THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

Working Paper SERIES

WP # 0048MSS-253-2013

Date May 29, 2013

Supervised classifiers of ultra high-dimensional higher-order data with locally doubly exchangeable covariance structure

Tatjana Pavlenko
Departament of Mathematics
Royal Institute of Technology KTH

Anuradha Roy
Department of Management Science and Statistics
The University of Texas at San Antonio
One UTSA Circle, San Antonio, TX 78249 USA

UTSA COLLEGE OF BUSINESS KNOWLEDGE FOR A NEW WORLD

ONE UTSA CIRCLE
SAN ANTONIO, TEXAS 78249-0631
210 458-4317 | BUSINESS.UTSA.EDU

Supervised classifiers of ultra high-dimensional higher-order data with locally doubly exchangeable covariance structure

Tatjana Pavlenko
Departament of Mathematics
Royal Institute of Technology KTH
SE-100 44 Stockholm, Sweden
Email: Pavlenko@math.kth.se

Anuradha Roy [*]
Department of Management Science and Statistics
The University of Texas at San Antonio
One UTSA Circle, San Antonio, TX 78249 USA
Email: Anuradha.roy@utsa.edu
Phone: +00-210-458-6343, Fax: +00-210-458-6350

We explore the performance accuracy of the linear and quadratic classifiers for ultra high-dimensional higher-order data, assuming that the class conditional distributions are multivariate normal with locally doubly exchangeable covariance structure. We derive a two-stage procedure for estimating the covariance matrix: at the first stage, the Lasso-based structure learning is applied to sparsifying the block components within the covariance matrix. At the second stage, the maximum likelihood estimators of all block-wise parameters are derived given that the within block covariance structure is doubly exchangeable and the mean vector has a Kronecker product structure. We also study the effect of the block size on the classification performance in the ultra high-dimensional setting and derive a class of asymptotically equivalent block structure approximations, in a sense that the choice of the block size is asymptotically negligible. Using synthetic data, we have shown that our new supervised decision rules are very efficient in learning by very small sized training samples and then successfully classifying the test samples.

Keywords classification rule; class of asymptotically equivalent structure approximations; locally doubly exchangeable covariance structure; graphical Lasso; maximum likelihood estimates; ultra high-dimensional higher-order data

JEL Classification: C33,C13

[*] Correspondence to: Anuradha Roy, Department of Management Science and Statistics, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA

1 Introduction

Analysis of ultra high-dimensional higher-order data (UHDHOD) is a mathematical challenge of this Century. High-dimensional data is data with anywhere from a few dozen to thousands of dimensions, whereas ultra high-dimensional data are those that goes beyond many thousands of dimensions and can usually be obtained in the same way as high-dimensional data. However, unlike the high-dimensional data, higher-order data can be arranged in hypercubes as opposed to vectors. In this article, we develop both linear and quadratic classifiers for ultra high-dimensional third-order data, assuming that $\mathcal{C}$ different normally distributed classes with a locally doubly exchangeable covariance structure (defined in Section 3.1) and with a constant mean vector over space (CMVOS) (defined in Section 3.2) are given.

Even though dimensionality is a curse, but at the same time a source of incredible information too. We circumvent the curse of dimensionality by developing successful algorithms to reduce the dimensionality in a natural and meaningful way. One of the challenging problems with the UHDHOD is to deal with the estimation of enormous variance-covariance matrix, which represents complex dependence structures among the variables and over the space-time points. In this framework, the number of samples is assumed to be much less than the total dimensions of the UHDHOD. One can achieve this by imposing some appropriate variance-covariance structure so that it captures the "natural" structure of the data with much less number of samples. This may be achieved by first selecting the essential variables (Yu and Liu, 2003) and then choosing the appropriate covariance structure over space-time points of the data. Reduction of dimensionality over space-time points of the data is also another option. In this article we combine the Lasso based covariance structure learning with imposing locally doubly exchangeability assumption.

In our current study we explore the block-wise sparsity, which leads to the additive structure of the resulting classifier and allows much simpler asymptotic theory for evaluating the classification accuracy, as it is shown in Pavlenko and Björkstrom (2010). Moreover, this approach allows for much simpler computational methods for estimating classifier; see details in Pavlenko, Björkstrom and Tillander (2012). They studied the classification problem in high dimensional first order data based on exploring sparsity patterns in the data dependence first, and then computing the estimate of the inverse variance-covariance matrix using constrained maximum likelihood. In their study, rather than restricting themselves to methods that completely ignore potential dependence structure they tried to recover it from the data and then used it to their advantage. They used the popular technique *graphical* Lasso or gLasso (Friedman, Hastie and Tibshirani, 2008) in learning the sparsity patterns. Structured covariance matrix essentially simplifies high-dimensional statistical procedures such as linear and quadratic discriminant analysis as well as Bayesian predictive classification (Corander et al., 2012).

