DISCRIMINANT ANALYSIS WITH A STRUCTURED COVARIANCE MATRIX

by

Lisa Tomasko and Ronald W. Helms
Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2157

May 1996

DISCRIMINANT ANALYSIS WITH A STRUCTURED COVARIANCE MATRIX

Lisa Tomasko and Ronald W. Helms, University of North Carolina
Lisa Tomasko, Dept. of Biostatistics, CB#7400, 3106E McGavran Greenberg, Chapel Hill, NC 27599

KEY WORDS: Linear Discriminant Function, Random Effects, Multivariate

SUMMARY

Discriminant analysis is commonly used to classify an observation into one of two (or more) populations on the basis of the correlated measurements. Traditionally a linear discriminant function assumes a common unstructured covariance matrix for both treatments, which may be taken from a multivariate model. We consider a model in which the covariance matrix is assumed to have a compound symmetric structure. Thus random subject effects are incorporated into the discrimination process in contrast to standard discriminant analysis methodology. In addition the usual multivariate expected value structure is altered. The impact on the discrimination process is contrasted when using the multivariate and random-effects covariance and expected value structures. To illustrate the procedures we consider repeated measurements data from a clinical trial comparing two active treatments; the goal is to classify patients into the two treatment groups.

1. INTRODUCTION

Classical discriminant analysis is a technique for creating a rule that can be used to assign a vector of observations $\underset{\sim}{y}_i$ to one of k groups based on the value of $\underset{\sim}{y}_i$. Discriminant analysis classifies observations in contrast to other statistical techniques that focus on hypothesis testing. The allocation process involves producing an optimum linear combination of correlated measurements that differentiates between groups or populations. There is not a mechanism in classical discriminant analysis to model the covariance structure. Thus information regarding the possible structure in the covariance for correlated measurements taken on the same individual is lost. Person level flexibility with respect to variability and unequal numbers of observations is the focus of our paper.

Our clinical trial application is primarily for illustration of the "new" discriminant analysis technique. There exist more meaningful classification problems than simply differentiating between two active treatments. For instance classifying a patient based on their repeated measures over time as a responder or non-responder to treatment. Other uses could be as an aid for classification of a type of disease severity to assist in treatment decisions. The extensions are plentiful. Before presenting new results for discriminant analysis methodology, it is necessary to review some important terminology in current statistical methods.

1.1 Discriminant Analysis

Traditional linear discriminant analysis is based on normality and equal covariance structures in the groups.[1-2] Quadratic discriminant functions allow for unequal variance structures between the groups.[3-5] Here we focus on the linear discriminant function for the 2 group case. Let $\underset{\sim}{Y}_{ik}$ denote the random $m \times 1$ vector of measurements from the $i-th$ subject, $i = 1, 2, ..., N_k$, $k = 1, 2$. Assume $\underset{\sim}{Y}_{ik} \sim N_m(\underset{\sim}{\mu}_k, \underset{\sim}{\Sigma}_k)$ where $\underset{\sim}{\mu}_k$ is the $m \times 1$ mean vector of the $k-th$ population and $\underset{\sim}{\Sigma}_k$ is the $m \times m$ covariance matrix of the $k-th$ population. We shall assume throughout that any distinct vectors of observations are stochastically independent, $i.e. \underset{\sim}{Y}_{ik} \perp \underset{\sim}{Y}_{i'k'}$ if $(i, k) \neq (i', k')$.

Discriminant analysis is based on likelihood priniciples. Given a new observation, $\underset{\sim}{Y}_i$, the estimated log likelihood function value of $\underset{\sim}{Y}$; $L(\underset{\sim}{y}\ ; \hat{\underset{\sim}{\mu}}_k, \hat{\underset{\sim}{\Sigma}}_k)$ is evaluated separately for each population, using estimated parameters for that population. $\underset{\sim}{Y}_i$ is assigned to the population with the largest value of $L(\underset{\sim}{y}\ ; \hat{\underset{\sim}{\mu}}_k, \hat{\underset{\sim}{\Sigma}}_k) * ln(p_k)$. Where p_k is the a priori probability that the subject is selected from population k.

