Computational Statistics and Data Analysis 56 (2012) 1644–1661

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Linear discrimination for three-level multivariate data with a separable additive mean vector and a doubly exchangeable covariance structure

Ricardo Leiva [a], Anuradha Roy [b,*]

[a] Departamento de Matemática, F.C.E., Universidad Nacional de Cuyo, 5500 Mendoza, Argentina

[b] Department of Management Science and Statistics, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 6 May 2011
Received in revised form 6 October 2011
Accepted 6 October 2011
Available online 15 October 2011

Keywords:
Additive mean structure
Doubly exchangeable covariance structure
Linear discriminant function
Maximum likelihood estimates

ABSTRACT

In this article, we study a new linear discriminant function for three-level m-variate observations under the assumption of multivariate normality. We assume that the m-variate observations have a doubly exchangeable covariance structure consisting of three unstructured covariance matrices for three multivariate levels and a separable additive structure on the mean vector. The new discriminant function is very efficient in discriminating individuals in a small sample scenario. An iterative algorithm is proposed to calculate the maximum likelihood estimates of the unknown population parameters as closed form solutions do not exist for these unknown parameters. The new discriminant function is applied to a real data set as well as to simulated data sets. We compare our findings with other linear discriminant functions for three-level multivariate data as well as with the traditional linear discriminant function.

1. Introduction

Even though the scientists in the present era agree with the need for flexibility in the classification technique for the multi-dimensional multivariate data, the elaboration of formal supporting statistical frameworks is just at the very beginning. An appropriate classification rule needs to be developed that is suitable for a particular data set. This fact motivated us to write a few articles on classification rules for three-level multivariate data with different mean vectors and different covariance structures for diverse data sets, so that they would be readily available whenever there is a need for them in future. The main idea of this article is to employ the information of a double exchangeability of a variance–covariance matrix, which allows to partition a covariance structure into three unstructured covariance matrices, corresponding to each of the three levels. The mean vector is assumed to have a separable additive structure. As a consequence, the number of estimated covariance parameters is substantially reduced, comparing to a classical approach, which enables us to apply the proposed procedure even to a very small number of observations. This is of critical importance to a variety of applied problems with repeated measures in medicine, biostatistics and social sciences. The classification rule for three-level multivariate data was first introduced by Roy and Leiva (2007). They used the constant mean vector over sites (CMVOS) in addition to the doubly exchangeable covariance structure or jointly equicorrelated covariance structure on the variance–covariance matrix. Later Leiva and Roy (2009a,b, 2011) studied the problem of classification of three-level multivariate data or triply multivariate data by using an "equicorrelated (partitioned) matrix" (Leiva, 2007) on the measurement vector over sites in addition to an AR(1) correlation structure on the repeated measurements over time, and the doubly exchangeable covariance structure respectively. In Leiva and Roy (2009a), they discussed structured mean

* Corresponding author. Tel.: +1 210 458 6343; fax: +1 210 458 6350.
E-mail addresses: rleiva@fcemail.uncu.edu.ar (R. Leiva), aroy@utsa.edu (A. Roy).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.10.007

vectors without time and site effects, with time effects, with both time and site effects and also with unstructured mean vectors, while in Leiva and Roy (2009b, 2011) they discussed separable multiplicative mean vectors and unstructured mean vectors. These classification rules for three-level or three-dimensional multivariate data demand ingenious approaches to the processes of study design, testings of hypotheses and estimation of unknown parameters. These problems result in models that are often not tractable analytically, and therefore their closed form solutions cannot be obtained. Hence, applying various iteration algorithms are almost always essential to find an exact or approximate solution to these models.

Multi-level multivariate data are very common in biological, biomedical, medical, environmental, engineering and many other fields. They require extraction of relevant information that is hidden in the data in order to model the data appropriately and accurately. Stanimirova and Simeonov (2005) modeled multiway data (Kroonenberg, 2008), especially a four-level environmental data set (*particle size fractions* × *concentrations of chemical components* × *seasons* × *sampling sites*) that came from monitoring of air quality in two industrial regions in Austria. In another application of stem cell systems biology Yener et al. (2008) applied multiway modeling and analysis techniques to model the dynamic activity of biological networks over time. In particular Yener et al. (2008) applied multiway modeling and analysis techniques to model two systems biology problems: (i) discovering functional clusters of gene/protein expression during stem cell differentiation, and (ii) dynamics of human mesenchymal stem cells osteogenic differentiation over time. In the first case they arranged their data in three-level fashion as *protein/gene locus link* × *gene ontology category* × *osteogenic stimulant*, and in the second case they arranged their data once more in three-level manner as *gene IDs* × *osteogenic stimulus* × *replicates*. Lange and Wu (2008) introduced a new method of multicategory vertex discriminant analysis using a primal majorization–minimization algorithm that relies on quadratic majorization and iteratively re-weighted least squares in a non-parametric set-up. These two authors Wu and Lange (2010) later explored an elaboration of this vertex discriminant analysis (VDA) for high-dimensional data that conducts classification and variable selection simultaneously where the number of predictor variables is comparable to or larger than the number of samples. Most recently, Akdemir and Gupta (2011) have developed classification techniques for high dimensional multi-level or multiway data. In their paper Akdemir and Gupta presented a technique called slicing for obtaining an approximate nonsingular estimate of the covariance matrix of high dimensional data when sample size is less than the dimension of the vector variate random variable.

This article is organized as follows. In Section 2, we define the problem and introduce the separable additive mean vector structure and doubly exchangeable covariance structure. The maximum likelihood estimates (MLEs) of the separable additive mean vector and the doubly exchangeable covariance matrix in a single population case are obtained in Section 3. The new classification rule is presented in Section 4. An example of a real data set is given in Section 5, and a simulation study is carried out in Section 6 to show the effectiveness of our new classification rule. Possible future extensions of our proposed classification rule to more general set-ups are discussed in Section 7. Finally, Section 8 concludes with several comments. Technical proofs of the MLEs of all unknown parameters, derivations of different classification rules and derivatives of Kronecker sum are presented in three appendices.

2. Problem statement

In this article we consider discrimination for three-level multivariate data among k different populations with a separable additive mean vector structure (defined in Section 2.1) in addition to a doubly exchangeable covariance structure (defined in Section 2.2). Let t and s denote a given point in time and a given site respectively. Let $\boldsymbol{x}_{ts}^{(p)} : (\Omega, P) \rightarrow \Re^m,\ 1 \le t \le v,\ 1 \le s \le u$, be the m-dimensional normally distributed random vector from the pth population. Then the random families $(\boldsymbol{x}_{1s}^{(p)})_{s\in\{1,\ldots,u\}}, \ldots, (\boldsymbol{x}_{vs}^{(p)})_{s\in\{1,\ldots,u\}}$ are assumed to be exchangeable. Furthermore, for fixed t, the family of random variables $(\boldsymbol{x}_{ts}^{(p)})_{s\in\{1,\ldots,u\}}$ is exchangeable. This assumption of double exchangeability reduces the number of unknown parameters considerably, thus allows more dependable or reliable parameter estimates. This covariance structure can capture the data arrangement or data pattern in a three-level multivariate data, and thus may offer more information about the true association of the data. One of the many advantages of this covariance structure is that the repeated measurements data need not be of equally spaced.

Let $\boldsymbol{x}_r^{(p)}$ be the muv-variate vector of all measurements corresponding to the rth individual in the pth population $r = 1, \ldots, n^{(p)}, p = 1, \ldots, k$. We partition this vector $\boldsymbol{x}_r^{(p)}$ as follows:

$$\boldsymbol{x}_r^{(p)} = \begin{pmatrix} \boldsymbol{x}_{r,1}^{(p)} \\ \vdots \\ \boldsymbol{x}_{r,v}^{(p)} \end{pmatrix}, \quad \text{where } \boldsymbol{x}_{r,t}^{(p)} = \begin{pmatrix} \boldsymbol{x}_{r,t1}^{(p)} \\ \vdots \\ \boldsymbol{x}_{r,tu}^{(p)} \end{pmatrix}, \text{ with } \boldsymbol{x}_{r,ts}^{(p)} = \begin{pmatrix} x_{r,ts,1}^{(p)} \\ \vdots \\ x_{r,ts,m}^{(p)} \end{pmatrix},$$

for $s = 1, \ldots, u, t = 1, \ldots, v$. The m-dimensional vector of measurements $\boldsymbol{x}_{r,ts}^{(p)}$ represents the rth replicate (individual) in the pth population on the sth location and at the tth time point.

Let $\boldsymbol{x}^{(p)}$ represent the muv-variate vector of all measurements corresponding to one individual in the pth population where we assume a distribution $N_{muv}\left(\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(p)}}\right)$, and let $\boldsymbol{x}_1^{(p)}, \ldots, \boldsymbol{x}_{n^{(p)}}^{(p)}$ be a random sample of size $n^{(p)}$ from this population. The unstructured variance–covariance matrix $\mathrm{Cov}\left[\boldsymbol{x}^{(p)}\right]$ has $q = muv\,(muv + 1)\,/2$ unknown parameters, which can be large for arbitrary values of m, u or v. In order to reduce the number of unknown parameters it is then essential to assume some appropriate structure on the variance–covariance matrix. One may assume a doubly exchangeable covariance structure in the situation, where the data is multivariate in three levels. Doubly exchangeable covariance structure consists of three $m \times m$-dimensional unstructured covariance matrices for three multivariate levels (Roy and Leiva, 2007). Thus, the resulting structure has only $3m\,(m+1)\,/2$ unknown parameters, which is much less than q. Moreover, this number does not even depend on the number of sites u and the number of time points v. The use of doubly exchangeable covariance structure provides a better insight into the three-level data structure. In the following sections we define the separable additive mean vector structure and the doubly exchangeable covariance structure respectively.

2.1. Separable additive mean vector structure

In the additive mean vector structure the mean for each of muv random variables $\boldsymbol{x}_r^{(p)}$, measured on the rth individual in the pth population can be expressed as

$$\mathrm{E}\big[x_r^{(p)}\big] = \mu_1^{(p)} z_1 + \cdots + \mu_m^{(p)} z_m + \lambda_1^{(p)} z_{(m+1)} + \cdots + \lambda_u^{(p)} z_{(m+u)} + \tau_1^{(p)} z_{(m+u+1)} + \cdots + \tau_v^{(p)} z_{(m+u+v)},$$

where $\boldsymbol{\tau}^{(p)} = \big(\tau_1^{(p)}, \ldots, \tau_v^{(p)}\big)' \in \Re^v$, $\boldsymbol{\lambda}^{(p)} = \big(\lambda_1^{(p)}, \ldots, \lambda_u^{(p)}\big)' \in \Re^u$, and $\boldsymbol{\mu}^{(p)} = \big(\mu_1^{(p)}, \ldots, \mu_m^{(p)}\big)' \in \Re^m$, with some identifiability constraints, for instance $\lambda_1^{(p)} = 0$ and $\tau_1^{(p)} = 0$. With these identifiability constraints the model reduces to

$$\mathrm{E}\big[x_r^{(p)}\big] = \mu_1^{(p)} z_1 + \cdots + \mu_m^{(p)} z_m + \lambda_2^{(p)} z_{(m+2)} + \cdots + \lambda_u^{(p)} z_{(m+u)} + \tau_2^{(p)} z_{(m+u+2)} + \cdots + \tau_v^{(p)} z_{(m+u+v)}.$$

This means that the mean vectors vary over time as well as over sites with suitably additive constants. The terms $z_1, \ldots, z_m$ are used to indicate the variable to which the mean $\mathrm{E}\big[x_r^{(p)}\big]$ belongs, the terms $z_{(m+1)}, \ldots, z_{(m+u)}$ are used to indicate the site to which the mean $\mathrm{E}\big[x_r^{(p)}\big]$ belongs, and finally, the terms $z_{(m+u+1)}, \ldots, z_{(m+u+v)}$ are used to indicate the time point to which the mean $\mathrm{E}\big[x_r^{(p)}\big]$ belongs. These indicator variables take the value one at a time and zero for others in each of these three levels. For example, if the mean $\mathrm{E}\big[x_r^{(p)}\big]$ comes from the first variable, then z_1 will be equal to one (corresponding to $\mu_1^{(p)}$, which represents the first variable effect), and $z_2, \ldots, z_m$ will all be equal to zero. Similarly, the other indicator variables for site and time effects. Note that, if $v = 1$, the measurements are taken only at one time point. Thus, there is no time effect, and as a result the above additive mean model reduces to

$$\begin{aligned}\mathrm{E}\big[x_r^{(p)}\big] &= \mu_1^{(p)} z_1 + \cdots + \mu_m^{(p)} z_m + \lambda_2^{(p)} z_{(m+2)} + \cdots + \lambda_u^{(p)} z_{(m+u)},\\ &= \mu_1^{(p)} z_1 + \cdots + \mu_m^{(p)} z_m + \lambda_1^{(p)} z_{(m+1)} + \cdots + \lambda_u^{(p)} z_{(m+u)}, \quad \text{with } \lambda_1^{(p)} = 0.\end{aligned}$$

A similar situation occurs for $u = 1$. Furthermore, if $u = 1$ and $v = 1$, the above mean model reduces to

$$\mathrm{E}\big[x_r^{(p)}\big] = \mu_1^{(p)} z_1 + \cdots + \mu_m^{(p)} z_m.$$

Thus, we see that our separable additive mean model generalizes the commonly used additive mean model for one-level multivariate data or univariate repeated measures data.

This additive mean model can be written using an alternative expression by applying the "Kronecker sum" of two vectors. Given the m-variate vector $\boldsymbol{\alpha} = (\alpha_h)$ and the n-variate vector $\boldsymbol{\beta} = (\beta_k)$, the Kronecker sum $\boldsymbol{\alpha} \oplus \boldsymbol{\beta}$ is the mn-variate vector $\boldsymbol{\gamma} = \big(\gamma_j\big)$, where

$$\gamma_j = \alpha_h + \beta_k \quad \text{if } j = (h-1)m + k,$$

for $h = 1, \ldots, m$, and $k = 1, \ldots, n$. That is, $\boldsymbol{\gamma} = \boldsymbol{\alpha} \oplus \boldsymbol{\beta}$ can be expressed as

$$\boldsymbol{\alpha} \oplus \boldsymbol{\beta} = \boldsymbol{\alpha} \otimes \mathbf{1}_n + \mathbf{1}_m \otimes \boldsymbol{\beta},$$

where $\mathbf{1}_a$ is the $(a \times 1)$-dimensional vector of ones. With this notation the separable additive mean model can be expressed as

$$\begin{aligned}\mathrm{E}\big[\boldsymbol{x}_r^{(p)}\big] &= \boldsymbol{\mu}_{\boldsymbol{x}^{(p)}} = \boldsymbol{\tau}^{(p)} \oplus \boldsymbol{\lambda}^{(p)} \oplus \boldsymbol{\mu}^{(p)}\\ &= \boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_u \otimes \mathbf{1}_m + \mathbf{1}_v \otimes \boldsymbol{\lambda}^{(p)} \otimes \mathbf{1}_m + \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}^{(p)},\end{aligned} \tag{1}$$

with $\tau_1^{(p)} = \lambda_1^{(p)} = 0$. Note that, if $\boldsymbol{\tau}^{(p)} = \mathbf{0}_v$ or $\boldsymbol{\lambda}^{(p)} = \mathbf{0}_u$, this model, as before, becomes the commonly used additive mean model.