Structural patterns in the covariance matrix can be anticipated in biological applications, where the dependencies among the genes reflect the underlying molecular mechanisms. For example, in tumor classification, genes can be grouped into pathways, so that the connection within a pathway is stronger than between pathways. This type of structure can provide some insight into the relationship between the gene expression level and a tumor type.

To tackle UHDHOD, we apply gLasso based blocked sparcifying covariance structure approximation at the first stage. Given blocked variables corrresponding to the gLasso step, we at the second stage integrate the higher-order variables (e.g., space-time points) into each blocked variables. Within each block we assume doubly exchangeable covariance structure which was extensively studied by Roy and Leiva (2007) and Leiva and Roy (2011, 2012). These two authors have introduced many covariance structures $\big($Roy and Leiva (2007), Leiva and Roy (2009, 2011, 2012)$\big)$ for high-dimensional third order $\big(\textit{variables } (p) \times \textit{sites } (u) \times \textit{time points } (v)\big)$ data in the context of classification problem. In the current study we assume that structure learning is performed with $u = 1$ and $v = 1$ at the first stage, and the structure remains stable over all space and time points. Roy and Leiva (2007) and Leiva and Roy (2011, 2012) used doubly exchangeable covariance structure, that allows to partition a covariance structure into three unstructured covariance matrices, corresponding to each of the three orders, in the classification problem. In this paper, we have shown by simulation study that our new classification rules performs better for smaller block sizes.

Kroonenberg (2008) discussed practical issues in applying higher-order or multiway component techniques to multiway data with an emphasis on methods for three-way data. Akdemir and Gupta (2011) have developed classification techniques for high dimensional multiway data. In their paper Akdemir and Gupta presented a technique called slicing for obtaining an approximate nonsingular estimate of the covariance matrix for high-dimensional data when sample size is less than the dimension of the observed vector. Dudoit et al. (2002) and Lai et al. (2006) developed classification rules for tumor samples using thousands of gene expression profiles with at most hundreds of samples. Bhattacharya et al. (2003) proposed a classifier called *Liknon* that simultaneously performs classification and relevant gene identification. Liknon is trained by optimizing a linear discriminant function with a penalty constraint via linear programming. Most recently, Kim and Simon (2011) developed probabilistic classifiers which use the probabilities in conjunction with other information such as treatment options and patient preferences for making complex integrated clinical decisions.

To the best of the authors' knowledge, classification of UHDHOD with limited samples has not yet been studied. To get the intensity levels of thousands of gene expressions (p), the lab scientists generally obtain the intensity levels in three probes (u) for each gene and then average them out to get one intensity level for each gene. Then they observe the intensity level for each gene over a period of time (v), generally over years. Gene expression from a diseased tissue would change over time, whereas gene expression from a healthy tissue would not change. Therefore, by allowing monitoring of expression levels in cells for thousands of genes simultaneously over the years, microarray experiments may lead to a more complete understanding of the molecular variations between the healthy and the diseased cells, or among different types of diseased cells, and hence to a finer and better classification, and ultimately to a more reliable diagnosis. Furthermore, classification accuracy can be further improved by including the intensity levels of the three probes separately (not averaged) in the classification rules, as it has been observed that the introduction of more orders of data in the discriminant analysis increases the classification accuracy (Leiva and Roy, 2009). Thus, in the end by analyzing ultra high-dimensional third order (*genes* × *probes* × *time points*) data one can classify different prognostic groups of patients by assessing disease heterogeneity and for design and stratification of future clinical trials. Patterns of cancer or any other life-threatening disease treatment are changing very rapidly, and it is important that the results of the present analysis be applicable to contemporary patients.

The rest of the article is organized as follows. We first set up a supervised classification problem and briefly describe learning block-diagonal covariance structure using gLasso in Section 2. We then introduce locally doubly exchangeable covariance structure and corresponding mean vector structure within each block in Section 3, and derive their maximum likelihood (ML) estimators in Section 4. Our new classification techniques, both linear and quadratic are derived in Section 5 together with some generalization of the linear rule. In Section 6, the asymptotic effect of the block size on the classification accuracy is studied in ultra high-dimensional framework. In Section 7, we report the results of some simulation studies that illustrate performance properties of our new classifiers. Finally, Section 8 concludes with several comments and the scope for the future research. Technical derivation of the MLEs of all unknown parameters and the proof of Proposition 1 are presented in two appendices.

