Computational Statistics and Data Analysis 56 (2012) 2782–2794

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust descriptive discriminant analysis for repeated measures data

Tolulope T. Sajobi [a], Lisa M. Lix [a,c,*], Bolanle M. Dansu [a,b], William Laverty [c], Longhai Li [c]

[a] School of Public Health, University of Saskatchewan, Saskatoon, Canada

[b] Department of Statistics, University of Agriculture, Abeokuta, Nigeria

[c] Department of Mathematics & Statistics, University of Saskatchewan, Saskatoon, Canada

ARTICLE INFO

Article history:
Received 27 July 2011
Received in revised form 26 January 2012
Accepted 29 February 2012
Available online 6 March 2012

Keywords:
Bias
Covariance structure
Discriminant function coefficients
Repeated measurements
Root mean square error

ABSTRACT

Discriminant analysis (DA) procedures based on parsimonious mean and/or covariance structures have recently been proposed for repeated measures data. However, these procedures rest on the assumption of a multivariate normal distribution. This study examines repeated measures DA (RMDA) procedures based on maximum likelihood (ML) and coordinatewise trimming (CT) estimation methods and investigates bias and root mean square error (RMSE) in discriminant function coefficients (DFCs) using Monte Carlo techniques. Study parameters include population distribution, covariance structure, sample size, mean configuration, and number of repeated measurements. The results show that for ML estimation, bias in DFC estimates was usually largest when the data were normally distributed, but there was no consistent trend in RMSE. For non-normal distributions, the average bias of CT estimates for procedures that assume unstructured group means and structured covariances was at least 40% smaller than the values for corresponding procedures based on ML estimators. The average RMSE for the former procedures was at least 10% smaller than the average RMSE for the latter procedures, but only when the data were sampled from extremely skewed or heavy-tailed distributions. This finding was observed even when the covariance and mean structures of the RMDA procedure were mis-specified. The proposed robust procedures can be used to identify measurement occasions that make the largest contribution to group separation when the data are sampled from multivariate skewed or heavy-tailed distributions.

1. Introduction

Linear discriminant analysis (DA; Fisher, 1936) is a multivariate procedure for predicting group membership (predictive discriminant analysis, PDA) and/or describing group separation (descriptive discriminant analysis, DDA) on a set of correlated variables. The former focuses on the accuracy of classification while the latter uses discriminant function coefficients (DFCs) to rank order variables according to their contributions to group separation (Rencher, 2002). The linear DA procedure makes no assumptions about the structures of the means or covariances of the variables other than the assumption of homoscedasticity (i.e., equality) of group covariances. Recently, several repeated measures DA (RMDA) procedures based on parsimonious mean and covariance structures, including constant means and compound symmetric (CS) or first-order autoregressive (AR-1) covariances, have been developed for PDA (Lix and Sajobi, 2010; Roy and Khattree, 2005a,b; Tomasko et al., 1999). These procedures are efficient when the sample size is small relative to the number of repeated measurements, although mis-specification of the mean or covariance structure may influence bias and accuracy.

* Correspondence to: School of Public Health, University of Saskatchewan, 107 Wiggins Road, Saskatoon, SK Canada, S7N 5E5. Tel.: +1 306 966 1617; fax: +1 306 966 7920.
E-mail address: lisa.lix@usask.ca (L.M. Lix).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.02.029

Previous research has shown that while the linear DA procedure will sometimes result in smaller misclassification error rates (MERs) for PDA in multivariate non-normal than normal data (Ashikaga and Chang, 1981; Baron, 1991; Lachenbruch et al., 1973), it will also frequently produce incorrect variable ranks for DDA when the data are non-normal (McLachlan, 1992). Thus, departures from the assumption of multivariate normality may have serious consequences for researchers who adopt the linear DA procedure. However, the effects of non-normality on the performance of RMDA procedures based on parsimonious mean and covariance structures have received little, if any, attention.

Several linear DA procedures that are robust (i.e., insensitive) to departures from the assumption of multivariate normality have been proposed (Todorov and Pires, 2007; Todorov et al., 1994) by replacing the conventional least-squares estimators of means and covariances with robust estimators, including M-estimators (Campbell, 1982), S-estimators (Croux and Dehon, 2001; He and Fung, 2000), minimum covariance determinant (MCD) estimators (Hubert and van Driessen, 2004; Rousseeuw, 1984), minimum volume ellipsoid (MVE) estimators (Rousseeuw, 1984; Rousseeuw and van Zomeren, 1990), and estimators based on the trimmed Mahalanobis distance (M-distance; Ahmed and Lachenbruch, 1977; Gnanadesikan and Kettenring, 1972). Some of these estimators have shown poor performance for PDA. Specifically, M-estimators may result in high MERs in high-dimensional data (Hawkins and McLachlan, 1997). Moreover, estimators based on trimmed M-distances may be sensitive to multivariate outliers (Campbell, 1982).