Thus, we see that a separable additive mean model is more elegant than a separable multiplicative mean model as it can be generated by three simple separable multiplicative mean models. The first one $\boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_u \otimes \mathbf{1}_m$ means that the mean vector over time remains constant over both sites and variables. The second one $\mathbf{1}_v \otimes \boldsymbol{\lambda}^{(p)} \otimes \mathbf{1}_m$ means that the mean vector over sites remains constant over both time and variables. Similarly, the third one. As a result, the simple separable multiplicative mean models can be thought of as the basis for the separable additive mean model. Separable additive mean model extends the traditional classification rules to a more general platform, but at the expense of computational complexity. The number of unknown parameters for this separable additive mean model is $(v + u + m - 2)$ in the pth population.

2.2. Doubly exchangeable covariance structure

Definition 1. Let $\boldsymbol{x}_r$ be an muv-variate partitioned real-valued random vector $\boldsymbol{x}_r = (\boldsymbol{x}'_{r,1}, \ldots, \boldsymbol{x}'_{r,v})'$, where $\boldsymbol{x}_{r,t} = (\boldsymbol{x}'_{r,t1}, \ldots, \boldsymbol{x}'_{r,tu})'$ for $t = 1, \ldots, v$, and $\boldsymbol{x}'_{r,ts} = (x_{r,ts,1}, \ldots, x_{r,ts,m})'$ for $s = 1, \ldots, u$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} \in \Re^{muv}$ be the mean vector, and $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ be the $(muv \times muv)$-dimensional partitioned covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}} = \mathrm{Cov}\,[\boldsymbol{x}] = \left(\boldsymbol{\Gamma}_{\boldsymbol{x}_{r,t},\boldsymbol{x}_{r,t^*}}\right) = \left(\boldsymbol{\Gamma}_{r,tt^*}\right)$, where $\boldsymbol{\Gamma}_{r,tt^*} = \mathrm{Cov}\left[\boldsymbol{x}_{r,t}, \boldsymbol{x}_{r,t^*}\right]$ for $t, t^* = 1, \ldots, v$. The m-variate vectors $\boldsymbol{x}_{r,11}, \ldots, \boldsymbol{x}_{r,1u}, \ldots, \boldsymbol{x}_{r,v1}, \ldots, \boldsymbol{x}_{r,vu}$ are said to be jointly equicorrelated if $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is given by

$$\boldsymbol{\Gamma}_{\boldsymbol{x}} = \mathbf{I}_{vu} \otimes (\mathbf{U}_0 - \mathbf{U}_1) + \mathbf{I}_v \otimes \mathbf{J}_u \otimes (\mathbf{U}_1 - \mathbf{W}) + \mathbf{J}_{vu} \otimes \mathbf{W}, \tag{2}$$

where $\mathbf{U}_0$ is a positive definite symmetric $m \times m$ matrix, and $\mathbf{U}_1$ and $\mathbf{W}$ are symmetric $m \times m$ matrices. The variance–covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is then said to have a jointly equicorrelated covariance structure with equicorrelation parameters $\mathbf{U}_0, \mathbf{U}_1$ and $\mathbf{W}$. The matrices $\mathbf{U}_0, \mathbf{U}_1$ and $\mathbf{W}$ are all unstructured. The symbol $\mathbf{I}_a$ is the $a \times a$ identity matrix and $\mathbf{J}_a = \mathbf{1}_a\mathbf{1}'_a$. Because of the doubly exchangeable nature of this covariance structure, $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is also called doubly exchangeable covariance structure.

Thus, the vectors $\boldsymbol{x}_{r,11}, \ldots, \boldsymbol{x}_{r,1u}, \ldots, \boldsymbol{x}_{r,v1}, \ldots, \boldsymbol{x}_{r,vu}$ are jointly equicorrelated if they have the following "jointly equicorrelated covariance" matrix

$$\mathrm{Cov}\left[\boldsymbol{x}_{r,ts}; \boldsymbol{x}_{r,t^*s^*}\right] = \begin{cases} \mathbf{U}_0 & \text{if } t = t^* \text{ and } s = s^*, \\ \mathbf{U}_1 & \text{if } t = t^* \text{ and } s \neq s^*, \\ \mathbf{W} & \text{if } t \neq t^*. \end{cases}$$

The $m \times m$ diagonal blocks $\mathbf{U}_0$ represent the variance–covariance matrix of the m response variables at any given site and at any given time point, whereas the $m \times m$ off-diagonal blocks $\mathbf{U}_1$ represent the covariance matrix of the m response variables between any two sites and at any given time point. We assume $\mathbf{U}_0$ is constant for all sites and time points, and $\mathbf{U}_1$ is same between any two sites and for all time points. The $m \times m$ off-diagonal blocks $\mathbf{W}$ represent the covariance matrix of the m response variables between any two time points. It is assumed to be the same for any pair of time points, irrespective of the same site or between any two sites.

Lemma 1. *Let $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ be a doubly exchangeable covariance matrix as in Eq. (2) of Definition 1.*

1. *If*

$$\boldsymbol{\Delta}_1 = \mathbf{U}_0 - \mathbf{U}_1,$$
$$\boldsymbol{\Delta}_2 = \mathbf{U}_0 + (u-1)\,\mathbf{U}_1 - u\mathbf{W} = (\mathbf{U}_0 - \mathbf{U}_1) + u\,(\mathbf{U}_1 - \mathbf{W}), \quad \text{and}$$
$$\boldsymbol{\Delta}_3 = \mathbf{U}_0 + (u-1)\,\mathbf{U}_1 + u\,(v-1)\,\mathbf{W} = (\mathbf{U}_0 - \mathbf{U}_1) + u\,(\mathbf{U}_1 - \mathbf{W}) + uv\mathbf{W},$$

are nonsingular matrices, the matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is nonsingular, and its inverse is given by

$$\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} = \mathbf{I}_{vu} \otimes \mathbf{A} + \mathbf{I}_v \otimes \mathbf{J}_u \otimes \mathbf{B} + \mathbf{J}_{vu} \otimes \mathbf{C}, \tag{3}$$

where

$$\mathbf{A} = \boldsymbol{\Delta}_1^{-1}, \mathbf{B} = \frac{1}{u}\left(\boldsymbol{\Delta}_2^{-1} - \boldsymbol{\Delta}_1^{-1}\right), \quad \text{and} \quad \mathbf{C} = \frac{1}{vu}\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right). \tag{4}$$

2. *The determinant of $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is given by*

$$|\boldsymbol{\Gamma}_{\boldsymbol{x}}| = |\boldsymbol{\Delta}_1|^{v(u-1)}\,|\boldsymbol{\Delta}_2|^{(v-1)}\,|\boldsymbol{\Delta}_3|\,. \tag{5}$$

See Roy and Leiva (2007) for the proof of this lemma. These results are used in Section 3 to obtain the maximum likelihood estimate (MLE) of the doubly exchangeable covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$.

We study the efficacy of our new classification rule by comparing the misclassification error rates (MERs) when the actual mean vectors have separable additive structure, and we estimate and perform the classification analyses by assuming them as separable additive structure, separable multiplicative structure as well as the unstructured. We also compare the performance of our new classification rule with the traditional one, where both mean vector and variance–covariance matrix are unstructured. We see that the assumption of wrong structure on mean vectors leads to very high MERs. Thus, it is important to test the hypothesis on the actual structure on the mean vector before any classification analysis. Nonetheless, testing only helps when there is high probability of detecting a meaningful deviation from the tested hypothesis when such a deviation exists, and low probability of detecting each deviation that is less than meaningful. In our small sample size situations, we expect low power in detecting meaningful deviations from separable additive or separable multiplicative structure. Thus, one could look for evidence of those deviations from comparisons of likelihood, Akaike information criterion (AIC) or Bayesian information criterion (BIC). In this article, with the help of simulations, we study the effectiveness of our new linear classification rule.

3. Maximum likelihood estimates of the separable additive mean vector and the doubly exchangeable covariance matrix in a single population

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be an muv-variate random sample of size n from a population with distribution $N_{muv}\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right)$. We assume that the covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ has the doubly exchangeable or the jointly equicorrelated covariance structure as defined in (2) with equicorrelation parameters $\mathbf{U}_0, \mathbf{U}_1$ and $\mathbf{W}$. As in Definition 1 we partition the muv-variate vector $\boldsymbol{x}_r$ for $r = 1, \ldots, n$ as $\boldsymbol{x}_r = (\boldsymbol{x}'_{r,1}, \ldots, \boldsymbol{x}'_{r,v})'$, where $\boldsymbol{x}_{r,t} = (\boldsymbol{x}'_{r,t1}, \ldots, \boldsymbol{x}'_{r,tu})'$ for $t = 1, \ldots, v$, and with $\boldsymbol{x}'_{r,ts} = (x_{r,ts,1}, \ldots, x_{r,ts,m})'$ for $s = 1, \ldots, u$. In this case we assume that the jth component for $j = 1, \ldots, m$, of the mean vector $\mathrm{E}\left[\boldsymbol{x}_{r,ts}\right] = \boldsymbol{\mu}_{ts} = \left(\mu_{ts,j}\right)$ is the sum $\mu_{ts,j} = \tau_t + \lambda_s + \mu_j$, where $\boldsymbol{\mu} = \left(\mu_j\right) \in \Re^m$ and $\tau_t, \lambda_s \in \Re$, with the identifiability constraints $\tau_1 = \lambda_1 = 0$. That is, the jth component $\mu_{ts,j}$ of the mean vector $\boldsymbol{\mu}_{ts}$ of $\boldsymbol{x}_{r,ts}$ is discomposed into a sum of three summands: the jth component of a (base) vector $\boldsymbol{\mu} = \mathrm{E}\left[\boldsymbol{x}_{r,11}\right]$, plus λ_s (an effect due to site s) plus τ_t (an effect due to time t). Therefore, $\boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu}$ where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_v)' \in \Re^v$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_u)' \in \Re^u$, with $\tau_1 = \lambda_1 = 0$, and $\boldsymbol{\mu} \in \Re^m$. The following theorem yields the exact expressions for the MLEs of the separable additive mean vector $\boldsymbol{\mu}_{\boldsymbol{x}}$ and the doubly exchangeable covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$.

Theorem 1. *Under the above assumptions, the maximum likelihood estimates of* $\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\mu}$ *and* $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ *are given by*

$$\widehat{\boldsymbol{\tau}} = \mathbf{D}_{\boldsymbol{\lambda\mu}}\left(\overline{\boldsymbol{x}} - \mathbf{1}_v \otimes \boldsymbol{\lambda} \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}\right), \tag{6}$$

$$\widehat{\boldsymbol{\lambda}} = \mathbf{D}_{\boldsymbol{\tau\mu}}\left(\overline{\boldsymbol{x}} - \boldsymbol{\tau} \otimes \mathbf{1}_v \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}\right), \tag{7}$$

$$\widehat{\boldsymbol{\mu}} = \mathbf{D}_{\boldsymbol{\tau\lambda}}\left(\overline{\boldsymbol{x}} - \boldsymbol{\tau} \otimes \mathbf{1}_v \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \boldsymbol{\lambda} \otimes \mathbf{1}_m\right), \tag{8}$$

and

$$\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}} = \mathbf{I}_{vu} \otimes \left(\widehat{\mathbf{U}}_0 - \widehat{\mathbf{U}}_1\right) + \mathbf{I}_v \otimes \mathbf{J}_u \otimes \left(\widehat{\mathbf{U}}_1 - \widehat{\mathbf{W}}\right) + \mathbf{J}_{vu} \otimes \widehat{\mathbf{W}}, \tag{9}$$

where $\mathbf{D}_{\boldsymbol{\lambda\mu}}, \mathbf{D}_{\boldsymbol{\tau\mu}}$ *and* $\mathbf{D}_{\boldsymbol{\tau\lambda}}$ *are given in* (A.2)–(A.4) *respectively in* Appendix A, *and*

$$\widehat{\mathbf{U}}_0 = \frac{1}{nuv}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)', \tag{10}$$

$$\widehat{\mathbf{U}}_1 = \frac{1}{nvu\,(u-1)}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s\neq s^*=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,ts^*} - (\tau_t + \lambda_{s^*})\,\mathbf{1}_m - \boldsymbol{\mu}\right)', \tag{11}$$

and

$$\widehat{\mathbf{W}} = \frac{1}{nu^2 v\,(v-1)}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{t\neq t^*=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,t^*s^*} - (\tau_{t^*} + \lambda_{s^*})\,\mathbf{1}_m - \boldsymbol{\mu}\right)'. \tag{12}$$

The proof of this theorem, which is simple but tedious, is given in Appendix A. We see that the MLEs of $\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{U}_0, \mathbf{U}_1$ and $\mathbf{W}$ have implicit equations, and therefore are not tractable analytically. The computation of the MLEs of $\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{U}_0, \mathbf{U}_1$ and $\mathbf{W}$ can be carried out by solving the above implicit equations simultaneously by the following fixed point iteration algorithm.

Iteration algorithm

We simultaneously calculate the maximum likelihood estimates (MLEs) of a total of $(v + u + m - 2) + 3m\,(m+1)\,/2$ unknown parameters in the separable additive mean vector and the doubly exchangeable variance–covariance matrix. The solutions satisfy the fully implicit and coupled equations with $\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{U}_0, \mathbf{U}_1$ and $\mathbf{W}$.

Algorithm outline:

Step 1: Calculate the global sample mean $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{r=1}^{n}\boldsymbol{x}_r$ as

$$\overline{\boldsymbol{x}} = \left(\overline{\boldsymbol{x}}'_{1,1}, \ldots, \overline{\boldsymbol{x}}'_{1,u}, \ldots, \overline{\boldsymbol{x}}'_{v,1}, \ldots, \overline{\boldsymbol{x}}'_{v,u}\right)',$$

where $\overline{\boldsymbol{x}}_{ts} = \frac{1}{n}\sum_{r=1}^{n}\boldsymbol{x}_{r,ts}$, for $t = 1, \ldots, v$, $s = 1, \ldots, u$. The initial value $\widehat{\boldsymbol{\mu}}^0$ of $\widehat{\boldsymbol{\mu}}$ is taken as $\widehat{\boldsymbol{\mu}}^0 = \overline{\boldsymbol{x}}_{1,1}$. Assume the initial values of $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ as $\boldsymbol{\tau}^0 = \mathbf{1}_v$ and $\boldsymbol{\lambda}^0 = \mathbf{1}_u$ respectively.