2 Background and problem set-up

In this article we focus on a *supervised* classification problem, where each observed individual $\boldsymbol{x}_r^{(c)}$ belongs to one of the $\mathcal{C}$ classes, $\Pi_1, \ldots, \Pi_c$. Let $\boldsymbol{x}_r^{(c)}$ be the *puv*-variate vector of all measurements corresponding to the r^{th} individual in the c^{th} class, $c = 1, \ldots, \mathcal{C}$, $r = 1, \ldots, n^{(c)}$. We partition this vector $\boldsymbol{x}_r^{(c)}$ as follows:

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

for $t = 1, \ldots, v$, $s = 1, \ldots, u$. The $(p \times 1)$ ultra high-dimensional vector of measurements $\boldsymbol{x}_{r,ts}^{(c)}$ represents the r^{th} replicate (individual) in the c^{th} class on the s^{th} site (space) and at the t^{th} time point. Each observation $\boldsymbol{x}_r^{(c)}$ for a fixed space-time point, e.g., for $u = 1$ and $v = 1$ is represented by a set of $(p \times 1)$ dimensional variables $(x_{r,1}^{(c)}, \ldots, x_{r,p}^{(c)})'$ in class c, and we assume that $\boldsymbol{x}_r^{(c)} \in N_p(\boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}})$. A decision rule, $y(\boldsymbol{x}_r^{(c)})$ is a function $y : \Re^p \to \{1, \ldots, \mathcal{C}\}$ defined for all

$\boldsymbol{x}_r^{(c)} \in \Re^p$. Assuming that $\Pi_1, \ldots, \Pi_c$ are modeled by normal distribution, we assign a new $(p \times 1)$ dimensional observation $\boldsymbol{x}_0$ to class $\Pi_{c'}$, i.e., $y(\boldsymbol{x}_0) = c'$ if $c' = \arg\max_{c=1,\ldots,\mathcal{C}} \ell^{(c)}(\boldsymbol{x}_0)$, where

$$\ell^{(c)}(\boldsymbol{x}_0; \boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}) = \boldsymbol{x}_0' \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}^{(c)}} - \frac{1}{2} \boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}' \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}^{(c)}} + \ln \pi_c, \tag{1}$$

with π_c is the a priori probability of the class c, and $\sum_{c=1}^{\mathcal{C}} \pi_c = 1$. This classifier is analogous to the well-known Fisher linear discriminant score that is optimal in the sense of minimum overall misclassification probability defined as $\mathcal{E} = \sum_{c=1}^{\mathcal{C}} \pi_c \mathcal{E}_c = \sum_{c=1}^{\mathcal{C}} \pi_c P(y(\boldsymbol{x}_0) \neq c | \boldsymbol{x}_0 \in \Pi_c)$.

For $\mathcal{C} = 2$ and $\boldsymbol{\Gamma}_{\boldsymbol{x}^{(1)}} = \boldsymbol{\Gamma}_{\boldsymbol{x}^{(2)}} = \boldsymbol{\Gamma}_{\boldsymbol{x}}$, (1) can be represented as

$$\ell(\boldsymbol{x}_0; \boldsymbol{\mu}_{\boldsymbol{x}^{(1)}}, \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}) = \left(\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}} + \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}) \right)' \cdot \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \cdot (\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}} - \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}) \lesseqgtr \ln \frac{\pi_2}{\pi_1}, \tag{2}$$

and to measure its performance accuracy we turn to the maximum conditional misclassification probability, $\mathcal{E}$ that is defined as

$$\max_{i=1,2}\{\mathcal{E}_i\} = \max_{i=1,2} \left\{ P\left(\ell(\boldsymbol{x}_0; \boldsymbol{\mu}_{\boldsymbol{x}^{(1)}}, \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}) \leq 0 | \boldsymbol{x}_0 \in \Pi_1\right), P(\ell(\boldsymbol{x}_0; \boldsymbol{\mu}_{\boldsymbol{x}^{(1)}}, \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}) > 0 | \boldsymbol{x}_0 \in \Pi_2) \right\}, \tag{3}$$

assuming that $\pi_1 = \pi_2 = 1/2$. Assume further that $\boldsymbol{x}_0 \in \Pi_1$, with known $\boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}$s and $\boldsymbol{\Gamma}_{\boldsymbol{x}}$, then $\ell(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}^{(1)}}, \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}})$ is also normally distributed and corresponding optimal misclassification probability can be expressed as