One approach that has not previously been used to develop robust DA procedures is to adopt estimators based on coordinatewise trimming (CT). In univariate data, trimmed means possess good theoretical properties for heavy-tailed and skewed distributions, are computationally efficient, and straightforward to implement (Wilcox, 1994). To implement CT, the coordinates of the multivariate data are independently trimmed by removing a pre-determined proportion of the observations from each tail of the distribution. Trimmed means and Winsorized covariances are then computed from the CT data; the latter are the theoretically correct estimators of variance corresponding to the trimmed mean. Robust estimators based on CT have been adopted in previous studies about multivariate and repeated measures procedures (Keselman et al., 2000a,b; Srivastava and Mudholkar, 2001).

RMDA procedures have a number of applications for describing group differences or predicting group membership. de Coster et al. (2005) used DA to develop a classification rule for stroke patients. Self-reported depression scores at one, three, six, and nine months post stroke were used to discriminate between patients with and without a clinical diagnosis of depression. RMDA procedures can also be used to identify the ages at which there are significant differences between the growth curves of male and female adolescents (Cutting et al., 2002; Knutsson et al., 1997).

In this study, we investigate robust RMDA procedures based on parsimonious means and/or covariances in which the conventional maximum likelihood (ML) estimates of the means and covariances are replaced by ML estimates of means and covariance parameters based on CT. The effects of population shape and other data characteristics on bias and error in DFCs are investigated using Monte Carlo techniques.

2. DFC estimation for RMDA procedures

This section focuses on the two-group problem, although all of the procedures can be generalized to multi-group designs. Let $\mathbf{y}_{ij}$ be the $p \times 1$ vector of observed measurements for the ith study participant ($i = 1, \ldots, n_j$) in the jth group ($j = 1, 2$). Initially we assume that $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the population mean and covariance for the jth group and are estimated by $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$, respectively. For the linear DA procedure, the DFC vector is estimated by

$$\hat{\mathbf{a}} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2), \tag{1}$$

where

$$\hat{\boldsymbol{\Sigma}} = \frac{(n_1 - 1)\hat{\boldsymbol{\Sigma}}_1 + (n_2 - 1)\hat{\boldsymbol{\Sigma}}_2}{n_1 + n_2 - 2}. \tag{2}$$

The number of uncorrelated discriminant functions that separates g groups is equal to $g - 1$. For the linear DA procedure, least-squares and ML estimators of the means and covariances are equivalent (McLachlan, 1992).

RMDA procedures based on constant mean vectors and compound symmetric (CS) or first-order autoregressive (AR-1) covariances for multivariate normal data have been described previously (Roy and Khattree, 2005a; Sajobi et al., in press), but are included here for completeness. For the CS structure, $\boldsymbol{\Sigma}$ has diagonal elements of σ^2 and off-diagonal elements of $\sigma^2\rho$, where ρ is the correlation parameter. For the AR-1 structure, the covariance elements are equal to $\sigma^2\rho^{|k-l|}$ for the kth and lth measurement occasions ($k, l = 1, \ldots, p$). The DFCs for each procedure are obtained by substituting ML estimates of $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ into Eq. (1). However, the assumption of constant $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ may not be tenable for repeated measures data and therefore procedures for unstructured means are developed next.

For a RMDA procedure that assumes a CS structure for $\boldsymbol{\Sigma}$ and unstructured means, let $\boldsymbol{\Theta}$ be a column vector of model parameters where the first $2p$ elements denote the mean parameters and the last two elements correspond to σ^2 and $\sigma^2\rho$, respectively. Let $\mathbf{Y}_j$ denote the $n_j \times p$ data matrix for the jth group and $n = n_1 + n_2$. Then the joint log-likelihood function is

$$\log l(\boldsymbol{\Theta}|\mathbf{Y}_1, \mathbf{Y}_2) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\left\{\sum_{j=1}^{2}\sum_{i=1}^{n_j}(\mathbf{y}_{ij} - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}_{ij} - \boldsymbol{\mu}_j)\right\}, \tag{3}$$