Step 2: Compute $\widehat{\mathbf{H}}_1, \widehat{\mathbf{H}}_2$ and $\widehat{\mathbf{H}}_3$ from (A.6)–(A.8) respectively.

Step 3: Compute $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}$ and $\widehat{\mathbf{C}}$ from (A.9).

Step 4: Compute the estimate $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1}$ from (A.10).

Step 5: Compute $\mathbf{D}_{\boldsymbol{\lambda\mu}}, \mathbf{D}_{\boldsymbol{\tau\mu}}$ and $\mathbf{D}_{\boldsymbol{\tau\lambda}}$ using $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1}$ in Step 4, from (A.2)–(A.4), respectively.

Step 6: Compute the estimate $\widehat{\boldsymbol{\mu}}$ from (8).

Step 7: Compute the estimate $\widehat{\boldsymbol{\lambda}}$ from (7) using the estimate $\widehat{\boldsymbol{\mu}}$ in Step 6.

Step 8: Compute the estimate $\widehat{\boldsymbol{\tau}}$ from (6) using the estimates $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\lambda}}$ in Steps 6 and 7 respectively.

Step 9: Compute $\widehat{\mathbf{U}}_0$, $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{W}}$ from (10)–(12) respectively.

Step 10: Repeat Steps 2–9 until convergence is attained. This is ensured by verifying if the maximum of the absolute difference among the L_1 distance between two successive values of $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\lambda}}$, and $\widehat{\boldsymbol{\tau}}$, and the absolute difference among the two successive values of trace of $\widehat{\mathbf{U}}_0$, $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{W}}$ is less than a pre-determined number ϵ. For our calculation we choose $\epsilon = 0.000001$.

3.1. Remarks

It is also possible to use the method of moments (MOM) estimates instead of MLEs. However, computations of MOM estimates $\widetilde{\boldsymbol{\tau}}$, $\widetilde{\boldsymbol{\lambda}}$ and $\widetilde{\boldsymbol{\mu}}$ are very intricate as the mean model becomes overparameterized. The greatest problem faced by MOM is the overparameterization. When we substitute $\mathrm{E}(\boldsymbol{x})$ with the first sample moment $\bar{\boldsymbol{x}}$, it leads to an over parameterization. The number of unknown parameters in the separable additive mean model is $(v + u + m - 2)$. Nevertheless, $\mathrm{E}(\boldsymbol{x}) \approx \bar{\boldsymbol{x}}$ gives a system of muv sub-equations. Since the structured additive means are sums of individual components of parameters as described in Section 2.1, a MOM approach that easily separates the parameters and handles their additive structure is a regression problem. To get MOM estimates of $(v + u + m - 2)$ unknown parameters one can write the system of muv sub-equations in a regression form as $\boldsymbol{X}\widetilde{\boldsymbol{\beta}} = \bar{\boldsymbol{x}}$, where $\widetilde{\boldsymbol{\beta}}$ is a MOM estimate vector of all $(v+u+m-2)$ unknown parameters in the separable additive mean model and $\boldsymbol{X}$ is the design matrix of zeros and ones. We thus get a MOM estimate $\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\bar{\boldsymbol{x}}$. From this estimate $\widetilde{\boldsymbol{\beta}}$, we can construct $\mathrm{E}\left[\boldsymbol{x}_r\right] = \widetilde{\boldsymbol{\mu}}_{\boldsymbol{x}} = \widetilde{\boldsymbol{\tau}} \oplus \widetilde{\boldsymbol{\lambda}} \oplus \widetilde{\boldsymbol{\mu}}$. By partitioning this vector as in Section 2 we find that $\mathrm{E}(\boldsymbol{x}_{r,ts}) = (\widetilde{\tau}_t + \widetilde{\lambda}_s)\mathbf{1}_m + \widetilde{\boldsymbol{\mu}}$. Now, substituting $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}')$ with the second sample moment $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$, we get the MOM estimate $\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' - \left(\widetilde{\boldsymbol{\tau}} \oplus \widetilde{\boldsymbol{\lambda}} \oplus \widetilde{\boldsymbol{\mu}}\right)\left(\widetilde{\boldsymbol{\tau}} \oplus \widetilde{\boldsymbol{\lambda}} \oplus \widetilde{\boldsymbol{\mu}}\right)'$. MOM estimates of $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$ can be found by suitably partitioning the matrix $\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{x}}$ and by taking appropriate averages of those partitioned matrices. The estimates $\widetilde{\mathbf{U}}_0$, $\widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{W}}$ are identical to Eqs. (10)–(12) respectively, but the component of means are replaced by the corresponding component of MOM estimates. Since we assume multivariate normal distribution, we have infinite support, thus we need not have to worry about the MOM estimates to lie outside the parameter space. The notable thing about the MOM estimates is that the parameters have closed form solutions. As a result, one need not have to use any iterative algorithm to compute them.

4. Discrimination with separable additive mean vectors and doubly exchangeable covariance matrix

In this section we derive the Bayesian linear decision rule for k populations with separable additive mean vectors. Using the same notations as in the introduction, we assume that the vectors $\boldsymbol{x}_{r,11}^{(p)}, \ldots, \boldsymbol{x}_{r,1u}^{(p)}, \ldots, \boldsymbol{x}_{r,v1}^{(p)}, \ldots, \boldsymbol{x}_{r,vu}^{(p)}$ are jointly equicorrelated with equicorrelation parameters $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$, and with separable additive mean vector $\mathrm{E}[\boldsymbol{x}_r^{(p)}] = \boldsymbol{\mu}_{\boldsymbol{x}^{(p)}} = \boldsymbol{\tau}^{(p)} \oplus \boldsymbol{\lambda}^{(p)} \oplus \boldsymbol{\mu}^{(p)}$, where $\boldsymbol{\tau}^{(p)} = \left(\tau_1^{(p)}, \ldots, \tau_v^{(p)}\right)' \in \Re^v$, $\boldsymbol{\lambda}^{(p)} = \left(\lambda_1^{(p)}, \ldots, \lambda_u^{(p)}\right)' \in \Re^u$, with $\tau_1^{(p)} = \lambda_1^{(p)} = 0$, and $\boldsymbol{\mu}^{(p)} \in \Re^m$. Let $\boldsymbol{x}_1^{(p)}, \ldots, \boldsymbol{x}_{n^{(p)}}^{(p)}$ be a random sample of size $n^{(p)}$ from the pth population with distribution $N_{muv}\left(\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right)$, for $p = 1, \ldots, k$. These k random training samples are independent among each other.

Now we consider the problem of assigning a new individual with muv-variate partitioned measurement vector $\boldsymbol{x}_o$ to one of the k classes in a Bayesian framework. The previous set-up leads to a linear discriminant function as follows:

Under the assumptions of equal prior probabilities and equal costs of misclassification, the sample classification rule is given by

Allocate an individual with response $\boldsymbol{x}_o$ to population i if

$$\widehat{l}^{(i)}(\boldsymbol{x}_o) = \text{largest of } \left\{\widehat{l}^{(p)}(\boldsymbol{x}_o) : p = 1, \ldots, k\right\}, \quad \text{for } i = 1, \ldots, k, \tag{13}$$

where the sample linear score $\widehat{l}^{(p)}$ is defined by

$$\widehat{l}^{(p)}(\boldsymbol{x}_o) = \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(p)}}' \cdot \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1} \cdot \boldsymbol{x}_o - \frac{1}{2}\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(p)}}' \cdot \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(p)}}, \tag{14}$$

and $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1}$ and $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(p)}} = \widehat{\boldsymbol{\tau}}^{(p)} \oplus \widehat{\boldsymbol{\lambda}}^{(p)} \oplus \widehat{\boldsymbol{\mu}}^{(p)}$, for $p = 1, \ldots, k$, are the MLEs of $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}$ respectively. $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1}$ is obtained from (3) by substituting the values of $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$ by $\widehat{\mathbf{U}}_0$, $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{W}}$ from (20)–(22) respectively, and $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(p)}}$ is obtained by substituting the values of $\widehat{\boldsymbol{\tau}}^{(p)}$, $\widehat{\boldsymbol{\lambda}}^{(p)}$ and $\widehat{\boldsymbol{\mu}}^{(p)}$ from (17)–(19) respectively. This linear rule (13) has been extensively studied by many authors. See McLachlan (1992). The theoretical linear score corresponding to (14) can be written as a function of $\boldsymbol{\tau}^{(p)}$, $\boldsymbol{\lambda}^{(p)}$, $\boldsymbol{\mu}^{(p)}$, $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$ as follows

$$l^{(p)}(\boldsymbol{x}_o) = m^{(p)}(\boldsymbol{x}_o) - \frac{1}{2}\kappa^{(p)}, \tag{15}$$

where

$$\begin{aligned} m^{(p)}(\boldsymbol{x}_o) &= \boldsymbol{\mu}'_{\boldsymbol{x}^{(p)}} \cdot \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \cdot \boldsymbol{x}_o, \\ &= \left(\mathbf{1}'_m \boldsymbol{\Delta}_2^{-1}\right) \sum_{t=1}^{v} u\tau_t^{(p)} \overline{\boldsymbol{x}}_{o,t\bullet} + vu\overline{\tau}^{(p)} \left(\mathbf{1}'_m \left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right)\right) \overline{\boldsymbol{x}}_o \\ &\quad + \left(\mathbf{1}'_m \boldsymbol{\Delta}_1^{-1}\right) \sum_{s=1}^{u} v\lambda_s^{(p)} \overline{\boldsymbol{x}}_{o,\bullet s} + vu\overline{\lambda}^{(p)} \left(\mathbf{1}'_m \left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right) \overline{\boldsymbol{x}}_o + vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \overline{\boldsymbol{x}}_o, \end{aligned}$$

$$\begin{aligned} \text{and} \quad \kappa^{(p)} &= \boldsymbol{\mu}'_{\boldsymbol{x}^{(p)}} \cdot \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \cdot \boldsymbol{\mu}_{\boldsymbol{x}^{(p)}} \\ &= vud_{2+} \left(\overline{\tau^2}^{(p)} - \left(\overline{\tau}^{(p)}\right)^2\right) + vud_{1+} \left(\overline{\lambda^2}^{(p)} - \left(\overline{\lambda}^{(p)}\right)^2\right) \\ &\quad + vud_{3+} \left(\overline{\tau}^{(p)} + \overline{\lambda}^{(p)}\right)^2 + 2vu \left(\overline{\tau}^{(p)} + \overline{\lambda}^{(p)}\right) \mathbf{1}'_m \boldsymbol{\Delta}_3^{-1} \boldsymbol{\mu}^{(p)} + vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \boldsymbol{\mu}^{(p)}, \end{aligned} \tag{16}$$

with $\overline{\boldsymbol{x}}_{o,t\bullet} = \frac{1}{u}\sum_{s=1}^{u} \boldsymbol{x}_{o,ts}$, $\overline{\boldsymbol{x}}_{o,\bullet s} = \frac{1}{v}\sum_{t=1}^{v} \boldsymbol{x}_{o,ts}$, $\overline{\boldsymbol{x}}_o = \frac{1}{vu}\sum_{t=1}^{v}\sum_{s=1}^{u} \boldsymbol{x}_{o,ts}$, $\overline{\tau}^{(p)} = \frac{1}{v}\sum_{t=1}^{v} \tau_t^{(p)}$, $\overline{\lambda}^{(p)} = \frac{1}{u}\sum_{s=1}^{u} \lambda_s^{(p)}$, $\overline{\tau^2}^{(p)} = \frac{1}{v}\sum_{t=1}^{v} \left(\tau_t^{(p)}\right)^2$, $\overline{\lambda^2}^{(p)} = \frac{1}{u}\sum_{s=1}^{u} \left(\lambda_s^{(p)}\right)^2$, and d_{j+} represents the sum of all the elements of $\boldsymbol{\Delta}_j^{-1}$, for $j = 1, 2, 3$. For detail see Appendix B. It is worthwhile to consider different cases for structured mean and variance–covariance matrix of this theoretical linear score (15) as follows.

Case 1: If we assume that the mean $\mathrm{E}\left[\boldsymbol{x}_{r,ts}^{(p)}\right] = \boldsymbol{\mu}_{ts}^{(p)}$ is constant for each time point $t = 1, \ldots, v$ (depends on site, but not on time), that is, $\boldsymbol{\tau}^{(p)} = \mathbf{0}_v$, then theoretical linear score (15) has the following summands:

$$m^{(p)}(\boldsymbol{x}_o) = \left(\mathbf{1}'_m \boldsymbol{\Delta}_1^{-1}\right) \sum_{s=1}^{u} v\lambda_s^{(p)} \overline{\boldsymbol{x}}_{o,\bullet s} + vu\overline{\lambda}^{(p)} \left(\mathbf{1}'_m \left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right) \overline{\boldsymbol{x}}_o + vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \overline{\boldsymbol{x}}_o,$$

and

$$\kappa^{(p)} = vud_{1+} \left(\overline{\lambda^2}^{(p)} - \left(\overline{\lambda}^{(p)}\right)^2\right) + vud_{3+} \left(\overline{\lambda}^{(p)}\right)^2 + 2vu\overline{\lambda}^{(p)} \mathbf{1}'_m \boldsymbol{\Delta}_3^{-1} \boldsymbol{\mu}^{(p)} + vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \boldsymbol{\mu}^{(\mathbf{p})}.$$

Case 2: If we assume that the mean $\mathrm{E}\left[\boldsymbol{x}_{r,ts}^{(p)}\right] = \boldsymbol{\mu}_{ts}^{(p)}$ is constant for each site $s = 1, \ldots, u$ (depends on time, but not on site), that is, $\boldsymbol{\lambda}^{(p)} = \mathbf{0}_u$, then theoretical linear score (15) has the following summands:

$$m^{(p)}(\boldsymbol{x}_o) = \left(\mathbf{1}'_m \boldsymbol{\Delta}_2^{-1}\right) \sum_{t=1}^{v} u\tau_t^{(p)} \overline{\boldsymbol{x}}_{o,t\bullet} + vu\overline{\tau}^{(p)} \left(\mathbf{1}'_m \left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right)\right) \overline{\boldsymbol{x}}_o + vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \overline{\boldsymbol{x}}_o,$$

and

$$\kappa^{(p)} = vud_{2+} \left(\overline{\tau^2}^{(p)} - \left(\overline{\tau}^{(p)}\right)^2\right) + vud_{3+} \left(\overline{\tau}^{(p)}\right)^2 + 2vu \left(\overline{\tau}^{(p)}\right) \mathbf{1}'_m \boldsymbol{\Delta}_3^{-1} \boldsymbol{\mu}^{(p)} + vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \boldsymbol{\mu}^{(\mathbf{p})}.$$

Case 3: Finally, if we assume that $\mathrm{E}\left[\boldsymbol{x}_{r,ts}^{(p)}\right] = \boldsymbol{\mu}_{ts}^{(p)}$ is constant for each time point $t = 1, \ldots, v$, and for each site $s = 1, \ldots, u$, that is, $\boldsymbol{\tau}^{(p)} = \mathbf{0}_v$, and $\boldsymbol{\lambda}^{(p)} = \mathbf{0}_u$, then the theoretical linear score (15) reduces to

$$l^{(p)}(\boldsymbol{x}_o) = vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \overline{\boldsymbol{x}}_o - \frac{1}{2} vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \boldsymbol{\mu}^{(p)}.$$

In particular, if $u = 1$ and $v = 1$, the above theoretical score reduces to the traditional score with the covariance matrix $\mathbf{U}_0$, i.e.,

$$l^{(p)}(\boldsymbol{x}_o) = \boldsymbol{\mu}^{(p)\prime} \mathbf{U}_0^{-1} \overline{\boldsymbol{x}}_o - \frac{1}{2} \boldsymbol{\mu}^{(p)\prime} \mathbf{U}_0^{-1} \boldsymbol{\mu}^{(p)}.$$

Thus, we see that our new classification rule is indeed an extension of the traditional classification rule.