$$\mathcal{E}_{opt} = \Phi\Big(-\frac{1}{2} \frac{\mathrm{E}\left[\ell(\boldsymbol{x}_0; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}) | \boldsymbol{x} \in \Pi_1\right]}{\sqrt{\mathrm{Var}\left[\ell(\boldsymbol{x}_0; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}) | \boldsymbol{x} \in \Pi_1\right]}} \Big) = \Phi\Big(-\frac{1}{2} \delta(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}) \Big), \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $\delta^2(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}) = \boldsymbol{\mu}_{\boldsymbol{x}}' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}}$ is the square of the Mahalanobis distance between the classes Π_1 and Π_2 with a shift vector $\boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{\mu}_{\boldsymbol{x}^{(1)}} - \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}$. Asymptotic properties of the estimated misclassification probability of the type (4) are extensively studied by McLachlan (2004), Shutoh (2011) and Shutoh et al. (2011).

In this article we extend the above consideration to the higher order data (e.g., space-time points). Now, let t and s denote a given point in time and a given site respectively. Let $\boldsymbol{x}_{ts}^{(c)} : (\Omega, C) \to \Re^p$, $1 \leq t \leq v$, $1 \leq s \leq u$, be the $p-$dimensional normally distributed random vector from the c^{th} population. Then the random families $(\boldsymbol{x}_{1s}^{(c)})_{s \in \{1,\ldots,u\}}, \ldots, (\boldsymbol{x}_{vs}^{(c)})_{s \in \{1,\ldots,u\}}$ are assumed to be exchangeable. Furthermore, for fixed t, the family of random variables $(\boldsymbol{x}_{ts}^{(c)})_{s \in \{1,\ldots,u\}}$ 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 third order multivariate data, and thus may offer more information about the true association of the data. The major advantage of this covariance structure is that the measurements over space and time need not be equally spaced over space and time. Observe that for $u = 1$ and $v = 1$ we arrive to the classifier (1).

2.1 Block-diagonal covariance structure approximation using gLasso: single class

As remarked in the Introduction, since computation of the sample based covariance matrix for ultra high-dimensional data is computer-intensive, it is always an advantage to reduce the dimension of the observed vectors. In particular, imposing a proper structure on the covariance matrix can essentially reduces the number of parameters to be estimated. For example, assuming that $\mathbf{\Gamma}$ is a $(p \times p)$ dimensional block diagonal matrix, i.e., $\mathbf{\Gamma} = \operatorname{diag}\left[\mathbf{\Gamma}_{[1]}, \ldots, \mathbf{\Gamma}_{[b]}\right]$ for any space-time point, where $\mathbf{\Gamma}_{[j]}$ has dimension $(p_j \times p_j)$ for $j = 1, \ldots, b$, we need to estimate only $\sum_{j=1}^{b} p_j(p_j + 1)/2$ unknown parameters, instead of $p(p+1)/2$, and assuming that $p_j \ll n$ a local estimation of each block-diagonal entry of $\mathbf{\Gamma}$ can be obtained using standard ML approach. However, assumption of the block-diagonal structure on $\mathbf{\Gamma}$ seems to be too strong; in the case of normal class conditional distributions, this assumption is equivalent to the independence of the corresponding sets of blocks of the observed vector $\boldsymbol{x}$. Therefore, instead of imposing the block diagonal structure on $\mathbf{\Gamma}$ we learn it from the data. In this section, we briefly review our results on the learning procedure developed in Pavlenko et al. (2012), which uses the Lasso-based technique (gLasso) that relates sparse covariance model selection in learning a Gaussian graph structure by using l_1 regularization. In what follows, we assume that the structure learning is applied at the first space-time point, that is $u = 1$ and $v = 1$, and then assume that the structure remains stable over all the space-time points. Interpretation of this assumption is very natural in the genetic data, where it is reasonable to assume that the genes are groupped into types, which have similar connectivity or correlation patterns. For example, genes can be grouped into pathways with certain biological interpretation, where the dense connection within a pathway is more likely than the connection between pathways, and this structure remains stable over probes and time.

