

## **Discriminant Analysis: A Powerful Classification Technique in Predictive Modelling**

**G. FERNANDEZ**

**University of Nevada – Reno**

### **ABSTRACT**

*Discriminant analysis is one of the classical classification techniques used to discriminate a single categorical variable using multiple attributes. Discriminant analysis also assigns observations to one of the pre-defined groups based on the knowledge of the multi-attributes. When the distribution within each group is multivariate normal, a parametric method can be used to develop a discriminant function using a generalized squared distance measure. The classification criterion is derived based on either the individual within-group covariance matrices or the pooled covariance matrix that also takes into account the prior probabilities of the classes. Non-parametric discriminant methods are based on non-parametric group-specific probability densities. Either a kernel or the k-nearest-neighbor method can be used to generate a non-parametric density estimate in each group and to produce a classification criterion. The performance of a discriminant criterion could be evaluated by estimating probabilities of mis-classification of new observations in the validation data. A user-friendly SAS application utilizing SAS macro to perform discriminant analysis is presented here. Chemical diabetes data containing multi-attributes is used to demonstrate the features of discriminant analysis in discriminating the three clinical types of diabetes.*

**Keywords:** Canonical variables, DISCRIM macro, Multivariate normality

### **INTRODUCTION**

Discriminant analysis (DA), a multivariate statistical technique is commonly used to build a predictive / descriptive model of group discrimination based on observed predictor variables and to classify each observation into one of the groups. Stepwise, canonical and discriminant function analyses are commonly used DA, techniques available in the SAS systems STAT module (SAS Inst. Inc. 2004). In DA multiple quantitative attributes are used to discriminate single classification variable. DA is different from the cluster analysis because prior knowledge of the class membership is required. The common objectives of DA are

- to investigate differences between groups

- to discriminate groups effectively;
- to identify important discriminating variables;
- to perform hypothesis testing on the differences between the expected groupings; and
- to classify new observations into pre-existing groups.

The main objective of this paper is to demonstrate features of the user-friendly SAS application DISCRIM (Fernandez, 2002) using the chemical diabetes data containing multi-attributes in discriminating the three clinical types of diabetes. The users can perform the discriminant analysis using their data by following the instructions included (Fernandez, 2002).

### **Diabetes Data**

Diabet2 data (Reaven and Miller, 1979) containing multi-attributes, X1: relative weight, X2: fasting plasma glucose level, X3: test plasma glucose, X4: Plasma insulin during test, X5: Steady state plasma glucose level and diabet1 (multivariate normally distributed simulated data with same mean vector and equal variance covariance matrix similar to diabet2) are used here to demonstrate the discriminant analysis features in classifying three types of diabetes, 1: Normal, 2: Overt diabetic, and 3: Chemical.

### **Parametric DA analysis using Diabet1 as the training data and diabet2 as the validation data**

*Checking for multivariate normality:* The right choice for selecting parametric vs. non-parametric discriminant analysis is dependent on the assumption of multivariate normality within each group. The diabetes data within each clinical group is assumed to have a multivariate normal distribution. This multivariate normality assumption can be checked by estimating multivariate skewness, kurtosis, and testing for their significance levels for each level. The Quantile-Quantile (Q-Q plot) plot of expected and observed distributions (Khattree and Naik 1995) of multi-attribute residuals can be used to graphically examine for multivariate normality for each group level. The estimated multivariate skewness and multivariate kurtosis clearly support the hypothesis, these four multi-attributes have a joint multivariate normal distribution. A non-significant departure from the 45° angle reference line in the Q-Q plot also supports this finding (Figures not shown). Thus, parametric discriminant analysis can be considered to be the appropriate technique for

discriminating the three clinical groups based on these four attributes (X1 to X4) for data discriml.

### **Canonical Discriminant Analysis (CDA):**

Canonical DA is a dimension-reduction technique similar to principal component analysis. The main objective of CDA is to extract a set of linear combinations of the quantitative variables that best reveal the differences among the groups. Given a nominal group variable and several quantitative attributes, the CDA extracts linear combinations of the quantitative variables (canonical variables) that capture between-class variation in much the same way that principal components summarize total variation (SAS Inst. Inc. 2004). Moreover, these canonical functions will be independent or orthogonal, that is, their contributions to the discrimination between groups will not overlap.