Case 4: Similarly, we can consider particular cases of the covariance matrix. For instance, if the vectors are only equicorrelated, or in other words if the vectors are only exchangeable, that is, $\mathbf{U}_1 = \mathbf{W}$, then $\boldsymbol{\Delta}_1 = \boldsymbol{\Delta}_2$, and $\boldsymbol{\Delta}_3 = \mathbf{U}_0 + (uv - 1)\,\mathbf{U}_1$, and the first summand of the linear score (15) reduces to

$$\begin{aligned} m^{(p)}(\boldsymbol{x}_o) &= \boldsymbol{\mu}'_{\boldsymbol{x}^{(p)}} \cdot \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \cdot \boldsymbol{x}_o \\ &= \left(\mathbf{1}'_m \boldsymbol{\Delta}_1^{-1}\right) \left[\sum_{t=1}^{v} u\tau_t^{(p)} \overline{\boldsymbol{x}}_{o,t\bullet} + \sum_{s=1}^{u} v\lambda_s^{(p)} \overline{\boldsymbol{x}}_{o,\bullet s}\right] + vu \left(\overline{\tau}^{(p)} + \overline{\lambda}^{(p)}\right) \left(\mathbf{1}'_m \left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right) \overline{\boldsymbol{x}}_o + vu\boldsymbol{\mu}^{(p)\prime} \boldsymbol{\Delta}_3^{-1} \overline{\boldsymbol{x}}_o. \end{aligned}$$

However, the second summand remains identical to (16). In real life applications most of the times the parameters are unknown, and they must be estimated from the data itself. This is done by taking random samples from each population.

Maximum likelihood estimates of $\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}$, $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$:

The log likelihood function can be written as $\ln(L) = \sum_{p=1}^{k} \ln(L^{(p)})$, where

$$\ln(L^{(p)}) = -\frac{n^{(p)}muv}{2}\ln(2\pi) - \frac{n^{(p)}}{2}\ln|\boldsymbol{\Gamma}_{\boldsymbol{x}}| - \sum_{r=1}^{n^{(p)}}\left(\boldsymbol{x}_r^{(p)} - \boldsymbol{\mu}_{\boldsymbol{x}}^{(p)}\right)'\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\left(\boldsymbol{x}_r^{(p)} - \boldsymbol{\mu}_{\boldsymbol{x}}^{(p)}\right).$$

We get the MLEs of $\boldsymbol{\tau}^{(p)}$, $\boldsymbol{\lambda}^{(p)}$ and $\boldsymbol{\mu}^{(p)}$, for $p = 1, \ldots, k$, from the above log likelihood function using the similar arguments as in Theorem 1 with different separable additive means and a common doubly exchangeable covariance matrix for k populations as follows

$$\widehat{\boldsymbol{\tau}}^{(p)} = \mathbf{D}_{\lambda\mu}\left(\overline{\boldsymbol{x}}^{(p)} - \mathbf{1}_v \otimes \boldsymbol{\lambda}^{(p)} \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}^{(p)}\right), \tag{17}$$

$$\widehat{\boldsymbol{\lambda}}^{(p)} = \mathbf{D}_{\tau\mu}\left(\overline{\boldsymbol{x}}^{(p)} - \boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_v \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}^{(p)}\right), \tag{18}$$

and

$$\widehat{\boldsymbol{\mu}}^{(p)} = \mathbf{D}_{\tau\lambda}\left(\overline{\boldsymbol{x}}^{(p)} - \boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_v \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \boldsymbol{\lambda}^{(p)} \otimes \mathbf{1}_m\right), \tag{19}$$

where the expressions of the matrix factors $\mathbf{D}_{\lambda\mu}$, $\mathbf{D}_{\tau\mu}$, and $\mathbf{D}_{\tau\lambda}$, in these equations are given in (A.2)–(A.4) respectively in the Appendix A. An estimate of $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}$ in these matrix factors can be calculated from the estimates of $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$. Now, The ML estimates of $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$ are given by

$$\widehat{\mathbf{U}}_0 = \frac{1}{nuv}\sum_{p=1}^{k}\sum_{r=1}^{n^{(p)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)', \tag{20}$$

$$\widehat{\mathbf{U}}_1 = \frac{1}{nuv(u-1)}\sum_{p=1}^{k}\sum_{r=1}^{n^{(p)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s\neq s^*=1}^{u}\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)\left(\boldsymbol{x}_{r,ts^*}^{(p)} - \left(\tau_t^{(p)} + \lambda_{s^*}^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)', \tag{21}$$

and

$$\widehat{\mathbf{W}} = \frac{1}{nu^2v(v-1)}\sum_{p=1}^{k}\sum_{r=1}^{n^{(p)}}\sum_{t=1}^{v}\sum_{t\neq t^*=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)\left(\boldsymbol{x}_{r,t^*s^*}^{(p)} - \left(\tau_{t^*}^{(p)} + \lambda_{s^*}^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)', \tag{22}$$

where (in this section) n denotes $n = \sum_{p=1}^{k} n^{(p)}$. The expressions of $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$ to estimate $\widehat{\mathbf{U}}_0$, $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{W}}$ are given by

$$\mathbf{H}_1 = \sum_{p=1}^{k}\sum_{r=1}^{n^{(p)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)',$$

$$\mathbf{H}_2 = \sum_{p=1}^{k}\sum_{r=1}^{n^{(p)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)\left(\boldsymbol{x}_{r,ts^*}^{(p)} - \left(\tau_t^{(p)} + \lambda_{s^*}^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)',$$

and

$$\mathbf{H}_3 = \sum_{p=1}^{k}\sum_{r=1}^{n^{(p)}}\sum_{t=1}^{v}\sum_{t^*=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,ts}^{(p)} - \left(\tau_t^{(p)} + \lambda_s^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)\left(\boldsymbol{x}_{r,t^*s^*}^{(p)} - \left(\tau_{t^*}^{(p)} + \lambda_{s^*}^{(p)}\right)\mathbf{1}_m - \boldsymbol{\mu}^{(p)}\right)'.$$

We see that the MLEs $\widehat{\boldsymbol{\mu}}^{(p)}$, $\widehat{\boldsymbol{\lambda}}^{(p)}$, $\widehat{\boldsymbol{\tau}}^{(p)}$, $\widehat{\mathbf{U}}_0$, $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{W}}$ do not have closed form solutions. These estimates are obtained by using a similar fixed point iteration algorithm as described in Section 3. The MLE $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}$ of $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is given in (9).

4.1. Remarks

In this article we consider only the linear classification rule where the doubly exchangeable covariance matrix is identical for all k populations. If instead the doubly exchangeable covariance parameters were different among these populations, the best classification rule would be the one using quadratic scores. However, the linear classification rule still has several advantages over the quadratic one. For instance, the distributional theory for the linear rule is simple, and may allow us to calculate the theoretical misclassification probabilities without difficulty. From a geometric point of view, the quadratic classification rule divides the sample space using k complicated quadratic surfaces, while the linear rule uses only hyper-planes. Moreover, the linear rule is robust under departures from the normality assumption, while the quadratic rule is very

Table 1
MERs (%) for glaucoma data for different training sample sizes.

Assumed mean vector	$(\text{trn}^{(1)}, \text{trn}^{(2)})$					
	(3, 3)	(5, 5)	(8, 8)	(13, 13)	(15, 15)	LOOCV
Additive	10.000	6.083	4.583	3.750	3.417	3.333
Multiplicative	17.250	13.417	9.417	4.583	4.000	1.667
Unstructured	16.167	10.750	8.667	5.833	5.750	3.333
CMVOS	7.667	5.167	4.083	3.667	3.333	3.333
Traditional	Failed	Failed	17.417	8.167	6.667	6.667

sensitive to the violations of this assumption. Due to these and many other reasons several authors (Chaudhuri et al., 1991; Park and Kshirsagar, 1994; Paranjpe and Gore, 1994) have considered linear classification rules with different covariance structures. In the two population case, most of their solutions use some linear combination of the two different covariance matrices from two populations to carry out the linear classification (Leiva and Herrera, 1999). It is easy to see that a linear combination of doubly exchangeable matrices is also doubly exchangeable. All these motivations inspire us to develop the linear classification rule in this article. However, since each summand in the log likelihood function $\ln(L) = \sum_{p=1}^{k} \ln(L^{(p)})$ has different parameters, one can calculate the ML estimates of all the parameters in the different covariance model using Theorem 1. More precisely, the theoretical quadratic scores in this case are

$$q^{(p)}(\boldsymbol{x}_o) = -\frac{1}{2}\left(\boldsymbol{x}_o - \boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}\right)' \cdot \boldsymbol{\Gamma}_{\boldsymbol{x}}^{(p)-1} \cdot \left(\boldsymbol{x}_o - \boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}\right) - \frac{1}{2}\ln\left|\boldsymbol{\Gamma}_{\boldsymbol{x}}^{(p)}\right|, \quad \text{for } p = 1, \ldots, k,$$

and the sample classification rule is given by
Allocate an individual with response $\boldsymbol{x}_o$ to population i if

$$\widehat{q}^{(i)}(\boldsymbol{x}_o) = \text{largest of } \left\{\widehat{q}^{(p)}(\boldsymbol{x}_o) : p = 1, \ldots, k\right\}, \quad \text{for } i = 1, \ldots, k, \tag{23}$$

where $\widehat{q}^{(p)}(\boldsymbol{x}_o)$ is obtained by replacing the required parameters in the expression of $q^{(p)}(\boldsymbol{x}_o)$ by their corresponding ML estimates which are obtained by using the same iteration algorithm outlined in Section 3. This is because the equations for calculating these ML estimates are obtained directly from Theorem 1 by replacing n, $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\mathbf{U}_0$, $\mathbf{U}_1$, $\mathbf{W}$, and $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}$ by $n^{(p)}$, $\boldsymbol{\tau}^{(p)}$, $\boldsymbol{\lambda}^{(p)}$, $\boldsymbol{\mu}^{(p)}$, $\mathbf{U}_0^{(p)}$, $\mathbf{U}_1^{(p)}$, $\mathbf{W}^{(p)}$, and $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{(p)-1}$ respectively, for $p = 1, \ldots, k$.

5. A real data example

To show the efficacy of our new classification rule we apply it to a medical data set, where the interest is in detecting the possible cases of glaucoma. We have $n^{(1)} = 30$ and $n^{(2)} = 30$ samples from two populations, the diseased patients with glaucoma (Pop1) and the individuals who do not have glaucoma (Pop2). Measurements of intraocular pressure (IOP) and central corneal thickness (CCT) were obtained from both the eyes (sites), each at three time points at an interval of three months. It is clear that for this data set $m = 2$, $u = 2$ and $v = 3$. The problem is to classify an unknown individual into one of the two populations using the new classification rule (13) as discussed in Section 4. Table 1 shows the misclassification error rates (MERs) of our new classification rule. This data set was analyzed before by assuming the same doubly exchangeable covariance structure, but with separable multiplicative and unstructured mean vectors (Leiva and Roy, 2009b, 2011) and with constant mean vector over sites (CMVOS) (Roy and Leiva, 2007). For comparison purpose MERs from these previous studies, i.e., for separable multiplicative mean vector, unstructured mean vector and CMVOS are also presented in Table 1. Additionally, MERs for the traditional linear classification rule (Seber, 1984, p. 293, 297, Johnson and Wichern, 2007, p. 586, 594), where both the mean vector and variance–covariance matrix are unstructured are also presented in Table 1. Our main aim of the analysis of this data set is to illustrate the computational facets of our new classification rule rather than giving any insight into the data set itself.

For this data set the unstructured variance–covariance matrix is (12×12)-dimensional; thus, the number of unknown parameters in the unstructured covariance matrix is 78. Therefore, estimation of the pooled unstructured variance–covariance matrix is not possible for small samples less than or equal to a total of 12 samples from both the populations. Thus, an assumption of structured covariance matrix is necessary for small sample situation. Our doubly exchangeable covariance structure has three (2×2)-dimensional unstructured covariance matrices, which gives only 9 unknown covariance parameters to estimate. Thus, estimation of these three unstructured covariance matrices is possible with only a total of three samples from both the populations. For this data set the separable additive mean vectors have 10 unknown parameters in the two populations which are simultaneously estimated along with 9 unknown covariance parameters, i.e. a total of 19 unknown parameters, using the fixed point iteration algorithm presented in Section 3.

To calculate the MERs, the data are split into a training set, upon which the classification rule is developed, and a test set upon which the classification rule is tested. The analysis is done with a number of pairs of training sample sizes $(\text{trn}^{(1)}, \text{trn}^{(2)}) = (3, 3), (5, 5), (8, 8), (13, 13)$ and $(15, 15)$, i.e., from very small to moderate pairs of training sample sizes, from the two populations. For each of these pairs of training samples, a pair of samples (15, 15) is randomly selected from the remaining data points and used as a test set. Based on the training samples, we simultaneously estimate the parameters

$(\boldsymbol{\tau}^{(1)}, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\tau}^{(2)}, \boldsymbol{\lambda}^{(2)}, \boldsymbol{\mu}^{(2)})$ and $(\mathbf{U}_0, \mathbf{U}_1, \mathbf{W})$ by using the maximum likelihood method as discussed in Section 4, and the sample classification rule is developed in the first stage. In the second stage the test samples are used to assess the efficacy of the new classification rule. To avoid the possible instability introduced by relying on just one single particular split into training and test sets, 40 different splits for each pair of training sample sizes are chosen. On the basis of these 40 different training samples, the unknown population parameters are estimated by the ML method. Using these estimates, the classification analysis is performed separately on each of the corresponding 40 test samples, and the MERs for these 40 choices are averaged. This method gives an error rate that is more stable than just doing it on one single particular split.