To describe the covariance learning procedure, we introduce an undirected graph $\mathcal{G}$ on p nodes which is defined by the ordered tuple $\mathcal{G} = \{\mathcal{X}, \mathbf{\Theta}\}$, where $\mathcal{X}$ is the set of nodes associated with the observed vector $\boldsymbol{x}$ and $\mathbf{\Theta}$ is the set of undirected edges. Then, the Gaussian random graph associated with $\mathcal{G}$ over the random vector $\boldsymbol{x}$ is the family of p-variate normal distributions with the inverse covariance, or *concentration* matrix $\mathbf{\Gamma}^{-1}$, that represents the edge structure of the graph, in the sense that $\gamma_{ij}^{-1} = 0$ if $(i,j) \notin \mathbf{\Theta}$. Hence, the sparsity pattern of $\mathbf{\Gamma}^{-1}$ reflects the conditional independence among the entries of $\boldsymbol{x}$. In particular, by the Hammersley-Clifford theorem (see Lauritzen, 1996), it holds that $\gamma_{ij}^{-1} = 0$ for all $(i,j) \notin \mathbf{\Theta}$. Therefore, the problem of learning the Gaussian graph structure is equivalent to estimating the off-diagonal zero pattern of the concentration matrix, i.e., the set

$$\mathbf{\Theta} := \{i, j \in \mathcal{X} | i \neq j,\ \gamma_{ij}^{-1} \neq 0\}.$$

For the graph structure learning, we focus on the minimizer of the negative l_1-penalized log-likelihood

$$\widetilde{\mathbf{\Gamma}}_{\lambda}^{-1} = \arg\min_{\mathbf{\Gamma}^{-1}\succ 0} \left[\operatorname{Tr}\left(\mathbf{\Gamma}^{-1}\widehat{\mathbf{\Gamma}} \right) - \ln|\mathbf{\Gamma}^{-1}| + \lambda \|\mathbf{\Gamma}^{-1}\|_1 \right], \tag{5}$$

where $\widehat{\mathbf{\Gamma}}$ is the ML estimator of $\mathbf{\Gamma}$, $\|\mathbf{\Gamma}^{-1}\|_1 = \sum\limits_{i<j} |\gamma_{ij}^{-1}|$, λ is a non-negative tuning parameter, and the minimization is taken over symmetric positive definite matrices. We refer to the optimization problem (5) as the gLasso (Friedman et al. 2007). The loss function in (5) is invariant to permutations of variables, and the problem always has a unique solution as the negative log-determinant is a strictly convex function. Moreover, the solution is positive definite for all $\lambda > 0$ even if $\mathbf{\Gamma}$ is singular, and for sufficiently large λ, the estimate $\tilde{\mathbf{\Gamma}}_{\lambda}^{-1}$ will be sparse due to its nature, l_1-penalty works by pushing the off-diagonal elements to zero, thereby inducing sparsity in the resulting estimator (Tibshirani, 1996). Similar estimates of $\mathbf{\Gamma}^{-1}$ were also considered in Rothman et al. (2008), where different strngth of penalization was applied for diagonal and off diagonal elements of $\mathbf{\Gamma}^{-1}$.

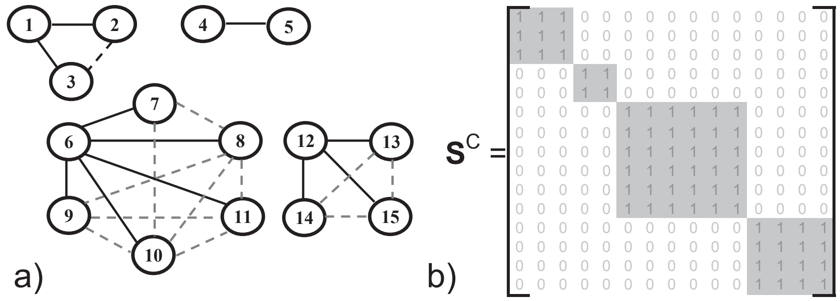
The sparsity pattern of the solution of (5) gives rise to the sparse symmetric edge *skeleton*, defined as $\mathcal{S}^{\lambda} := \mathbf{1}_{\{\tilde{\gamma}_{ij}^{-1} > \lambda\}}$ which in turn generates the estimated concentration graph $\mathcal{G}^{\lambda} = \{\mathcal{X}, \mathbf{\Theta}^{\lambda}\}$. Suppose now that for a specific λ, $\mathcal{G}^{\lambda}$ allows for a decomposition into b_{λ} connected components as $\mathcal{G}^{\lambda} = \bigcup\limits_{j=1}^{b_{\lambda}} \{\mathcal{X}_j^{\lambda}, \mathbf{\Theta}_j^{\lambda}\}$, where by a connected component we mean a maximal connected subgraph of $\mathcal{G}^{\lambda}$. Observe that $b_{\lambda} \in \{1, \dots, p\}$, so that $b_{\lambda} = p$ for large λ and $b_{\lambda} = 1$ for small λ. The former case implies all the components are isolated, i.e., have size 1, whereas the latter case implies there is only one component of size p. The connected components obtained from the decomposition of $\mathcal{G}^{\lambda}$ lead to the block-diagonal form of the edge-matrix concentration graph

$$\mathbf{\Theta}^{\lambda} = \begin{pmatrix} \mathbf{\Theta}_1^{\lambda} & 0 & \dots & 0 \\ 0 & \mathbf{\Theta}_2^{\lambda} & 0 & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \mathbf{\Theta}_{b_{\lambda}}^{\lambda} \end{pmatrix}, \tag{6}$$

where the different components represent blocks of indices specified by $\mathcal{X}_i^{\lambda}, i = 1, \dots, b_{\lambda}$. Figure 1 illustrates correspondence between the graph structure (panel(a)), and the block-wise sparsity pattern of the concentration matrix (panel (b)).