In CDA, canonical variables that have the highest possible multiple correlations with the groups are extracted. The un-standardized coefficients used in computing the raw canonical variables are called the canonical coefficients or canonical weights. The standardized discriminant function coefficients indicate the partial contribution of each variable to the discriminant function(s), controlling for other attributes entered in the equation. The total standardized discriminant functions given in Table 5 indicate that the predictor variable, test plasma glucose (X3), contributed significantly to the first canonical variable (CAN1). The fasting plasma glucose (X2) and test plasma glucose (X3) equally contributed in a negative way to the second canonical variable (CAN2).

These canonical variables are independent or orthogonal to each other; that is, their contributions to the discrimination between groups will not overlap. This maximal multiple correlation between the first canonical variable and the group variables are called the first canonical correlation. The second canonical correlation is obtained by finding the linear combination uncorrelated with the CAN1 that has the highest possible multiple correlations with the groups. In CDA, the process of extracting canonical variables is repeated until you extract the maximum number of canonical variables which is equal to the number of groups minus one, or the number of variables in the analysis, whichever is smaller.

The correlation between the CAN1 and the clinical group is very high ( $>0.9$ ) and about 87% of the variation in the first canonical variable can be attributed to the differences among the three clinical groups (Table 1). The first Eigen value

measures the variability in the CAN1 and accounts for 93% of the variability among the three group members in four predictor variables. The correlation between the CAN2 and the clinical group is moderate (0.5) and about 33% of the variation in the second canonical variable can be attributed to the differences among the three clinical groups (Table 2). The second Eigen value measures the variability in the second canonical variable and accounts for remaining 6% of the variability among the three group members in the four predictor variables. Both canonical variables are statistically highly significant based on the Wilk's lambda test ( $P < 0.001$ ). However, the statistical validity might be questionable if the multivariate normality or the equal variance-covariance assumptions are violated.

**Table 1:** Canonical discriminant analysis using SAS MACRO: DISCRIM- Total canonical structure loadings

Variable	Label	Can1	Can2
x1	Relative wt	0.049	0.599
x2	Fasting Plasma Glucose	0.926	-0.327
x3	Test plasma glucose	0.994	-0.096
x4	Plasma insulin during test	-0.229	0.623

**Table 2:** Canonical discriminant analysis using SAS MACRO: DISCRIM- Canonical correlations

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of $\text{Inv}(E)^*H = \text{CanRsq}/(1-\text{CanRsq})$			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.936	0.934	0.010	0.877	7.130	6.628	0.934	0.934
2	0.578	0.572	0.056	0.334	0.502		0.0658	1.000

For each observation in the training dataset, we can compute standardized canonical variable scores.

These standardized canonical variable scores and the structure loadings can be used in two-dimensional bi-plots (Gabriel, 1981) to aid visual interpretation of the group differences. Inter-relationships among the four multi-attributes and the discriminations of the three groups are presented in Figure 1. The first canonical variable that has the largest loadings on X2 and X3 discriminated the NORMAL (1), OVERT (2) and the CHEMICAL diabetic groups effectively. The CAN2, which has a moderate size loading on X1 and X4, discriminated the NORMAL (1) and the OVERT group. But, CAN2 is not effective in separating the CHEMICAL diabetic group. The narrow angle between the X2 and X3 variable vector in the same

direction indicates that the plasma glucose variables are positively highly correlated. The correlations between X1 and X4 are moderate in size and act in the opposite direction from the plasma glucose variables.