The evaluation of the classification is done through error rates. There are several types of error rates in the literature.[2] The *apparent error rate* or *resubstitution error rate* is defined as the fraction of observations in the initial sample which were misclassified by the discriminant function. This method has an optimistic bias and can perform poorly in small samples.[2] However for large samples the resubstitution error rate is quite satisfactory.[2]

To nearly eliminate the bias the jackknife estimator was introduced.[2] The jackknife estimator of the error rate calculates a discriminant function with $N - 1$ observations and classifies the one observation remaining. This process is continued until each observation is classified.

1.2 The Mixed Model

Mixed models have a long history dating back to ANOVA settings. In particular the expected mean squares and variance components used in complicated ANOVA settings.[6] Current mixed models methodology has been extended to settings outside of the initial ANOVA.

The following notation and assumptions will be used to describe the mixed models that accommodate both mixed and random effects.[7] Consider the model:

$$\underset{\sim}{Y}_i = \underset{\sim}{X}_i \underset{\sim}{\beta} + \underset{\sim}{Z}_i \underset{\sim}{d}_i + \underset{\sim}{e}_i \quad i = 1, \ldots, N$$

where

$\underset{\sim}{Y}_i$ is a $m_i \times 1$ vector of measurements (or observations) of a dependent variable or response variable from the $i-$th subject, $i = 1, \ldots, N$.

$\underset{\sim}{\beta}$ is a $p \times 1$ vector of unknown, constant, fixed effect primary parameters.

$\underset{\sim}{X}_i$ is an $m_i \times p$ fixed effects design matrix for the $i-$th subject. The values in $\underset{\sim}{X}_i$ are fixed known constants without appreciable error.

$\underset{\sim}{d}_i$ is a $q \times 1$ vector of unobservable random effects or subject effect coefficients.

$\underset{\sim}{Z}_i$ denotes an $m_i \times q$ random effects design matrix for the $i-$th subject

$\underset{\sim}{e}_i$ is the $m_i \times 1$ vector of unobservable within-subject error terms.

The model assumptions include:

$$E\begin{bmatrix} \underset{\sim}{d}_i \\ \underset{\sim}{e}_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$V\begin{bmatrix} \underset{\sim}{d}_i \\ \underset{\sim}{e}_i \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\Delta} & 0 \\ 0 & \sigma^2 \underset{\sim}{V} \end{bmatrix}$$

and for $i' \neq i$

$$\operatorname{cov}[d_{i'}, d_i] = 0; \ \operatorname{cov}[d_{i'}, e_i] = 0; \ \operatorname{cov}[e_{i'}, e_i] = 0$$

which lead to:

$$\mathrm{E}(\underset{\sim}{Y}_i) = \underset{\sim}{X}_i \underset{\sim}{\beta}$$
$$\mathrm{Var}(\underset{\sim}{Y}_i) = \underset{\sim}{\Sigma}_i = \underset{\sim}{Z}_i \underset{\sim}{\Delta} \underset{\sim}{Z}_i' + \sigma^2 \underset{\sim}{V}_i$$

Note $\underset{\sim}{Y}_i$ is independent of $\underset{\sim}{Y}_j$ for $i \neq j$.

Maximum likelihood estimation of $\hat{\underset{\sim}{\beta}}$ involves $\hat{\underset{\sim}{\Sigma}}_i$ resulting in non-linear estimation equations. Maximum likelihood or restricted maximum likelihood solutions are attainable through several algorithms.[8-9] The most common are: (1) EM- estimation and maximization, (2) Newton- Raphson, and (3) Method of Scoring. Each of these algorithms has been explored in the literature. [10-12]

2. LINEAR DISCRIMINANT ANALYSIS WITH LONGITUDINAL DATA

In a longitudinal experimental design the natural model for the multiple measurements is the general linear multivariate model. Under the general linear multivariate model each individual's response vector is assumed to be distributed multivariate normal as $\underset{\sim}{Y}_{ik} \sim N_m(\underset{\sim}{\mu}_k, \underset{\sim}{\Sigma}_k)$. The respective population parameters from this distribution are estimated using standard general linear multivariate model procedures. The linear discriminant function, assuming $\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2$, then becomes

$$D(\underset{\sim}{y}) = \{(\underset{\sim}{y} - 1/2(\hat{\underset{\sim}{\mu}}_1 + \hat{\underset{\sim}{\mu}}_2)\}' \hat{\underset{\sim}{\Sigma}}^{-1} (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2). \qquad (1)$$