We then construct a matrix $\widehat{\mathbf{\Gamma}}_{\lambda} = \operatorname{diag}\left[\widehat{\mathbf{\Gamma}}_{\lambda,[1]}, \dots, \widehat{\mathbf{\Gamma}}_{\lambda,[b]}\right]$ having the same block-diagonal structure as (6), by the solution of (5), so that under certain assumptions on the maximal size of the connected components, the problem of estimation of $\mathbf{\Gamma}$ can be reduced to b_{λ} constrained ML estimators of $\mathbf{\Gamma}_i$s, that can be solved independently, The resulting covariance structure has only $\sum\limits_{j=1}^{b_{\lambda}} p_j(p_j+1)/2$ unknown parameters, which is much less than $p(p+1)/2$ as mentioned before.

A correlation based version of (5) is explored in Pavlenko et al. (2012), and a two-stage estimation procedure yielding a block-diagonal estimator of $\mathbf{\Gamma}^{-1}$ is suggested. Authors develop an



$$\mathbf{S}^{\mathrm{C}} = \begin{bmatrix}
1&1&1&0&0&0&0&0&0&0&0&0&0&0&0\\
1&1&1&0&0&0&0&0&0&0&0&0&0&0&0\\
1&1&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&1&1&1&1
\end{bmatrix}$$

Figure 1: (a) A Gaussian graph having $p = 15$ nodes that is decomposed into a set $b = 4$ connected components. (b) The block-wise sparse covariance structure (stabilized skeleton, $\mathcal{S}^{\mathrm{C}}$) where zero pattern of the inverse covariance matrix is associated with the graph in (a).

empirical test procedure that uses the gLasso-based sparsity pattern of the solution of (5) and generates the *stabilized* edge skeleton, $\mathcal{S}^{\lambda}$ meaning that only significant non-zeros are included; see Algorithm 1 in Pavlenko et al. (2012). They also have shown that gLasso-based edge skeleton followed by a suitable reordering of the graph nodes (see Algorithm 2 in Pavlenko et al., 2012), induces a *quasi-decomposition* of $\mathcal{G}^{\lambda}$ into connected components, meaning that after omitting a small number/portion of edges a decomposition $\{\mathcal{X}, \mathbf{\Theta}^{\lambda}\} = \overset{b_{\lambda}}{\underset{j=1}{\cup}} \{\mathcal{X}_j, \mathbf{\Theta}_j^{\lambda}\}$ holds. Observe also that the step from a quasi-decomposition to an exact partition of $\mathcal{G}^{\lambda}$ is not unique, i.e., it yields a family of decompositions; see details in Pavlenko et al. (2012)) where the sensitivity of Cuthil-McKee reordering transform used in Algorithm 2 in Pavlenko et al. (2012) to the choice of the initial node is discussed. It is also shown that for a given regularization strength λ and with certain constraints on the maximum size of connected component, the exact choice of the decomposition is asymptotically negligible.

In the present paper we explore the two-stage estimation procedure as described in the introduction in the classification framework. Our focus is mainly be on the second stage, i.e., we assume that the block structure of $\mathbf{\Gamma}^{-1}$ is established in (6) for a specific λ under constrained block size,

$$\max_{j=1,\dots,b_{\lambda}} p_j < n/uv, \tag{7}$$

where n is the size of the training data. Note that if $\mathbf{\Gamma}^{-1}$ is block-diagonal then so is its inverse, $\mathbf{\Gamma}$, therefore consideration in what follows will be focused on modeling $\mathbf{\Gamma}$ given that the decomposition $\{\mathcal{X}, \mathbf{\Theta}^{\lambda}\} = \overset{b_{\lambda}}{\underset{j=1}{\cup}} \{\mathcal{X}_j, \mathbf{\Theta}_j^{\lambda}\}$ holds. Unlike Pavlenko et al. (2012), where the within-block covariance components are assumed to be unstructured, in this article we impose jointly equicorrelated covariance structure for all space-time data within each block entry, $\mathbf{\Theta}_j^{\lambda}$ thereby allowing a modeling of a higher-order data. In the following sections we introduce locally jointly equicorrelated covariance structure and a Kronecker product structured mean vector for third-order multivariate data.