Plot of Canonical Discriminant Analysis- diabet1

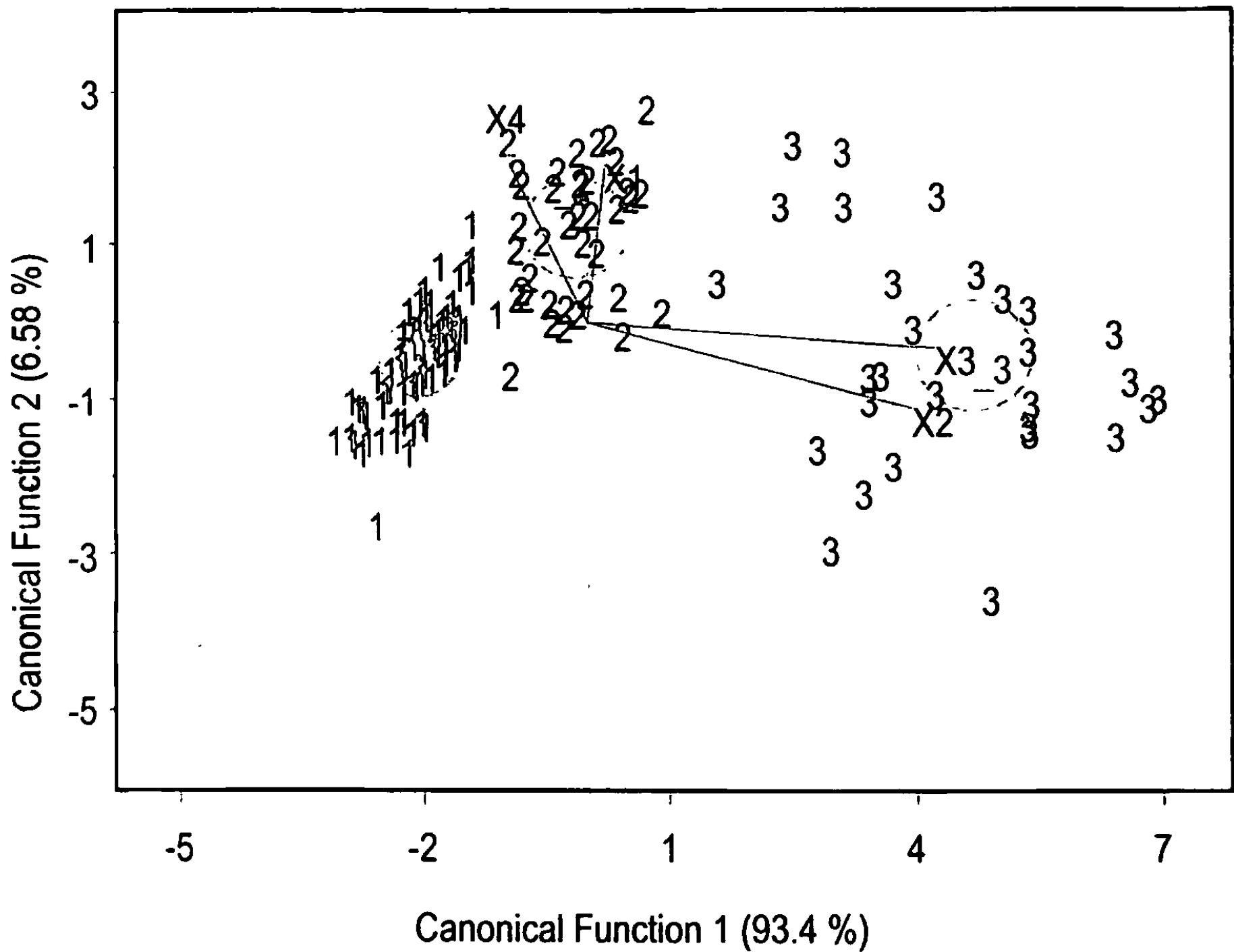


Figure 1: Biplot display of Canonical function scores

The two canonical variables extracted from the CDA effectively discriminated the three clinical diabetic groups. The difference between the normal and the chemical group is distinct. The discrimination between the NORMAL and the OVERT group is effective when both CAN1 and CAN2 are used simultaneously. Therefore, the CDA can be considered as an effective descriptive tool in discriminating groups based on continuous predictor variables. If the variance-covariance between the groups is assumed to be equal and the predictor variables have joint multivariate normal distributions within each group, then the group differences can be tested statistically for group differences.

***Predictive Discriminant Analysis (PDA):*** PDA is a predictive classification technique deals with a set of multi-attributes and one classification variable, the latter being a grouping variable with two or more levels. Predictive discriminant analysis is similar to multiple regression analysis except that PDA is used when the criterion variable is categorical and nominally scaled. As in multiple regression, in PDA a set of rules is formulated which consists of as many linear combinations of predictors as there are categories, or groups. A PDA is commonly used for classifying observations to pre-defined groups based on knowledge of the quantitative attributes. When the distribution within each group is assumed to be multivariate normal, a parametric method can be used to develop a discriminant function using a measure of generalized squared distance. The discriminant function, also known as a classification criterion, is estimated by measuring generalized squared distance (SAS Inst. Inc., 2004). The classification criterion can be derived based on either the individual within-group covariance matrices (a quadratic function) or the pooled covariance matrix (a linear function). This classification criterion also takes into account the prior probabilities of the discriminating groups. Each observation is classified in the group from which it has the smallest generalized squared distance. The posterior probability of an observation belonging to each class could be also estimated in PDA.