Using the above estimators from the multivariate model each individual is classified into one of two groups. If $D(\underset{\sim}{y}) > ln(p_2/p_1)$ then subject is classified into population 1; otherwise the subject is classified into population 2. The assumptions for this model are $\underset{\sim}{Y}$ is a vector of random variables, $\underset{\sim}{X}$ is a design matrix of fixed constants and $\underset{\sim}{\Sigma}$ is an unstructured covariance matrix. Every individual is assumed to have the same covariance matrix, *i.e.*the homoscedasticity assumption. Here the normality assumption of the $\underset{\sim}{y}$ is needed for the discriminant function to be applied.

The primary goal of discriminant analysis is to classify individuals based on the data at hand. The two groups that an individual can be placed into are commonly indentified as high or low risk groups. In conjunction with the classification process there is additional interest in indentifying a probability of group membership for each individual. An approach to this problem is to formulate a logistic model where the discriminant function value is a covariate and group membership is the dependent value. The predicted probabilities from such a model can be viewed as the probability of group membership.

2.1 Modifications to the Expected Value Structure

The expected value structure of the multivariate model allows each parameter to change over time. Hence all time interactions are implicitly included in the model. $\underset{\sim}{\widehat{\beta}}$ has $p \times m$ parameters where p is the number of predictors and m is the number of timepoints. If it is reasonable to average effects across timepoints then it maybe of interest to remove the interactions with time from the expected value structure. The structure of the discriminant function is the same as shown in (1) but $\underset{\sim}{\Sigma}$ is now estimated with an additional $p \times (m\text{-}1)$ degrees of freedom and the $\underset{\sim}{X}$ matrix columns are reduced by $p \times (m\text{-}1)$.

2.2 Modifications to the Covariance Structure

The unstructured covariance matrix in the general linear multivariate model requires each subject to have the same covariance structure. The estimation requires each subject to have complete data so that a subject with incomplete records is not included in the estimation. Utilizing mixed model methodology enables all case records, regardless of completeness of longitudinal measurements, to contribute to the analysis. Additionally the covariance structure is allowed to vary, in dimension and structure, from subject to subject. In contrast to the unstructured matrix in the multivariate model, the mixed model has the ability to model the covariance structure. Thus instead of having to estimate $\{m \times (m\text{-}1)\}/2$ parameters for the covariance matrix the number of estimated parameters can be substantially reduced.

The covariance structure in the mixed model is defined as $\text{Var}(y_{ik}) = \underset{\sim}{Z}_i \underset{\sim}{\Delta} \underset{\sim}{Z}_i' + \sigma^2 \underset{\sim}{V}_i$. This variance is for one particular subject and thus it can vary by subject and can allow the model to capture the variability in a more precise manner. Some popular choices for the covariance structure are:

1. Random Intercept Model (Compound Symmetry)

$$\text{Var}(y_{ik}) = \underset{\sim}{Z}_{i\,(m\times1)} \underset{\sim}{\Delta}_{(1\times1)} \underset{\sim}{Z}_i{}'_{(1\times m)} + \sigma^2 \underset{\sim}{V}_{i(m\times m)}$$

Subject is the only variance component in $\underset{\sim}{\Delta}$ or equivalently intercept is the only random effect. Additionally, a common within-subject variance component σ^2 is estimated while the off diagonal elements of the $\underset{\sim}{V}$ matrix are zero. The off diagonal elements of the variance of y_{ik} all have equal covariance while the diagonal components have variance equal to $\delta + \sigma^2$. Thus this compound symmetric structure contains two parameters.

2. Random Intercept and Random Slope Model

$$\text{Var}(y_{ik}) = \underset{\sim}{Z}_{i\,(m\times2)} \underset{\sim}{\Delta}_{(2\times2)} \underset{\sim}{Z}_i{}'_{(m\times2)} + \sigma^2 \underset{\sim}{V}_{i(m\times m)}$$

Here the intercept and slope vary for each subject. Thus $\underset{\sim}{\Delta}$ has 3 random effects. An effect for subject , an effect for time, and an effect for the variability of subject by time. The within-subject variance, σ^2, is constant across time while the off-diagonal elements of the $\underset{\sim}{V}$ matrix are zero. Aside from symmetry, the elements of the $Var(y_{ik})$ may all be unique where none of the variances or covariances are equal to each other. This covariance structure contains a total of 3 random-effect parameters and 1 fixed-effect parameter.