3 Covariance and mean vector structures

In this section, we introduce the locally jointly equicorrelated covariance structure and a Kronecker product structured mean vector for third-order multivariate data. We also present some auxiliary matrix results.

3.1 Covariance structure with local double exchangeability

Definition 1. *Let $\boldsymbol{x}_r$ be an puv−variate partitioned real-valued random vector $\boldsymbol{x}_r = \left(\boldsymbol{x}'_{r,[1]},\ldots,\boldsymbol{x}'_{r,[b]}\right)'$, where $\boldsymbol{x}_{r,[j]} = \left(\boldsymbol{x}'_{r,[j],1},\ldots,\boldsymbol{x}'_{r,[j],v}\right)'$, for $j = 1,\ldots,b$ with $\boldsymbol{x}'_{r,[j],t} = \left(\boldsymbol{x}'{}_{r,[j],t1},\ldots,\boldsymbol{x}'{}_{r,[j],tu}\right)'$ for $t = 1,\ldots,v$ and $\boldsymbol{x}'_{r,[j],ts} = \left(x_{r,[j],ts,1},\ldots,x_{r,[j],ts,p_j}\right)'$ for $s = 1,\ldots,u$. Let $E[\boldsymbol{x}_r] = \boldsymbol{\mu}_{\boldsymbol{x}} \in \Re^{puv}$, and $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ be the $(puv\times puv)-$dimensional partitioned covariance matrix $Cov\left[\boldsymbol{x}_r\right] = \left(\boldsymbol{\Gamma}_{\boldsymbol{x}_{r,t},\boldsymbol{x}_{r,t^*}}\right) = \left(\boldsymbol{\Gamma}_{r,tt^*}\right)$, where $\boldsymbol{\Gamma}_{r,tt^*} = Cov\left[\boldsymbol{x}_{r,t},\boldsymbol{x}_{r,t^*}\right]$ for $t, t^* = 1,\ldots,v$.*

The $p-$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 locally jointly equicorrelated if $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is given by

$$\boldsymbol{\Gamma}_{\boldsymbol{x}} = diag\left(\boldsymbol{\Gamma}_{[1]},\boldsymbol{\Gamma}_{[2]},\ldots,\boldsymbol{\Gamma}_{[b]}\right), \tag{8}$$

where

$$\boldsymbol{\Gamma}_{[j]} = \left[\begin{array}{cccc|cccc|c|cccc}
\boldsymbol{U}_{0[j]} & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{U}_{1[j]} & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} \\
\boldsymbol{U}_{1[j]} & \boldsymbol{U}_{0[j]} & \cdots & \boldsymbol{U}_{1[j]} & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{U}_{1[j]} & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{U}_{0[j]} & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} \\
\hline
\boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{U}_{0[j]} & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} \\
\boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{U}_{1[j]} & \boldsymbol{U}_{0[j]} & \cdots & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{U}_{1[j]} & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{U}_{0[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} \\
\hline
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
\boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{U}_{0[j]} & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{U}_{1[j]} \\
\boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{U}_{1[j]} & \boldsymbol{U}_{0[j]} & \cdots & \boldsymbol{U}_{1[j]} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{W}_{[j]} & \cdots & \boldsymbol{U}_{1[j]} & \boldsymbol{U}_{1[j]} & \cdots & \boldsymbol{U}_{0[j]}
\end{array}\right], \tag{9}$$

$\boldsymbol{U}_{0[j]}$ is a positive definite symmetric $p_j \times p_j$ matrix, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ are symmetric $p_j \times p_j$ matrices, and $j = 1,\ldots,b$. The variance covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is then said to have a locally jointly equicorrelated covariance structure with sets of equicorrelation parameters $\left\{\boldsymbol{U}_{0[1]},\ldots,\boldsymbol{U}_{0[b]}\right\}$, $\left\{\boldsymbol{U}_{1[1]},\ldots,\boldsymbol{U}_{1[b]}\right\}$ and $\left\{\boldsymbol{W}_{[1]},\ldots,\boldsymbol{W}_{[b]}\right\}$ such that $\sum_{j=1}^{b} p_j = p$. The matrices $\boldsymbol{U}_{0[j]},\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ for $j = 1,\ldots,b$ are all unstructured.