The performance of a discriminant criterion in the classification of new observations in the validation data could be evaluated by estimating the probabilities of misclassification or error rates in the SAS DISCRIM procedure. When the input data set is a SAS data set, the error rate can also be estimated by cross validation. SAS uses two types of error-rate estimates to evaluate the derived classification criterion based on parameters estimated by the training sample: i) error-count estimates and ii) posterior probability error-rate estimates. The error count estimate is calculated by applying the discriminant criterion derived from the training sample to a test set and then counting the number of misclassified observations. The group-specific error-count estimate is the proportion of misclassified observations in the group. If the test sample set is independent of the training sample, the estimate is unbiased. However, it can have a large variance, especially if the test sample size is small (SAS Inst. Inc., 2004).

When no independent test sets are available, the same data set can be used both to calibrate and to evaluate the classification criterion. The resulting error-count estimate has an optimistic bias and is called an apparent error rate. To reduce the bias, the data can be split into two sets, one set for deriving the discriminant function and the other set for estimating the error rate. Such a split-sample method has the unfortunate effect of reducing the effective sample size.

Another way to reduce bias in estimating the classification error is cross validation (Lachenbruch and Mickey, 1968). In cross validation,  $n-1$  out of  $n$  training observations in the calibration sample are treated as a training set. It determines the discriminant functions based on these  $n-1$  observations and then applies them to classify the one observation left out. This is performed for each of the  $n$  training observations. The misclassification rate for each group is the proportion of sample observations in that group that are misclassified. This method achieves a nearly unbiased estimate but with a relatively large variance. Classification results based on parametric Quadratic DF and error rates based on cross validation are presented in Table 3.

**Table 3:** Parametric discriminant function analysis using SAS MACRO: DISCRIM-classification table and error count estimates by groups in cross-validation using quadratic discriminant function

a) Training data

From group	To Group			Total
	1	2	3	
1	69 <sup>a</sup> 95.83 <sup>b</sup>	3 4.17	0 0.00	72 100.00
2	1 2.78	35 97.22	0 0.00	36 100.00
3	0 0.00	1 3.03	32 96.97	33 100.00
Total	70 49.65	39 27.66	32 22.70	141 100.00

b) Validation data

1	75 <sup>a</sup> 98.68 <sup>b</sup>	1 1.32	0 0.00	76 100.00
2	2 5.56	34 94.44	0 0.00	36 100.00
3	0 0.00	3 9.09	30 90.91	33 100.00
Total	77 53.10	38 26.21	30 20.69	145 100.00
Error Count Estimates for group				
Rate	0.013	0.055	0.090	0.042
Priors	0.510	0.255	0.234	

To reduce the variance in an error-count estimate Glick (1978) suggested a smoothed error-rate estimates. Instead of summing values that are either zero or one as in the error-count estimation, the smoothed estimator uses a continuum of values between zero and one in the terms that are summed. The resulting estimator has a smaller variance than the error-count estimate. The posterior probability error-rate estimates are smoothed error-rate estimates. The posterior probability estimates for each group are based on the posterior probabilities of the observations classified into that same group. The posterior probability estimates provide good estimates of the error rate when the posterior probabilities are accurate. When a parametric classification criterion (linear or quadratic discriminant function) is derived from a non normal population, the resulting posterior probability error-rate estimators may not be appropriate.

The overall error rate is estimated through a weighted average of the individual group-specific error-rate estimates, where the prior probabilities are used as the weights. To reduce both the bias and the variance of the estimator, Hora and Wilcox (1982) compute the posterior probability estimates based on cross validation. The resulting estimates are intended to have both low variance from using the posterior probability estimate and low bias from cross validation. They use Monte Carlo studies on two-group multivariate normal distributions to compare the cross validation posterior probability estimates with three other estimators: the apparent error rate, cross validation estimator, and posterior probability estimator. They conclude that the cross validation posterior probability estimator has a lower mean squared error in their simulations. Classification results based on parametric Quadratic DF and smoothed error rates based on cross validation are not presented here.

### **Non-parametric discriminant function analysis**

When no distribution assumptions within each group can be made, or when the distribution is not assumed to have multivariate normal, nonparametric methods can be used to estimate the group-specific densities. Non-parametric discriminant methods are based on nonparametric estimates of group-specific probability densities. Either a kernel method or the k-nearest-neighbor method can be used to generate a non-parametric density estimate in each group and to produce a classification criterion. The kernel method in SAS systems uses uniform, normal, Epanechnikov, biweight, or triweight kernels in the density estimation (SAS Inst. Inc., 2004). Either Mahalanobis or euclidean distance can be used to determine proximity in the SAS DISCRIM procedure (SAS Inst. Inc., 2004) When the knearest-neighbor method is used, the Mahalanobis distances are estimated based on the pooled covariance matrix. Whereas in the kernel method, the Mahalanobis

distances based on either the individual within-group covariance matrices or the pooled covariance matrix is estimated.

