012). Maternal lipids at

midpregnancy and risk of preterm delivery. Acta Obstet. Gynecol. Scand. 91, 726-735

10. Murphy, T.B., Dean, N. and Raftery, A.E. (2010). Variable Selection and updating in Model-Based Discriminant Analysis for High Dimensional Data with Food Authenticity Applications. The Annals of Applied Statistics, Vol.4, No.1, 396-421
11. Pfeiffer, K.P. (1985). Stepwise Variable Selection and Maximum Likelihood Estimation of Smoothing Factors of Kernel Functions for Nonparametric Dsicriminant Functions evaluated by Different Criteria. J. Biomed. Informatics. 18, 46-61
12. Raftery, A.E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. J. Amer. Statist. Assoc. 101, 168-178
13. Weismann-Brenner, A., O'Reilly-Green, C. and Ferber, A. (2009). Values of amniotic fluid index in cases of preterm premature rupture of membranes. J. Perinatal Medicine. 37, 232-235.