We have also used the leave-one-out cross validation (LOOCV) method which is normally regarded as more reliable than any other method to calculate the MERs. Leave-one-out cross validation removes each observation, constructs the classification rule from the remaining $n-1$ observations, and then computes whether the deleted observation is correctly classified by leave-one-out classification rule. The average MER on the deleted observation over the n possible ways is computed to evaluate the performance of the classification rule.

The MLEs of $(\boldsymbol{\tau}^{(1)}, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\mu}^{(1)})$ and $(\boldsymbol{\tau}^{(2)}, \boldsymbol{\lambda}^{(2)}, \boldsymbol{\mu}^{(2)})$ in separable additive mean vector case for the pair of training sample sizes $(15, 15)$ are

$$\widehat{\boldsymbol{\tau}}^{(1)} = [0, -3.6912, -3.1291]', \qquad \widehat{\boldsymbol{\lambda}}^{(1)} = [0, 0.0217]', \qquad \widehat{\boldsymbol{\mu}}^{(1)} = [24.2848, 534.2959]',$$

and

$$\widehat{\boldsymbol{\tau}}^{(2)} = [0, 1.1162, 0.5590]', \qquad \widehat{\boldsymbol{\lambda}}^{(2)} = [0, 0.8351]' \quad \text{and} \quad \widehat{\boldsymbol{\mu}}^{(2)} = [12.8907, 502.1018]',$$

respectively. The MLEs of $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$ are

$$\widehat{\mathbf{U}}_0 = \begin{bmatrix} 9.445 & 9.779 \\ 9.779 & 322.068 \end{bmatrix}, \qquad \widehat{\mathbf{U}}_1 = \begin{bmatrix} 5.778 & 7.500 \\ 7.500 & -18.690 \end{bmatrix} \quad \text{and} \quad \widehat{\mathbf{W}} = \begin{bmatrix} 1.772 & 9.774 \\ 9.774 & 143.020 \end{bmatrix}.$$

Thus, the corresponding correlation matrices are

$$\text{Corr\_}\widehat{\mathbf{U}}_0 = \begin{bmatrix} 1 & 0.177 \\ 0.177 & 1 \end{bmatrix}, \qquad \text{Corr\_}\widehat{\mathbf{U}}_1 = \begin{bmatrix} 0.612 & 0.136 \\ 0.136 & -0.058 \end{bmatrix} \quad \text{and} \quad \text{Corr\_}\widehat{\mathbf{W}} = \begin{bmatrix} 0.188 & 0.177 \\ 0.177 & 0.444 \end{bmatrix}.$$

Therefore, we see that the estimated correlation coefficient between IOP and CCT in any eye and at any time point is 0.177, whereas the estimated correlation coefficient between IOP and CCT for the two eyes is 0.136 and that for any two time points is 0.177. As expected, the correlation coefficient do not increase with both the eyes and over the time points.

From Table 1 we see that for all chosen pairs of training samples the MERs corresponding to our new classification rule with separable additive mean vectors are less than that of the MERs corresponding to the classification rule with separable multiplicative mean vectors and with unstructured mean vectors. The data do not seem to have separable multiplicative mean vector structure; see Leiva and Roy (2011). For training samples of sizes $(3, 3)$, $(5, 5)$, and $(8, 8)$ we see that MERs are more than that of Roy and Leiva's (2007) classification rule with CMVOS. However, for moderate training samples of sizes we see that our new classification rule with the separable additive mean vector has comparable MERs to classification rule with CMVOS. Further, LOOCV error rates are almost same for all the methods.

This may perhaps be due to the fact that the estimates of a total of 19 unknown parameters (10 for the mean vectors and 9 for the doubly exchangeable covariance structure) in the separable multiplicative and separable additive mean vector cases become unstable for small sample sizes, and thus increase the MERs. The parameter estimates in these two classification rules with separable mean vectors, do not have any closed form solutions for the parameters and have to compute all 19 parameters simultaneously using fixed point iteration algorithm from the training samples, hence need more samples to obtain the reliable or dependable estimate of the parameters. Thus, the performance of these classification rules for small sample sizes are much poorer than the moderate sample sizes. It is well known that the MER increases at the cost of estimating more parameters. On the other hand, the classification rule with unstructured mean vectors has a total of 33 unknown parameters (24 for the mean vectors and 9 for the doubly exchangeable covariance matrix), and Roy and Leiva's (2007) classification rule with CMVOS has 21 unknown parameters (12 for the mean vectors and 9 for the doubly exchangeable covariance matrix), have closed form solutions; thus the parameter estimates are more reliable, hence offer less MERs. Traditional classification rule fails for training samples of sizes $(3, 3)$, $(5, 5)$ owing to very small sample sizes, and other times its performance is much poorer than the other methods.

To demonstrate that the new classification rule with separable additive mean vectors would perform better if the data have the separable additive mean vector structures than the one that do not have, we perform the following simulation study. We also compare the results with the traditional linear classification rule.

6. A simulation study

To test the performance of our new classification rule, data sets are simulated separately with different training sample sizes from two populations. The values of v, number of repeated measurements over time, are chosen as 3 and 5, $u = 2$ and $m = 3$. We assume that the structure of the mean vectors as $\boldsymbol{\mu}_x^{(p)} = \boldsymbol{\tau}^{(p)} \oplus \boldsymbol{\lambda}^{(p)} \oplus \boldsymbol{\mu}^{(p)}$ for $p = 1, 2$, where $\boldsymbol{\tau}^{(1)} = (0, 0.9, 0.75, 0.70, 0.70)'$, $\boldsymbol{\tau}^{(2)} = (0, 0.6, 0.6, 0.4, 0.4)'$, $\boldsymbol{\lambda}^{(1)} = (0, 1.5)'$, $\boldsymbol{\lambda}^{(2)} = (0, 2.2)'$, $\boldsymbol{\mu}^{(1)} = (2, 1, 1)'$ and

Table 2
MERs (%) for the simulated data with separable additive mean model for different training sample sizes.

Assumed mean vector	$(\text{trn}^{(1)}, \text{trn}^{(2)})$															
	(3, 3)		(5, 5)		(6, 6)		(8, 8)		(10, 10)		(15,15)		(20,20)		(50,50)	
	v															
	3	5	3	5	3	5	3	5	3	5	3	5	3	5	3	5
Additive	16.90	7.73	11.68	13.45	14.30	8.73	13.30	7.53	12.35	6.38	10.90	7.45	9.60	6.25	9.28	4.55
Multiplicative	16.78	9.05	11.15	14.58	14.15	10.15	14.00	13.40	12.98	12.23	11.90	11.10	11.40	9.58	11.18	7.33
Unstructured	16.73	9.95	16.55	14.05	15.53	14.63	13.85	9.40	12.85	8.40	12.18	8.08	11.23	6.55	10.20	5.13
Traditional	Failed	Failed	Failed	Failed	Failed	Failed	Failed	Failed	29.80	Failed	15.93	Failed	21.15	14.58	10.58	11.65

$\boldsymbol{\mu}^{(2)} = (0, 1, 0)'$. We also assume that the two populations have the same doubly exchangeable covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ with $\mathbf{U}_0, \mathbf{U}_1$ and $\mathbf{W}$ as

$$\mathbf{U}_0 = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 4 & 3 \\ 2 & 3 & 5 \end{bmatrix}, \qquad \mathbf{U}_1 = \begin{bmatrix} 0.40 & 0.11 & 0.40 \\ 0.11 & 0.60 & 0.15 \\ 0.40 & 0.15 & 0.60 \end{bmatrix}, \quad \text{and} \quad \mathbf{W} = \begin{bmatrix} 0.3 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.3 \end{bmatrix}.$$

Training samples of sizes $(\text{trn}^{(1)}, \text{trn}^{(2)}) = (3, 3), (5, 5)$ (very small), (6, 6), (8, 8), (10, 10) (small), (15, 15), (20, 20) (moderate) and (50, 50) (large), and a pair of test samples (2000, 2000) are generated from the muv-variate normal populations $N_{muv}(\boldsymbol{\mu}_{\boldsymbol{x}}^{(p)}, \boldsymbol{\Gamma}_{\boldsymbol{x}}),\ \ p = 1, 2$, where $\boldsymbol{\mu}_{\boldsymbol{x}}^{(p)}$ and $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ are defined as above. Based on these samples, we estimate $(\boldsymbol{\mu}_{\boldsymbol{x}}^{(1)}, \boldsymbol{\mu}_{\boldsymbol{x}}^{(2)})$ and $(\mathbf{U}_0, \mathbf{U}_1, \mathbf{W})$ using the maximum likelihood method as discussed in Section 4.

To see the effectiveness of the separable additive mean vector structure in the classification rule we study at the same time the outcome of the classification rules when the actual mean vectors have separable additive structure, and we estimate and perform the classification analyses by assuming them as separable additive structure, separable multiplicative structure as well as unstructured, and calculate the MERs separately. Finally, we calculate the MERs for the traditional linear classification rule, where both the mean vectors and the pooled variance–covariance matrix are unstructured.

Our doubly exchangeable covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ has a blocked constant covariance structure over time, like compound symmetric covariance structure has constant variance over time for the univariate repeated measures data. The number of unknown parameters in $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is $3m\,(m + 1)\,/2$, a number independent of u and v. However, the number of unknown parameters in the separable additive as well as separable multiplicative mean structure in the classification rule is $2(v+u+m-2)$. Thus, in the separable structured mean cases the MER is influenced by two circumstances which are at odds with each other. First, the estimation of our doubly exchangeable covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ becomes better with the increase of v, as the number of unknown parameters is independent of the number of repeated measures v. The increase in v offers more information in each observation, and thus MER is decreased. Second, the estimation of the structured mean vectors becomes poorer since the number of unknown parameters in them increases with v, and accordingly MER is increased. However, the first one overshadows the second one after a 'burnout period'. We see from Table 2 that after the 'burnout period' of very small sample sizes, MER always decreases with the increase of v. This is due to the stable and thus reliable parameter estimates after the 'burnout period'. The number of unknown parameters are 30 and 34 for $v = 3$ and 5, respectively; and, the total number of samples are only 6 and 10 for the pairs of samples (3, 3) and (5, 5) respectively. Therefore, for $v = 5$ more parameters are in the model; as a result, the MER can be increased due to the cost of estimating more parameters when the number of samples is only 10. Separable mean vectors do not give reliable parameter estimates for very small samples, thus behave erroneously. Thus, we will only discuss the results after the 'burnout period'.

Table 2 shows the MERs for the simulated data. We see that the MERs are relatively higher when we overlook the structure of the mean vectors that is present in the data. We see that with the increase of training sample sizes MER decreases as we get much reliable estimates of the unknown parameters. MERs are smallest when we analyze the data assuming the separable additive mean vector structure. For small sample sizes the MERs for the separable additive structure, separable multiplicative structure, and unstructured are all comparable. We see that the assumption of the additive structured mean vector for large v offers maximum benefit when the same structure is present in the data.

When repeated measurements do not have any structure on the variance–covariance matrix, repeated measurements over time are sometimes disturbing in discriminating the groups. As mentioned before we also perform the simulation study for the traditional linear classification rule. We see from Table 2 that the MER increases with v for the pair of samples (50, 50), as the number of parameters are more for $v = 5$. Moreover, traditional linear classification rule fails most of the time due to the lack of proper sample size. Here also we see that our new classification rule outperforms the traditional one.

7. Possible future extensions of our proposed classification rule

Although multivariate normal distribution has been the fundamental focus in multivariate analysis, statisticians have been trying to extend the theory of multivariate analysis to more general case. Family of elliptical distributions with tractable radius or multivariate exponential family of distributions provides a useful generalization of the multivariate normal distribution for the modeling of repeated measure data. The covariance matrix retains its interpretation (at least up to a constant in the case of elliptical distributions) so that it can easily be structured for uniform and serial dependence and several levels of variance components (Lindsey, 1999).

Under mild assumptions, extension of our classification rule is possible to the family of elliptical distributions (Fang and Anderson, 1990) when they possess density functions, i.e., their contours are of the same shape as the multivariate normal density; namely elliptical. The main attraction of these densities are that they are more flexible than the normal density in that they cover both thick- and thin-tailed distributions (relative to the normal), and so to supply robust alternatives to many statistical procedures. Thus, the usual heavy tails of most members of the class of elliptical distributions make them natural candidates in modeling the data with outliers. An N-dimensional random vector $\boldsymbol{x}$ is distributed according to an elliptically contoured distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} > 0$ (positive definite), if it has the density function of the form

$$f(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} g\big((\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big),$$

where g is a density generator. In our scheme if the matrices $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$, and $\boldsymbol{\Delta}_3$ are positive definite (or equivalently, if $\boldsymbol{\Gamma_x}$ is a positive definite matrix) and the assumptions of Theorem 1 of Anderson et al. (1986) (see also Theorem 7.1.4, P. 236 of Gupta and Varga, 1993) are satisfied then it is possible to find the MLEs of our model parameters under the elliptical distribution. More precisely, we find the MLEs $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\tau}}, \widehat{\mathbf{U}}_0, \widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{W}}$ in Section 3, based on a sample of size n of muv-variate random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a population with distribution $N_{muv}\left(\boldsymbol{\mu_x}, \boldsymbol{\Gamma_x}\right)$. That is, we obtain the MLEs $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \widehat{\boldsymbol{\tau}} \oplus \widehat{\boldsymbol{\lambda}} \oplus \widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu_x}$ and $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}} = \mathbf{I}_{vu} \otimes \big(\widehat{\mathbf{U}}_0 - \widehat{\mathbf{U}}_1\big) + \mathbf{I}_v \otimes \mathbf{J}_u \otimes \big(\widehat{\mathbf{U}}_1 - \widehat{\mathbf{W}}\big) + \mathbf{J}_{vu} \otimes \widehat{\mathbf{W}}$ of $\boldsymbol{\Gamma_x}$, based on a sample of size 1 of the N-variate random vector $\boldsymbol{x} = \left(\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n'\right)'$ from a population with distribution $N_N\left(\mathbf{1}_n \otimes \boldsymbol{\mu_x}, \mathbf{I}_n \otimes \boldsymbol{\Gamma_x}\right)$, where $N = nmuv$. Now, under the assumptions that $g(\boldsymbol{x}'\boldsymbol{x})$ is a density in $\Re^N$ and $y^{N/2}g(y)$ has a finite positive maximum y_g, Theorem 1 of Anderson et al. (1986) states: on the basis of an observation $\boldsymbol{x}$ from $f(\boldsymbol{x})$ if the MLEs of the parameters $\boldsymbol{\nu} = 1_n \otimes \boldsymbol{\mu_x}$ and $\mathbf{V} = \mathbf{I}_n \otimes \boldsymbol{\Gamma_x}$ under normality exists and are unique and that MLE $\widehat{\mathbf{V}} = \mathbf{I}_n \otimes \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}$ of $\mathbf{I}_n \otimes \boldsymbol{\Gamma_x}$ is positive definite, then the MLEs of the parameters $\boldsymbol{\nu}$ and $\mathbf{V}$ for g also exist and they are given by $\widetilde{\boldsymbol{\nu}} = \widehat{\boldsymbol{\nu}} = \mathbf{1}_n \otimes \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\widetilde{\mathbf{V}} = \frac{N}{y_g}\widehat{\mathbf{V}} = \frac{N}{y_g}\mathbf{I}_n \otimes \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}$, and the maximum of the likelihood is $|\widetilde{\mathbf{V}}|^{-\frac{1}{2}} g\left(y_g\right)$. As a result, we can extend our classification rule to a wider class of distributions such as the multivariate power exponential and student t families, as these two are subfamilies of the elliptically contoured distributions and include the multivariate normal distribution as a special case. They can have the heavy tails required for handling extreme observations (Lindsey and Jones, 2000). MLEs can be derived for the unknown parameters $\beta(> 0)$, $\boldsymbol{\Sigma}$ (positive definite) and $\boldsymbol{\mu}$ of the multivariate power exponential distribution, as the distribution is continuous over the entire space of the kurtosis parameter β when $\beta > 0$. For $\beta < 1$, the distribution has heavier tails than the multivariate normal distribution and can be helpful in providing robustness against outliers. In the case of Multivariate student t distribution, MLEs can be derived for the unknown parameters $\beta(>2)$, $\boldsymbol{\Sigma}$ (positive definite) and $\boldsymbol{\mu}$, as in this case also the distribution is continuous over the entire space of the parameter β when $\beta > 2$.