In non-parametric DA estimation, with the estimated group-specific densities and their associated prior probabilities, the posterior probability estimates of group membership for each class can be evaluated. The classification of an observation vector  $x$  is based on the estimated group specific densities from the calibration or training sample. From these estimated densities, the posterior probabilities of group membership at  $x$  are evaluated.

### **Non-parametric DA analysis using Diabet2 as the training data and diabet1 as the validation data**

**Checking for multivariate normality:** The estimated multivariate skewness and multivariate kurtosis for each group clearly support the hypothesis that, these four multi-attributes do not have a joint multivariate normal distribution. A significant departure from the  $45^\circ$  angle reference line in the Q-Q plot also supports this finding (figures are not included). Thus, non-parametric discriminant analysis must be considered to be the appropriate technique for discriminating the three clinical groups based on these four attributes (X1 to X4).

**Non-parametric DA:** Compare the classification summary and the misclassification rates of these four different non-parametric DFA methods and pick one, that gives the smallest classification error in the cross-validation. Among the three NN-DFA, classification results based on the 2<sup>nd</sup> NN non-parametric DFA gave the smallest classification error. The classification summary and the error rates for NN ( $k=2$ ) are presented in Table 4. When the  $k$ -nearest-neighbor method is used, the *Mahalanobis* distances are estimated based on the pooled covariance matrix. Classification results based on NN ( $k=2$ ) and error rates based on cross-validation are presented in Table 4. The misclassification rates in groups 1, 2, and 3 are 1.3%, 0%, and 12.0% respectively. The overall discrimination is quite satisfactory since the overall error rate is very low, at 3.45%. The posterior probability estimates based on cross validation reduces both the bias and the variance of classification function. The resulting overall error estimates are intended to have both a low variance from using the posterior probability estimate and a low bias from cross validation.

The DISCRIM macro also output a table of the  $i$ th group posterior probability estimates for all observations in the training data set. These posterior probability values are very useful estimates since these estimates can be successfully used in

developing scorecards and ranking the observations in the data set. The posterior probability error-rate estimates for each group are based on the posterior probabilities of the observations classified into that same group. The posterior probability estimates provide good estimates of the error rate when the posterior probabilities are accurate.

If the classification error rate obtained for the validation data is small and similar to the classification error rate for the training data, then we can conclude that the derived classification function has good discriminative potential. Classification results for the validation dataset based on NN(K=2) classification functions are presented in Table 4. The misclassification rates in groups 1, 2, and 3 are 4.1%, 25%, and 15.1.0% respectively. The overall discrimination in the validation dataset is moderately good since the weighted error rate is 11.2%. A total of 17 observations in the validation dataset are misclassified. The misclassification error rate estimated for the validation data set is relatively higher than the error rate obtained from the training data. We can conclude that the classification criterion derived using NN (k=2) performed poorly in validating the independent validation data set. The presence of multivariate influential observations in the training data set might be one of the contributing factors for this poor performance in validation. Using larger K values in NN DFA might do a better job in classifying validation data set.

The classification summary using KD (normal, un-equal bandwidth) non-parametric DFA and the error rates using cross-validation are presented in Table 5. The misclassification rates in groups 1, 2, and 3 are 7.8%, 16.6%, and 9.0% respectively. Thus an overall success rate of correct discrimination is about 90% since the overall error rate is about 10.3%, slightly lower than the overall error rate for the K=2 NN method. The posterior probability error-rate estimates for each group are based on the posterior probabilities of the observations classified into that same group. The smoothed posterior probability error rate estimates based on cross-validation DF are not presented. The overall error rate for stratified and un-stratified estimates are equal since group proportion was used as the prior probability estimate. The overall discrimination is quite satisfactory since the overall error rate using the smoothed posterior probability error rate is relatively low, at 4.7%.