3. First-order Autoregressive

$$\text{Var}(y_{ik}) = \sigma^2 \underset{\sim}{V}_{i(m\times m)}\ ;\quad v_{ij} = \rho^{\,|i-j|}$$

Here $\underset{\sim}{V}$ has a structure other than the identity matrix in contrast to the first two covariance structures. The variances are homogenous across time while the covariances depend on the time lag, $|i-j|$, between pairs of measures. This variance structure of y_{ik} has two covariance parameters.

Regardless of covariance structure chosen, the more precise covariance structure is then used in the estimation of the linear discriminant function as shown in (1), replacing the unstructured matrix from the general linear multivariate model. The expected value structure is then estimated with an increase in degrees of freedom.

3. CLINICAL TRIAL EXAMPLE

The data for this example are taken from a randomized, multicenter clinical trial in patients with diagnosis of general anxiety disorder. Anxiety is measured by the Hamilton Anxiety Scale where higher scores mean higher anxiety. There were baseline measures and post-treatment measures at each of four weeks following randomization. Two active treatments and one placebo treatment were evaluated. For this illustration only data from the active treatment arms are investigated. Only patients with complete records, i.e. no missing visits, are used in this example.

However the "new" discriminant analysis methodology, developed here, could handle all records regardless of completeness.

Change from baseline was used as the response. Figure 1 displays the mean changes for the two treatment groups. The standard deviations at each timepoint were narrow (pooled across treatments $\hat{\sigma}$ = 1.72) There is clear distinction between the two groups suggesting discriminant analysis could be a powerful tool for these data.

Figure 1
Mean Change from Baseline (Ham-A) Across Time

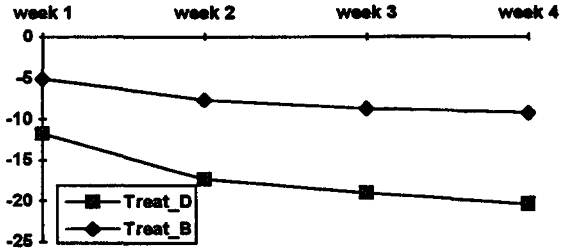

Table 1 shows the covariance and correlation matrices at all 5 timepoints. The unstructured covariance matrix taken from the multivariate model and the compound symmetric covariance taken from the random-effects model are shown in Table 2. The random intercept (compound symmetric $\underset{\sim}{\Sigma}$) model was found to be a good fit for the data.

Table 1 Covariance and Correlation Matrices

Covariance at baseline and 4 post treatment measures

$$\underset{\sim}{\hat{\Sigma}} = \begin{bmatrix} 4.44 & 3.59 & 3.10 & 2.83 & 2.17 \\ & 4.37 & 3.78 & 3.17 & 2.80 \\ & & 4.37 & 3.60 & 2.98 \\ & & & 4.48 & 3.43 \\ & & & & 4.08 \end{bmatrix}$$

Correlation between all 5 timepoints

$$\underset{\sim}{\hat{P}} = \begin{bmatrix} 1.0 & .81 & .70 & .63 & .51 \\ & 1.0 & .86 & .72 & .66 \\ & & 1.0 & .81 & .70 \\ & & & 1.0 & .81 \\ & & & & 1.0 \end{bmatrix}$$

Table 2. Unstructured and Structured Covariance Matrices

Change from Baseline **Unstructured** Covariance Matrix

$$\underset{\sim}{\hat{\Sigma}}_{(m\times m)} = (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\hat{\beta}})'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\hat{\beta}})/(N-p)$$

$$\underset{\sim}{\hat{\Sigma}} = \begin{bmatrix} 1.64 & 1.54 & 1.20 & 1.49 \\ & 2.62 & 2.11 & 2.15 \\ & & 3.26 & 2.88 \\ & & & 4.18 \end{bmatrix}$$