Under mild assumptions, extension of our classification rule to the family of multivariate simple linear exponential family of distributions is also possible. The distribution of an N-dimensional random variable $\boldsymbol{x}$ belongs to the N-dimensional simple linear exponential family (Ziegler, 2011), if its density (meant to include probability mass functions for discrete data) is given by

$$f(\boldsymbol{x}) = \exp\left(\boldsymbol{\vartheta}'\boldsymbol{x} + b(\boldsymbol{x}, \boldsymbol{\Psi}) - d(\boldsymbol{\vartheta}, \boldsymbol{\Psi})\right),$$

where $\boldsymbol{\vartheta} \in \Re^N$ be the parameter vector of interest, $\boldsymbol{\Psi} \in \Re^{N\times N}$ be a positive definite matrix of fixed nuisance parameters, $b : \Re^N \times \Re^{N\times N} \to \Re$ and $d : \Re^N \times \Re^{N\times N} \to \Re$ some functions. The parameter vector $\boldsymbol{\vartheta}$ is termed as the natural parameter. It can be shown that the function $d(\boldsymbol{\vartheta}, \boldsymbol{\Psi})$ is the cumulant generating function of $f(\boldsymbol{x})$. As before, if the matrices $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$, and $\boldsymbol{\Delta}_3$ are positive definite (or equivalently, if $\boldsymbol{\Gamma_x}$ is a positive definite matrix) then by Theorem 1.2, P. 2 of Ziegler (2011) we have

$$\mathrm{E}(\boldsymbol{x}) = \frac{\partial d(\boldsymbol{\vartheta}, \boldsymbol{\Psi})}{\partial \boldsymbol{\vartheta}}$$

and

$$\mathrm{Var}(\boldsymbol{x}) = \frac{\partial^2 d(\boldsymbol{\vartheta}, \boldsymbol{\Psi})}{\partial \boldsymbol{\vartheta}\partial \boldsymbol{\vartheta}'}.$$

Using the invariance principle of the MLEs one can always calculate the ML estimates of $\frac{\partial d(\boldsymbol{\vartheta}, \boldsymbol{\Psi})}{\partial \boldsymbol{\vartheta}}$ and $\frac{\partial^2 d(\boldsymbol{\vartheta}, \boldsymbol{\Psi})}{\partial \boldsymbol{\vartheta}\partial \boldsymbol{\vartheta}'}$, and thus the MLEs $\widehat{\mathrm{E}(\boldsymbol{x})}$ and $\widehat{\mathrm{Var}(\boldsymbol{x})}$. As a result, we can extend our classification rule to a wider class of distributions such as the multinomial distribution, as it belongs to the multivariate simple linear exponential family of distributions and include the multivariate normal distribution as a special case.

8. Concluding remarks

Our study concludes that the appropriate or relevant structures on mean vectors and variance–covariance matrix are very important for the classification of three-level multivariate data. We see that the traditional linear classification rule is not at all useful when some structure is present in the data. The traditional linear classification rule not only gives much higher MERs, but in fact, also fails for small and very small sample sizes. Thus, the shift from traditional linear classification rule to our new classification rule is necessitated for some data set with particular structure on mean vectors and variance–covariance matrix.

Many future research directions can arise out of our proposed methodology in this paper. For instance, the present work can easily be extended to more than three levels. Also, one could generalize our classification rule to elliptical distributions with tractable radius, or multivariate exponential family of distributions as justified in Section 7. Generalization of our classification rule to elliptical distributions with tractable radius is under progress and will be presented in a future correspondence.

Acknowledgments

The authors would like to thank the associate editor for his thoughtful comments, and the two reviewers for their constructive comments of an earlier version of this article, and for providing valuable recommendations that led to the much improved version of the article. The second author thanks the support for the summer research grant from the College of Business at the University of Texas at San Antonio.

Appendix A. Maximum likelihood estimation of $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\boldsymbol{U_0}$, $\boldsymbol{U_1}$ and $\boldsymbol{W}$

Proof of Theorem 1. The likelihood function $L = L\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right) = L\left(\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{U}_0, \mathbf{U}_1, \mathbf{W}\right)$ can be written as

$$L\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right) = \frac{\exp -\frac{1}{2}\sum\limits_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)}{(2\pi)^{\frac{nmuv}{2}} \left|\boldsymbol{\Gamma}_{\boldsymbol{x}}\right|^{\frac{n}{2}}}.$$

Thus, the log likelihood function can be written as

$$\ln(L) = -\frac{nmuv}{2}\ln(2\pi) - \frac{n}{2}\ln\left|\boldsymbol{\Gamma}_{\boldsymbol{x}}\right| - \frac{1}{2}\sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right). \tag{A.1}$$

Let $\dot{\boldsymbol{x}}_r = \boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{x}_r - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu}$, and let $\overline{\boldsymbol{x}}_{ts}$ denotes the sample mean vector corresponding to time point t and site s, that is, $\overline{\boldsymbol{x}}_{ts} = \frac{1}{n}\sum_{r=1}^{n}\boldsymbol{x}_{r,ts}$. Let $\overline{\boldsymbol{x}} = \left(\overline{\boldsymbol{x}}'_{11}, \ldots, \overline{\boldsymbol{x}}'_{1u}, \ldots, \overline{\boldsymbol{x}}'_{v1}, \ldots, \overline{\boldsymbol{x}}'_{vu}\right)'$ be the global sample mean vector, that is, $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{r=1}^{n}\boldsymbol{x}_r$. Since $\dot{\boldsymbol{x}}_r = \boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}} = \left(\boldsymbol{x}_r - \overline{\boldsymbol{x}}\right) + \left(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}}\right)$, the sum of quadratic terms in the above log likelihood function can be expressed as

$$\sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right) = \operatorname{tr}\left[\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)'\right] = \operatorname{tr}\left[\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\mathbf{V}\right],$$

where

$$\mathbf{V} = \sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)' = \mathbf{W} + \mathbf{Z},$$

with

$$\mathbf{W} = \sum_{r=1}^{n}\left(\boldsymbol{x}_r - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_r - \overline{\boldsymbol{x}}\right)',$$

and

$$\begin{aligned}\mathbf{Z} &= n\left(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}}\right)\left(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}}\right)', \\ &= n\left(\overline{\boldsymbol{x}} - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu}\right)\left(\overline{\boldsymbol{x}} - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu}\right)'.\end{aligned}$$

Therefore, the log likelihood function (A.1) can be written as

$$\ln(L) = -\frac{nmuv}{2}\ln(2\pi) - \frac{n}{2}\ln\left|\boldsymbol{\Gamma}_{\boldsymbol{x}}\right| - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\mathbf{W}\right), -\frac{n}{2}\left(\overline{\boldsymbol{x}} - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu}\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\left(\overline{\boldsymbol{x}} - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu}\right).$$

We will first find the maximum likelihood estimates of the parameters $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ for a fixed covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$. For that we will find the partial derivatives of $\ln(L)$ with respect to $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ respectively. Now, the partial derivative $\frac{\partial}{\partial \boldsymbol{\tau}}\ln(L)$ is given by

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\tau}} \ln(L) &= -\frac{n}{2}\frac{\partial}{\partial \boldsymbol{\tau}}\left[(\bar{\boldsymbol{x}} - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu})' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu})\right], \\ &= n (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \oplus \boldsymbol{\mu}), \quad \text{using Property 2 of Appendix C}, \\ &= n (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\bar{\boldsymbol{x}} - \mathbf{1}_v \otimes \boldsymbol{\lambda} \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}) \\ &\quad - n (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\boldsymbol{\tau} \otimes \mathbf{1}_u \otimes \mathbf{1}_m), \quad \text{using Eq. (1)}, \\ &= n (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\bar{\boldsymbol{x}} - \mathbf{1}_v \otimes \boldsymbol{\lambda} \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}) - n (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m) \boldsymbol{\tau}.\end{aligned}$$

Equating the above derivative to zero we get

$$\boldsymbol{\tau} = \mathbf{D}_{\boldsymbol{\lambda}\boldsymbol{\mu}} (\bar{\boldsymbol{x}} - \mathbf{1}_v \otimes \boldsymbol{\lambda} \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}),$$

where

$$\mathbf{D}_{\boldsymbol{\lambda}\boldsymbol{\mu}} = \left[(\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)\right]^{-1} \cdot (\mathbf{I}_v \otimes \mathbf{1}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}. \tag{A.2}$$

Using the same technique we get

$$\boldsymbol{\lambda} = \mathbf{D}_{\boldsymbol{\tau}\boldsymbol{\mu}} (\bar{\boldsymbol{x}} - \boldsymbol{\tau} \otimes \mathbf{1}_v \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}),$$

where

$$\mathbf{D}_{\boldsymbol{\tau}\boldsymbol{\mu}} = \left[(\mathbf{1}_v \otimes \mathbf{I}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\mathbf{1}_v \otimes \mathbf{I}_u \otimes \mathbf{1}_m)\right]^{-1} \cdot (\mathbf{1}_v \otimes \mathbf{I}_u \otimes \mathbf{1}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}, \tag{A.3}$$

and

$$\boldsymbol{\mu} = \mathbf{D}_{\boldsymbol{\tau}\boldsymbol{\lambda}} (\bar{\boldsymbol{x}} - \boldsymbol{\tau} \otimes \mathbf{1}_v \otimes \mathbf{1}_m - \mathbf{1}_v \otimes \boldsymbol{\lambda} \otimes \mathbf{1}_m),$$

where

$$\mathbf{D}_{\boldsymbol{\tau}\boldsymbol{\lambda}} = \left[(\mathbf{1}_v \otimes \mathbf{1}_u \otimes \mathbf{I}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} (\mathbf{1}_v \otimes \mathbf{1}_u \otimes \mathbf{I}_m)\right]^{-1} \cdot (\mathbf{1}_v \otimes \mathbf{1}_u \otimes \mathbf{I}_m)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}. \tag{A.4}$$

We will now maximize the log likelihood function (A.1) with respect to $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$ for fixed $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ to get the MLEs of $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$. Since $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}$ and $|\boldsymbol{\Gamma}_{\boldsymbol{x}}|$ can be expressed as a function of $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$, maximizing with respect to $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$ is equivalent to maximizing with respect to $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$. Now substituting the values of $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}$ from (3) in the log likelihood function (A.1) and simplifying we get

$$\begin{aligned}\ln(L) &= -\frac{nmuv}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Gamma}_{\boldsymbol{x}}| - \frac{1}{2}\sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)'(\mathbf{I}_{vu} \otimes \mathbf{A})\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right) \\ &\quad - \frac{1}{2}\sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)'(\mathbf{I}_v \otimes \mathbf{J}_u \otimes \mathbf{B})\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right) - \frac{1}{2}\sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)'(\mathbf{J}_{vu} \otimes \mathbf{C})\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right).\end{aligned}$$

Substituting the values of $\mathrm{E}\left[\boldsymbol{x}_{r,ts}\right] = (\tau_t + \lambda_s)\,\mathbf{1}_m + \boldsymbol{\mu}$, and $|\boldsymbol{\Gamma}_{\boldsymbol{x}}|$ from (5) in the above log likelihood equation we get

$$\ln(L) = -\frac{nmuv}{2}\ln(2\pi) - \frac{nv(u-1)}{2}\ln|\boldsymbol{\Delta}_1| - \frac{n(v-1)}{2}\ln|\boldsymbol{\Delta}_2| - \frac{n}{2}\ln|\boldsymbol{\Delta}_3| - \frac{1}{2}\mathrm{tr}\mathbf{A}\mathbf{H}_1 - \frac{1}{2}\mathrm{tr}\mathbf{B}\mathbf{H}_2 - \frac{1}{2}\mathrm{tr}\mathbf{C}\mathbf{H}_3, \tag{A.5}$$

where

$$\mathbf{H}_1 = \sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)', \tag{A.6}$$

$$\mathbf{H}_2 = \sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,ts^*} - (\tau_t + \lambda_{s^*})\,\mathbf{1}_m - \boldsymbol{\mu}\right)', \tag{A.7}$$

and

$$\mathbf{H}_3 = \sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{t^*=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\,\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,t^*s^*} - (\tau_{t^*} + \lambda_{s^*})\,\mathbf{1}_m - \boldsymbol{\mu}\right)'. \tag{A.8}$$

Now using the identity

$$\mathbf{A}\mathbf{H}_1 + \mathbf{B}\mathbf{H}_2 + \mathbf{C}\mathbf{H}_3 = \boldsymbol{\Delta}_1^{-1}\left(\mathbf{H}_1 - \frac{1}{u}\mathbf{H}_2\right) + \boldsymbol{\Delta}_2^{-1}\left(\frac{1}{u}\mathbf{H}_2 - \frac{1}{vu}\mathbf{H}_3\right) + \boldsymbol{\Delta}_3^{-1}\left(\frac{1}{vu}\mathbf{H}_3\right),$$

the log likelihood (A.5) can be written as

$$\begin{aligned}\ln(L) &= -\frac{nmuv}{2}\ln(2\pi) - \frac{nv(u-1)}{2}\ln|\boldsymbol{\Delta}_1| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Delta}_1^{-1}\left(\mathbf{H}_1 - \frac{1}{u}\mathbf{H}_2\right)\\ &\quad - \frac{n(v-1)}{2}\ln|\boldsymbol{\Delta}_2| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Delta}_2^{-1}\left(\frac{1}{u}\mathbf{H}_2 - \frac{1}{vu}\mathbf{H}_3\right) - \frac{n}{2}\ln|\boldsymbol{\Delta}_3| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Delta}_3^{-1}\left(\frac{1}{vu}\mathbf{H}_3\right),\\ &= -\frac{nmuv}{2}\ln(2\pi) + nv(u-1)\left\{-\frac{1}{2}\ln|\boldsymbol{\Delta}_1| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Delta}_1^{-1}\left[\frac{1}{nv(u-1)}\left(\mathbf{H}_1 - \frac{1}{u}\mathbf{H}_2\right)\right]\right\}\\ &\quad + n(v-1)\left\{-\frac{1}{2}\ln|\boldsymbol{\Delta}_2| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Delta}_2^{-1}\left[\frac{1}{n(v-1)}\left(\frac{1}{u}\mathbf{H}_2 - \frac{1}{vu}\mathbf{H}_3\right)\right]\right\}\\ &\quad + n\left\{-\frac{1}{2}\ln|\boldsymbol{\Delta}_3| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Delta}_3^{-1}\left(\frac{1}{nuv}\mathbf{H}_3\right)\right\}.\end{aligned}$$