If the classification error rate obtained for the validation data is small and similar to the classification error rate for the training data, then we can conclude that the derived classification function has good discriminative potential. Classification results for the validation data set based on KD (normal, un-equal bandwidth) non

parametric DFA classification functions are presented in Table 5. The misclassification rates in groups 1, 2, and 3 are 4.1%, 19.4%, and 12.1% respectively. The overall discrimination in the validation data set is moderately good since the weighted error rate is 9.7%. A total of 14 observations in the validation data set are misclassified. The misclassification error rate estimated for the validation data set is very much similar to the error rate obtained from the training data. We can conclude that the classification criterion derived using KD (normal, un-equal bandwidth) performed good in validating the independent validation data set.

**Table 5: Unequal bandwidth kernel density discriminant function analysis using SAS MACRO DISCRIM: Classification summary using cross-validation results**

**a) Training data**

<b>Number of Observations and Percent Classified into group</b>				
<b>From group</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>Total</b>
1	70 92.11	5 6.58	1 1.32	76 100.00
2	2 5.56	30 83.33	4 11.11	36 100.00
3	0 0.00	3 9.09	30 90.91	33 100.00
Total	72 49.66	38 26.21	35 24.14	145 100.00
<b>Error Count Estimates for group</b>				
Error Rate	0.078	0.166	0.090	0.103
Prior probability	0.524	0.248	0.227	

**b) Validation data**

_1	69 95.83	3 4.17	0 0.00	72 100.00
_2	7 19.44	29 80.56	0 0.00	36 100.00
_3	2 6.06	2 6.06	29 87.88	33 100.00
Total	78 55.32	34 24.11	29 20.57	141 100.00
Priors	0.52414	0.24828	0.22759	
<b>Error Count Estimates for GROUP</b>				
Rate	0.0417	0.1944	0.1212	0.0977
Priors	0.5241	0.2483	0.2276	

## **User-friendly SAS macro application- DISCRIM**

The DISCRIM macro is a powerful user-friendly SAS application for performing complete discriminant analysis. A screen-shot of the DISCRIM application is not presented. Options are available for obtaining various exploratory and diagnostic graphs and for performing different types of discriminant analyses. SAS procedures, STEPDISC, and DISCRIM are the main tools used in the DISCRIM macro. In addition to these SAS procedures, GPLOT, BOXPLOT procedures, and IML modules are also utilized in the DISCRIM macro. The enhanced features implemented in the DISCRIM macro are:

1. Exploratory bi-variate plots to check for group discrimination in a simple scatter plot between two predictor variables are generated.
2. Plots for checking for multivariate normality and influential observations within each group are also generated.
3. Test statistics and *P-values* for testing equality in variance and covariance matrices within each group level are automatically produced.
4. In the case of CDA, box plots of canonical discriminant functions by groups and biplot display of canonical discriminant function scores of observations and the structure loadings for the predictors are generated.
5. When you fit DFA, box plots of the  $i^{\text{th}}$  level posterior probability by groups are produced.
6. Options are available for validating the discriminant model obtained from a training dataset using an independent validation dataset by comparing classification errors.

### **Software requirements for using the DISCRIM macro**

SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at your site. SAS/IML is required to check for multivariate normality. Active Internet connection is required for downloading DISCRIM macro from the book's website.

## **SUMMARY**

A user-friendly SAS application developed by the author utilizes the advanced analytical and graphical features in SAS software to perform stepwise, canonical and

parametric and non-parametric discriminant function analysis with data exploration is presented here. Chemical diabetes data containing multi-attributes is used to demonstrate the features of discriminant analysis in discriminating the three clinical types of diabetes. The users can perform complete discriminant analysis using their data by using the DISCRIM macro applications and following the instructions included in the book (Fernandez, 2002).

## REFERENCES

- Fernandez, G. (2002). *Data Mining Using SAS applications*. Chapman and Hall, Florida.
- Gabriel, K.R (1981) Bi-plot display of multivariate matrices for inspection of data and diagnosis. (In) *Interpreting Multivariate Data*. V. Barnett (Ed.). John Wiley & Sons: London.
- Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition*, 10(1): 211-222.
- Hora S.C. and J.B. Wilcox. (1982). Estimation of error rates in several population discriminant analyses. *Journal of Marketing Research*, 19: 57-61.
- Khattree R. and D.N. Naik (1995). *Applied Multivariate Statistics with SAS software*. SAS Institute Inc., Cary NC.
- Lachenbruch P.A. and M.A. Mickey (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10: 1-10.
- SAS Institute Inc. (2004). *SAS/STAT 9.1 Users Guide*. SAS Institute Inc., Cary NC.
- Reaven, G.M. and R.G. Miller (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*: 16: 17-24.