Change from Baseline **Structured** Covariance Matrix- Compound Symmetry

$$\underset{\sim}{\hat{\Sigma}}_{(m\times m)} = \underset{\sim}{Z}_{i\,(m\times 1)}\underset{\sim}{\hat{\Delta}}_{(1\times 1)}\underset{\sim}{Z}_{i\,(1\times m)}' + \sigma^2\underset{\sim}{I}_{(m\times m)}$$

$$\underset{\sim}{\hat{\Sigma}} = \begin{bmatrix} 2.92 & 1.89 & 1.89 & 1.89 \\ & 2.92 & 1.89 & 1.89 \\ & & 2.92 & 1.89 \\ & & & 2.92 \end{bmatrix}$$

The discriminant analysis to classify a patient into one of the two treatment groups based on $\underset{\sim}{Y}_{ik}$ will be implemented using four separate scenarios described below:

(1) Case 1- Usual general linear multivariate estimates for expected value and covariance structures.

(2) Case 2- Expected value structure altered by removing interactions across time when appropriate. Covariance structure identical to Case 1, *i.e.* unstructured covariance matrix.

(3) Case 3- Expected value structure identical to Case 1, *i.e.* including all interactions with time. Covariance structure modeled as compound symmetric.

(4) Case 4- Expected value structure altered by removing interactions across time when appropriate. In addition, covariance structure modeled as compound symmetric.

Additionally, the expected value structure can be modeled using only the model's fixed effects. Thus, the expected values would be the population values and not contain a random effect component.

To evaluate the discriminant function in each of the 4 cases the resubstitution error rate was calculated. This error rate is intuitive and performs adequately in our sample of (N=104). The results are listed in Table 3.

Table 3. Results of the Discriminant Analysis for each of the 4 Cases

Case	$E(\underset{\sim}{Y})$	$V(\underset{\sim}{Y})$	Resubstitution Error Rate
1	GLMM- interactions	Unstructured	29/104 (28%)
2	no interactions	Unstructured	26/104 (25%)
3	GLMM- interactions	Compound Symmetry	0/104 (0%)
4	no interactions	Compound Symmetry	0/104 (0%)

4. CONCLUSIONS

It is clear that the discriminant function is affected by modifications to its covariance matrix. The results are quite encouraging. Noteworthy, the resubstitution error rate was used here and is expected to be optimistically biased. However in relative terms the error rate is substantially improved when using a structured covariance matrix. Additionally these results are shown in a complete data setting. The data were restricted to complete cases to enable the multivariate model to be comparable to the random-effects model. Clearly the random-effects model is helpful in missing data at random situations. This opens up many paths for future applications.

REFERENCES

1. Lachenbruch, P. A. (1979), "Discriminant Analysis," *Biometrics*, 35, 69-85.

2. Lachenbruch, P. A. (1982), *Encyclopedia of Statistical Sciences, Volume 2*, New York: John Wiley, pp.389-397.

3. Marks, S. and Dunn, O. J. (1974), "Discriminant Functions When Covariance Matrices are Unequal," *Journal of the American Statistical Association*, 69, 555-559.

4. Flury, B. W. and Schmid, M. J. (1992), "Quadratic Discriminant Functions with Constraints on the Covariance Matrices: Some Asymptotic Results," *Journal of Multivariate Analysis*, 40, 244-261.

5. Wahl, P. W. and Kronmal, R. A. (1977), "Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate," *Biometrics*, 33, 479-484.

6. Scheffe', H. (1959), *The Analysis of Variance*. New York: John Wiley & Sons.

7. Helms, R. W. (1995), *The Joy of Modeling: General Linear Models for the Analysis of Continuous Data,* Copyrighted unpublished manuscript, The University of North Carolina, Chapel Hill, NC 27599-7400.

8. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM algorithm (with Discussion)," *Journal of the Royal Statistical Society* (B), 39, 1-38.

9. Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320-340.

10. Laird, N.M. and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963-974.

11. Laird, N. M., Lange, N., and Stram, D. (1987), "Maximum Likelihood Computations With Repeated Measures: Application of the EM Algorithm," *Journal of the American Statistical Association*, 82, 97-105.

12. Lindstrom, M. J. and Bates, D. M. (1988), "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated Measures Data," *Journal of the American Statistical Association*, 83, 1014-1022.