Therefore, using Lemma 3.2.2 of Anderson (2003), the MLEs $\widehat{\boldsymbol{\Delta}}_1$, $\widehat{\boldsymbol{\Delta}}_2$ and $\widehat{\boldsymbol{\Delta}}_3$ for each $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are given by

$$\widehat{\boldsymbol{\Delta}}_1 = \frac{1}{nv(u-1)}\left(\mathbf{H}_1 - \frac{1}{u}\mathbf{H}_2\right),$$
$$\widehat{\boldsymbol{\Delta}}_2 = \frac{1}{n(v-1)}\left(\frac{1}{u}\mathbf{H}_2 - \frac{1}{vu}\mathbf{H}_3\right),$$

and

$$\widehat{\boldsymbol{\Delta}}_3 = \frac{1}{nuv}\mathbf{H}_3.$$

Thus, from Eq. (4) we can write MLEs of $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{C}}$ as follows:

$$\widehat{\mathbf{A}} = \widehat{\boldsymbol{\Delta}}_1^{-1}, \qquad \widehat{\mathbf{B}} = \frac{1}{u}\left(\widehat{\boldsymbol{\Delta}}_2^{-1} - \widehat{\boldsymbol{\Delta}}_1^{-1}\right), \quad \text{and} \quad \widehat{\mathbf{C}} = \frac{1}{vu}\left(\widehat{\boldsymbol{\Delta}}_3^{-1} - \widehat{\boldsymbol{\Delta}}_2^{-1}\right). \tag{A.9}$$

Using the fact that ML estimations of transformed parameters are the transformed parameter estimates of the MLEs, from (3) we have the estimate of $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}$ as

$$\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1} = \mathbf{I}_{vu}\otimes\widehat{\mathbf{A}} + \mathbf{I}_v\otimes\mathbf{J}_u\otimes\widehat{\mathbf{B}} + \mathbf{J}_{vu}\otimes\widehat{\mathbf{C}}. \tag{A.10}$$

Now the MLEs of $\mathbf{U}_0$, $\mathbf{U}_1$ and $\mathbf{W}$ can be obtained from $\widehat{\boldsymbol{\Delta}}_1$, $\widehat{\boldsymbol{\Delta}}_2$ and $\widehat{\boldsymbol{\Delta}}_3$ by

$$\begin{aligned}\widehat{\mathbf{W}} &= \frac{1}{vu}\left(\widehat{\boldsymbol{\Delta}}_3 - \widehat{\boldsymbol{\Delta}}_2\right) = \frac{1}{vu}\left[\frac{1}{nvu}\mathbf{H}_3 - \frac{1}{n(v-1)}\left(\frac{1}{u}\mathbf{H}_2 - \frac{1}{vu}\mathbf{H}_3\right)\right],\\ &= \frac{\mathbf{H}_3 - \mathbf{H}_2}{nu^2v(v-1)},\\ &= \frac{1}{nu^2v(v-1)}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{t\neq t^*=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,t^*s^*} - (\tau_{t^*} + \lambda_{s^*})\mathbf{1}_m - \boldsymbol{\mu}\right)'\\ \widehat{\mathbf{U}}_1 &= \frac{1}{u}\left\{\widehat{\boldsymbol{\Delta}}_2 - \widehat{\boldsymbol{\Delta}}_1 + u\widehat{\mathbf{W}}\right\},\\ &= \frac{1}{u}\left\{\frac{1}{n(v-1)}\left(\frac{1}{u}\mathbf{H}_2 - \frac{1}{vu}\mathbf{H}_3\right) - \frac{1}{nv(u-1)}\left(\mathbf{H}_1 - \frac{1}{u}\mathbf{H}_2\right) + \frac{\mathbf{H}_3 - \mathbf{H}_2}{nuv(v-1)}\right\},\\ &= \frac{\mathbf{H}_2 - \mathbf{H}_1}{nu(u-1)v},\\ &= \frac{1}{nvu(u-1)}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s\neq s^*=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,ts^*} - (\tau_t + \lambda_{s^*})\mathbf{1}_m - \boldsymbol{\mu}\right)',\end{aligned}$$

and

$$\begin{aligned}\widehat{\mathbf{U}}_0 &= \widehat{\boldsymbol{\Delta}}_1 + \widehat{\mathbf{U}}_1 = \frac{1}{nv(u-1)}\left(\mathbf{H}_1 - \frac{1}{u}\mathbf{H}_2\right) + \frac{\mathbf{H}_2 - \mathbf{H}_1}{nu(u-1)v},\\ &= \frac{\mathbf{H}_1}{nuv},\\ &= \frac{1}{nuv}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\mathbf{1}_m - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_{r,ts} - (\tau_t + \lambda_s)\mathbf{1}_m - \boldsymbol{\mu}\right)'. \quad \square\end{aligned}$$

Appendix B. Derivation of different classification rules

$$\begin{aligned}
\text{Let } \boldsymbol{\alpha}' &= \left(\boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_u \otimes \mathbf{1}_m\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \\
&= \left(\boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_u \otimes \mathbf{1}_m\right)' \left(\mathbf{I}_v \otimes \mathbf{I}_u \otimes \boldsymbol{\Delta}_1^{-1}\right) + \left(\boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_u \otimes \mathbf{1}_m\right)' \left(\mathbf{I}_v \otimes \mathbf{J}_u \otimes \frac{1}{u}\left(\boldsymbol{\Delta}_2^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right) \\
&\quad + \left(\boldsymbol{\tau}^{(p)} \otimes \mathbf{1}_u \otimes \mathbf{1}_m\right)' \left(\mathbf{J}_v \otimes \mathbf{J}_u \otimes \frac{1}{vu}\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right)\right) \\
&= \boldsymbol{\tau}^{(p)\prime} \otimes \mathbf{1}_u' \otimes \mathbf{1}_m' \boldsymbol{\Delta}_1^{-1} + \boldsymbol{\tau}^{(p)\prime} \otimes u\mathbf{1}_u' \otimes \frac{1}{u}\mathbf{1}_m'\left(\boldsymbol{\Delta}_2^{-1} - \boldsymbol{\Delta}_1^{-1}\right) + \boldsymbol{\tau}^{(p)\prime}\mathbf{J}_v \otimes u\mathbf{1}_u' \otimes \frac{1}{vu}\mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right) \\
&= \boldsymbol{\tau}^{(p)\prime} \otimes \mathbf{1}_u' \otimes \mathbf{1}_m' \boldsymbol{\Delta}_2^{-1} + \overline{\tau}^{(p)}\mathbf{1}_v' \otimes \mathbf{1}_u' \otimes \mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right),
\end{aligned}$$

where

$$\overline{\tau}^{(p)} = \frac{1}{v}\sum_{t=1}^{v} \tau_t^{(p)}$$

$$\begin{aligned}
\text{Similarly, let } \boldsymbol{\beta}' &= \left(\mathbf{1}_v \otimes \boldsymbol{\lambda}^{(p)} \otimes \mathbf{1}_m\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \\
&= \mathbf{1}_v' \otimes \boldsymbol{\lambda}^{(p)\prime} \otimes \mathbf{1}_m'\boldsymbol{\Delta}_1^{-1} + \mathbf{1}_v' \otimes \boldsymbol{\lambda}^{(p)\prime}\mathbf{J}_u \otimes \frac{1}{u}\mathbf{1}_m'\left(\boldsymbol{\Delta}_2^{-1} - \boldsymbol{\Delta}_1^{-1}\right) + v\mathbf{1}_v' \otimes \boldsymbol{\lambda}^{(p)\prime}\mathbf{J}_u \otimes \frac{1}{vu}\mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right) \\
&= \mathbf{1}_v' \otimes \boldsymbol{\lambda}^{(p)\prime} \otimes \mathbf{1}_m'\boldsymbol{\Delta}_1^{-1} + \mathbf{1}_v' \otimes u\overline{\lambda}^{(p)}\mathbf{1}_u' \otimes \frac{1}{u}\mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1} - \left(\boldsymbol{\Delta}_2^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right) \\
&= \mathbf{1}_v' \otimes \boldsymbol{\lambda}^{(p)\prime} \otimes \mathbf{1}_m'\boldsymbol{\Delta}_1^{-1} + \mathbf{1}_v' \otimes \overline{\lambda}^{(p)}\mathbf{1}_u' \otimes \mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_1^{-1}\right),
\end{aligned}$$

where

$$\overline{\lambda}^{(p)} = \frac{1}{u}\sum_{s=1}^{u} \lambda_s^{(p)},$$

and

$$\begin{aligned}
\boldsymbol{\eta}' &= \left(\mathbf{1}_v \otimes \mathbf{1}_u \otimes \boldsymbol{\mu}^{(p)}\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \\
&= \mathbf{1}_v' \otimes \mathbf{1}_u' \otimes \boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_1^{-1} + \mathbf{1}_v' \otimes u\mathbf{1}_u' \otimes \frac{1}{u}\boldsymbol{\mu}^{(p)\prime}\left(\boldsymbol{\Delta}_2^{-1} - \boldsymbol{\Delta}_1^{-1}\right) + v\mathbf{1}_v' \otimes u\mathbf{1}_u' \otimes \frac{1}{vu}\boldsymbol{\mu}^{(p)\prime}\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right) \\
&= \mathbf{1}_v' \otimes \mathbf{1}_u' \otimes \boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}.
\end{aligned}$$

Therefore, the first summand of $l^{(p)}(\boldsymbol{x}_o)$ in (15) is

$$\begin{aligned}
m^{(p)}(\boldsymbol{x}_o) &= \boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_o \\
&= (\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\eta})'\boldsymbol{x}_o \\
&= \left(\boldsymbol{\tau}^{(p)\prime} \otimes \mathbf{1}_u' \otimes \mathbf{1}_m'\boldsymbol{\Delta}_2^{-1}\right)\boldsymbol{x}_o + \left(\overline{\tau}^{(p)}\mathbf{1}_v' \otimes \mathbf{1}_u' \otimes \mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right)\right)\boldsymbol{x}_o \\
&\quad + \left(\mathbf{1}_v' \otimes \boldsymbol{\lambda}^{(p)\prime} \otimes \mathbf{1}_m'\boldsymbol{\Delta}_1^{-1}\right)\boldsymbol{x}_o + \left(\mathbf{1}_v' \otimes \overline{\lambda}^{(p)}\mathbf{1}_u' \otimes \mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right)\boldsymbol{x}_o + \left(\mathbf{1}_v' \otimes \mathbf{1}_u' \otimes \boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\right)\boldsymbol{x}_o \\
&= \sum_{t=1}^{v}\left[\tau_t^{(p)}\left(\mathbf{1}_m'\boldsymbol{\Delta}_1^{-1}\right)\sum_{s=1}^{u}\boldsymbol{x}_{o,ts}\right] + \left(\frac{1}{v}\sum_{t=1}^{v}\tau_t^{(p)}\right)\left(\mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right)\right)\sum_{t=1}^{v}\sum_{s=1}^{u}\boldsymbol{x}_{o,ts} + \left(\mathbf{1}_m'\boldsymbol{\Delta}_1^{-1}\right)\left(\sum_{s=1}^{u}\lambda_s^{(p)}\right) \\
&\quad \times \sum_{t=1}^{v}\boldsymbol{x}_{o,ts} + \left(\frac{1}{u}\sum_{s=1}^{u}\lambda_s^{(p)}\right)\left(\mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right)\sum_{t=1}^{v}\sum_{s=1}^{u}\boldsymbol{x}_{o,ts} + \boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\sum_{t=1}^{v}\sum_{s=1}^{u}\boldsymbol{x}_{o,ts} \\
&= \left(\mathbf{1}_m'\boldsymbol{\Delta}_2^{-1}\right)\sum_{t=1}^{v}u\tau_t^{(p)}\overline{\boldsymbol{x}}_{o,t\bullet} + vu\overline{\tau}^{(p)}\left(\mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_2^{-1}\right)\right)\overline{\boldsymbol{x}}_o \\
&\quad + \left(\mathbf{1}_m'\boldsymbol{\Delta}_1^{-1}\right)\sum_{s=1}^{u}v\lambda_s^{(p)}\overline{\boldsymbol{x}}_{o,\bullet s} + vu\overline{\lambda}^{(p)}\left(\mathbf{1}_m'\left(\boldsymbol{\Delta}_3^{-1} - \boldsymbol{\Delta}_1^{-1}\right)\right)\overline{\boldsymbol{x}}_o + vu\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\overline{\boldsymbol{x}}_o,
\end{aligned}$$

where

$$\overline{\boldsymbol{x}}_{o,t\bullet} = \frac{1}{u}\sum_{s=1}^{u}\boldsymbol{x}_{o,ts},$$

$$\overline{\boldsymbol{x}}_{o,\bullet s} = \frac{1}{v}\sum_{t=1}^{v}\boldsymbol{x}_{o,ts},$$

and

$$\bar{\boldsymbol{x}}_o = \frac{1}{vu}\sum_{t=1}^{v}\sum_{s=1}^{u}\boldsymbol{x}_{o,ts}.$$

The quadratic form in the second summand of $l^{(p)}(\boldsymbol{x}_o)$ in (15) is

$$\begin{aligned}\kappa^{(p)} &= \boldsymbol{\mu}'_{\boldsymbol{x}^{(p)}}\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}} \\ &= (\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\eta})'\left(\boldsymbol{\tau}^{(p)}\oplus\boldsymbol{\lambda}^{(p)}\oplus\boldsymbol{\mu}^{(p)}\right) \\ &= \left[\left(\left(\boldsymbol{\tau}^{(p)}-\overline{\tau}^{(p)}\mathbf{1}_v\right)'\otimes\mathbf{1}'_u\otimes\mathbf{1}'_m\boldsymbol{\Delta}_2^{-1}\right)+\left(\overline{\tau}^{(p)}\mathbf{1}'_v\otimes\mathbf{1}'_u\otimes\mathbf{1}'_m\boldsymbol{\Delta}_3^{-1}\right)\right]\left(\boldsymbol{\tau}^{(p)}\oplus\boldsymbol{\lambda}^{(p)}\oplus\boldsymbol{\mu}^{(p)}\right) \\ &\quad+\left[\left(\mathbf{1}'_v\otimes\left(\boldsymbol{\lambda}^{(p)}-\overline{\lambda}^{(p)}\mathbf{1}_u\right)'\otimes\mathbf{1}'_m\boldsymbol{\Delta}_1^{-1}\right)+\left(\mathbf{1}'_v\otimes\overline{\lambda}^{(p)}\mathbf{1}'_u\otimes\mathbf{1}'_m\boldsymbol{\Delta}_3^{-1}\right)\right]\left(\boldsymbol{\tau}^{(p)}\oplus\boldsymbol{\lambda}^{(p)}\oplus\boldsymbol{\mu}^{(p)}\right) \\ &\quad+\left(\mathbf{1}'_v\otimes\mathbf{1}'_u\otimes\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\right)\left(\boldsymbol{\tau}^{(p)}\oplus\boldsymbol{\lambda}^{(p)}\oplus\boldsymbol{\mu}^{(p)}\right).\end{aligned}$$

Replacing $\boldsymbol{\tau}^{(p)}\oplus\boldsymbol{\lambda}^{(p)}\oplus\boldsymbol{\mu}^{(p)}$ and noting that $\left(\boldsymbol{\tau}^{(p)}-\overline{\tau}^{(p)}\mathbf{1}_v\right)'\mathbf{1}_v=0$ and $(\boldsymbol{\lambda}^{(p)}-\overline{\lambda}^{(p)}\mathbf{1}_u)'\mathbf{1}_u=0$, it follows that

$$\begin{aligned}\kappa^{(p)} &= \boldsymbol{\mu}'_{\boldsymbol{x}^{(p)}}\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}} \\ &= \left(\boldsymbol{\tau}^{(p)}-\overline{\tau}^{(p)}\mathbf{1}_v\right)'\boldsymbol{\tau}^{(p)}\otimes\mathbf{1}'_u\mathbf{1}_u\otimes\mathbf{1}'_m\boldsymbol{\Delta}_2^{-1}\mathbf{1}_m+\mathbf{1}'_v\mathbf{1}_v\otimes\left(\boldsymbol{\lambda}^{(p)}-\overline{\lambda}^{(p)}\mathbf{1}_u\right)'\boldsymbol{\lambda}^{(p)}\otimes\mathbf{1}'_m\boldsymbol{\Delta}_1^{-1}\mathbf{1}_m \\ &\quad+\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\left(\mathbf{1}'_v\otimes\mathbf{1}'_u\otimes\mathbf{1}'_m\boldsymbol{\Delta}_3^{-1}\right)\left(\boldsymbol{\tau}^{(p)}\oplus\boldsymbol{\lambda}^{(p)}\oplus\boldsymbol{\mu}^{(p)}\right)+\left(\mathbf{1}'_v\otimes\mathbf{1}'_u\otimes\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\right)\left(\boldsymbol{\tau}^{(p)}\oplus\boldsymbol{\lambda}^{(p)}\oplus\boldsymbol{\mu}^{(p)}\right).\end{aligned}$$

Denoting by d_{j+} the sum of all the elements of $\boldsymbol{\Delta}_j^{-1}$, for $j=1,2,3$, we have

$$\begin{aligned}\kappa^{(p)} &= \boldsymbol{\mu}'_{\boldsymbol{x}^{(p)}}\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}} \\ &= ud_{2+}\left(\boldsymbol{\tau}^{(p)}-\overline{\tau}^{(p)}\mathbf{1}_v\right)'\boldsymbol{\tau}^{(p)}+vd_{1+}\left(\boldsymbol{\lambda}^{(p)}-\overline{\lambda}^{(p)}\mathbf{1}_u\right)'\boldsymbol{\lambda}^{(p)}+ud_{3+}\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\mathbf{1}'_v\boldsymbol{\tau}^{(p)} \\ &\quad+vd_{3+}\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\mathbf{1}'_u\boldsymbol{\lambda}^{(p)}+vu\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\mathbf{1}'_m\boldsymbol{\Delta}_3^{-1}\boldsymbol{\mu}^{(p)}+\mathbf{1}'_v\boldsymbol{\tau}^{(p)}u\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\mathbf{1}_m \\ &\quad+v\mathbf{1}'_u\boldsymbol{\lambda}^{(p)}\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\mathbf{1}_m+vu\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\boldsymbol{\mu}^{(p)} \\ &= ud_{2+}\left(\boldsymbol{\tau}^{(p)}-\overline{\tau}^{(p)}\mathbf{1}_v\right)'\boldsymbol{\tau}^{(p)}+vd_{1+}\left(\boldsymbol{\lambda}^{(p)}-\overline{\lambda}^{(p)}\mathbf{1}_u\right)'\boldsymbol{\lambda}^{(p)}+vud_{3+}\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\overline{\tau}^{(p)} \\ &\quad+vud_{3+}\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\overline{\lambda}^{(p)}+vu\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\mathbf{1}'_m\boldsymbol{\Delta}_3^{-1}\boldsymbol{\mu}^{(p)}+vu\overline{\tau}^{(p)}\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\mathbf{1}_m \\ &\quad+vu\overline{\lambda}^{(p)}\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\mathbf{1}_m+vu\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\boldsymbol{\mu}^{(p)}.\end{aligned}$$

Now, by denoting $\overline{\tau^2}^{(p)}=\frac{1}{v}\sum_{t=1}^{v}\left(\tau_t^{(p)}\right)^2$ and $\overline{\lambda^2}^{(p)}=\frac{1}{u}\sum_{s=1}^{u}\left(\lambda_s^{(p)}\right)^2$ we have

$$\begin{aligned}\kappa^{(p)} &= \boldsymbol{\mu}'_{\boldsymbol{x}^{(p)}}\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}} \\ &= vud_{2+}\left(\overline{\tau^2}^{(p)}-\left(\overline{\tau}^{(p)}\right)^2\right)+vud_{1+}\left(\overline{\lambda^2}^{(p)}-\left(\overline{\lambda}^{(p)}\right)^2\right) \\ &\quad+vud_{3+}\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)^2+2vu\left(\overline{\tau}^{(p)}+\overline{\lambda}^{(p)}\right)\mathbf{1}'_m\boldsymbol{\Delta}_3^{-1}\boldsymbol{\mu}^{(p)}+vu\boldsymbol{\mu}^{(p)\prime}\boldsymbol{\Delta}_3^{-1}\boldsymbol{\mu}^{(p)}.\end{aligned}$$

Appendix C. Kronecker sum derivatives

Let $\underset{u_h\times 1}{\boldsymbol{x}_h}=\left(x_{h1},\ldots,x_{hu_h}\right)'$ be an $(u_h\times 1)$-dimensional vector of real variables for $h=1,\ldots,n$. Let $\bigoplus_{h=1}^{n}\boldsymbol{x}_h=\boldsymbol{x}_1\oplus\boldsymbol{x}_2\oplus\cdots\oplus\boldsymbol{x}_n$ represents the Kronecker sum of them, and $\bigotimes_{h=1}^{n}\boldsymbol{x}_h=\boldsymbol{x}_1\otimes\boldsymbol{x}_2\otimes\cdots\otimes\boldsymbol{x}_n$ represents the Kronecker product of them. Then $\boldsymbol{v}=\bigoplus_{h=1}^{n}\boldsymbol{x}_h$ be a $\left(\sum_{h=1}^{n}u_h\right)\times 1$ dimensional vector, and

$$\begin{aligned}\underset{\left(\sum_{h=1}^{n}u_h\right)\times 1}{\boldsymbol{v}} &= \bigoplus_{k=1}^{n}\boldsymbol{x}_k=\boldsymbol{x}_1\oplus\boldsymbol{x}_2\oplus\cdots\oplus\boldsymbol{x}_n \\ &= \sum_{k=1}^{n}\left(\bigotimes_{h=1}^{k-1}\mathbf{1}_{u_h}\otimes\boldsymbol{x}_k\otimes\bigotimes_{h=k+1}^{n}\mathbf{1}_{u_h}\right) \\ &= \sum_{k=1}^{n}\left(\bigotimes_{h=1}^{k-1}\mathbf{1}_{u_h}\otimes\mathbf{I}_{u_k}\otimes\bigotimes_{h=k+1}^{n}\mathbf{1}_{u_h}\right)\boldsymbol{x}_k.\end{aligned}$$

Now, the following properties can be proved easily:

1. The quantity $\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}_j}$ can be calculated as follows:

$$\frac{\partial\left(\bigoplus_{k=1}^{n} \boldsymbol{x}_k\right)}{\partial \boldsymbol{x}_j} = \left(\bigotimes_{k=1}^{j-1} \mathbf{1}'_{u_k}\right) \otimes \mathbf{I}_{u_j} \otimes \left(\bigotimes_{h=j+1}^{n} \mathbf{1}'_{u_k}\right) \quad \text{for } j = 1, \ldots, n,$$

   where it is assumed that

$$\bigotimes_{h=k}^{i} \boldsymbol{x}_h = 1 \quad \text{if } k > i.$$

2. Let $\boldsymbol{y} = \boldsymbol{a} - \boldsymbol{v} = \boldsymbol{a} - \bigoplus_{k=1}^{n} \boldsymbol{x}_k$, where $\boldsymbol{a}$ is a constant vector and let $\mathbf{D}$ be a $\left(\sum_{h=1}^{n} u_h \times \sum_{h=1}^{n} u_h\right)$-dimensional symmetric matrix, if $\boldsymbol{q} = \boldsymbol{y}' \cdot \mathbf{D} \cdot \boldsymbol{y} = (\boldsymbol{a} - \boldsymbol{v})' \cdot \mathbf{D} \cdot (\boldsymbol{a} - \boldsymbol{v})$ then

$$\begin{aligned}
\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{x}_j} &= \frac{\partial\left(\boldsymbol{y}' \cdot \mathbf{D} \cdot \boldsymbol{y}\right)}{\partial \boldsymbol{x}_j} \\
&= \frac{\partial\left[\left(\boldsymbol{a} - \bigoplus\limits_{k=1}^{n} \boldsymbol{x}_k\right)' \cdot \mathbf{D} \cdot \left(\boldsymbol{a} - \bigoplus_{k=1}^{n} \boldsymbol{x}_k\right)\right]}{\partial \boldsymbol{x}_j} \\
&= \left(\frac{\partial\left(\boldsymbol{a} - \bigoplus\limits_{k=1}^{n} \boldsymbol{x}_k\right)}{\partial \boldsymbol{x}_j}\right) \cdot \left(\frac{\partial\left[\boldsymbol{y}' \cdot \mathbf{D} \cdot \boldsymbol{y}\right]}{\partial \boldsymbol{y}}\right) \\
&= -2\left[\left(\bigotimes_{k=1}^{j-1} \mathbf{1}'_{u_k}\right) \otimes \mathbf{I}_{u_j} \otimes \left(\bigotimes_{h=j+1}^{n} \mathbf{1}'_{u_k}\right)\right] \cdot \mathbf{D} \cdot \left(\boldsymbol{a} - \bigoplus_{k=1}^{n} \boldsymbol{x}_k\right).
\end{aligned}$$

References

Akdemir, D., Gupta, A.K., 2011. Array variate random variables with multiway Kronecker delta covariance matrix structure. Tech. Report. Department of Mathematics and Statistics at the Bowling Green State University.

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. John Wiley & Sons, Inc., New Jersey.

Anderson, T.W., Fang, K., Hsu, H., 1986. Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions. Canadian Journal of Statistics 14 (1), 55–59.

Chaudhuri, G., Borwankar, J.D., Rao, P.R.K., 1991. Bhattacharyya distance based linear discriminant function for stationary time series. Communications in Statistics—Theory and Methods 20, 2195–2205.

Fang, K., Anderson, T.W., 1990. Statistical Inference in Elliptically Contoured and Related Distributions. Allerton Press Inc., New York.

Gupta, A.K., Varga, T., 1993. Elliptically Contoured Models in Statistics. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis, sixth ed. Pearson Prentice Hall, New Jersey.

Kroonenberg, P.M., 2008. Applied Multiway Data Analysis. John Wiley & Sons, Inc., New Jersey.

Lange, K., Wu, T.T., 2008. An MM algorithm for multicategory vertex discriminant analysis. Journal of Computational and Graphical Statistics 17, 527–544.

Leiva, R., 2007. Linear discrimination with equicorrelated training vectors. Journal of Multivariate Analysis 98 (2), 384–409.

Leiva, R.A., Herrera, M., 1999. Generalización de la distancia de Mahalanobis para el análisis discriminante lineal en poblaciones con matrices de covarianza desiguales. Revista de la Sociedad Argentina de Estadística 3, 64–85.

Leiva, R., Roy, A., 2009a. Classification rules for triply multivariate data with an AR(1) correlation structure on the repeated measures over time. Journal of Statistical Planning and Inference 139 (8), 2598–2613.

Leiva, R., Roy, A., 2009b. Linear discrimination for three-level multivariate data with structured mean vectors and doubly exchangeable covariance structure, The Paper Presented at the MAT-TRIAD 2009, Bedlewo, Poland, March 23–27.

Leiva, R., Roy, A., 2011. Linear discrimination for multi-level multivariate data with separable means and jointly equicorrelated covariance structure. Journal of Statistical Planning and Inference 141 (5), 1910–1924.

Lindsey, J.K., 1999. Multivariate elliptically contoured distributions for repeated measurements. Biometrics 55, 1277–1280.

Lindsey, J.K., Jones, B., 2000. Modeling pharmacokinetic data using heavy-tailed multivariate distributions. Journal of Biopharmaceutical Statistics 10 (3), 369–381.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.

Paranjpe, S.A., Gore, A.P., 1994. Selecting variables for discrimination when covariance matrices are unequal. Statistics and Probability Letters 21 (5), 417–419.

Park, P.S., Kshirsagar, A.M., 1994. Distances between normal populations when covariance matrices are unequal. Communications in Statistics—Theory and Methods 23 (12), 3549–3556.

Roy, A., Leiva, R., 2007. Discrimination with jointly equicorrelated multi-level multivariate data. Advances in Data Analysis and Classification 1 (3), 175–199.

Seber, G.A.E., 1984. Multivariate Observations. Wiley, New York.

Stanimirova, I., Simeonov, V., 2005. Modeling of environmental four-way data from air quality control. Chemometrics and Intelligent Laboratory Systems 77, 115–121.

Wu, T.T., Lange, K., 2010. Multicategory vertex discriminant analysis for high-dimensional data. The Annals of Applied Statistics 4 (4), 1698–1721.

Yener, B., Acar, E., Aguis, P., Bennett, K., 2008. Multiway modeling and analysis in stem cell systems biology. BMC Systems Biology 2.

Ziegler, A., 2011. Generalized Estimating Equations. Springer, New York.