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

$$\text{Cov}\left[\boldsymbol{x}_{r,ts}; \boldsymbol{x}_{r,t^*s^*}\right] = \begin{cases} \boldsymbol{U}_{0[j]} & \text{if} \quad t = t^* \quad \text{and} \quad s = s^*, \\ \boldsymbol{U}_{1[j]} & \text{if} \quad t = t^* \quad \text{and} \quad s \neq s^*, \\ \boldsymbol{W}_{[j]} & \text{if} \quad t \neq t^*, \end{cases}$$

for all $j = 1, \ldots, b$, that is,

$$\begin{aligned} \boldsymbol{\Gamma}_{[j]} &= \boldsymbol{I}_{uv} \otimes \boldsymbol{U}_{0[j]} + \left[\boldsymbol{I}_v \otimes (\boldsymbol{J}_u - \boldsymbol{I}_u)\right] \otimes \boldsymbol{U}_{1[j]} + \left[\boldsymbol{J}_{uv} - (\boldsymbol{I}_v \otimes \boldsymbol{J}_u)\right] \otimes \boldsymbol{W}_{[j]} \\ &= \boldsymbol{I}_{uv} \otimes \left(\boldsymbol{U}_{0[j]} - \boldsymbol{U}_{1[j]}\right) + \boldsymbol{I}_v \otimes \boldsymbol{J}_u \otimes \left(\boldsymbol{U}_{1[j]} - \boldsymbol{W}_{[j]}\right) + \boldsymbol{J}_{uv} \otimes \boldsymbol{W}_{[j]}, \end{aligned}$$

for all $j = 1, \ldots, b$,, where $\boldsymbol{I}_a$ is the $a \times a$ identity matrix, and $\boldsymbol{J}_a = \mathbf{1}_a \mathbf{1}_a'$.

The $p_j \times p_j$ diagonal blocks $\boldsymbol{U}_{0[j]}$ in (9) represent the variance-covariance matrix of the p_j response variables at any given site and at any given time point, whereas the $p_j \times p_j$ off-diagonal blocks $\boldsymbol{U}_{1[j]}$ in (9) represent the covariance matrix of the p_i response variables between any two sites (probes) and at any given time point. We assume $\boldsymbol{U}_{0[j]}$ is constant for all sites and time points, and $\boldsymbol{U}_{1[j]}$ is same for all site pairs and for all time points. The $p_j \times p_j$ off-diagonal blocks $\boldsymbol{W}_{[j]}$ represent the covariance matrix of the p_j 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.

Observe that due to its doubly exchangeable nature, each component diagonal block of $\boldsymbol{\Gamma}_{\boldsymbol{x}}$, $\boldsymbol{\Gamma}_{[j]}$ is also called doubly exchangeable covariance structure; see e.g., Roy and Leiva (2007). However, within-block double exchangeability does not imply that the entire $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ has doubly exchangeable structure since the block size varies with j. Thus, we call the structure $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ defined in (8-9) as *locally* doubly exchangeable. Thus, we see that our locally jointly equicorrelated covariance structure generalizes Roy and Leiva's (2007) jointly equicorrelated covariance structure.

3.2 Mean vector structure

Like covariance structure mean vector can have some structure too. In this article in addition to locally jointly equicorrelated covariance structure we consider the following Kronecker product structure on the mean vector

$$\boldsymbol{\mu}_{\boldsymbol{x}} = \left(\boldsymbol{\mu}_{\boldsymbol{x}[1]}', \ldots, \boldsymbol{\mu}_{\boldsymbol{x}[b]}'\right)', \quad \text{where} \quad \boldsymbol{\mu}_{\boldsymbol{x}[j]} = \left(\boldsymbol{\mu}_{\boldsymbol{x}[j],1}', \ldots, \boldsymbol{\mu}_{\boldsymbol{x}[j],v}'\right)',$$

with $\boldsymbol{\mu}_{\boldsymbol{x}[j],t} = \mathbf{1}_u \otimes \boldsymbol{\mu}_{[j],t}$ and $\boldsymbol{\mu}_{[j],t} \in \Re^{p_j}$, for $t = 1, \ldots, v$, $j = 1, \ldots, b$, and $c = 1, \ldots, \mathcal{C}$ for the third-order multivariate data. Gene expression levels are expected to be constant over all the three probes (sites), so we consider the above 'Constant mean vector structure over sites' (CMVOS) as a natural mean structure for this article.

3.3 Matrix results

Lemma 1. *Let $\mathbf{\Gamma}_{[j]}$ be the doubly exchangeable covariance matrix for the j^{th} block as in equation (9) of Definition 1.*

1. *If*

$$\begin{aligned}
\mathbf{\Delta}_{1[j]} &= \boldsymbol{U}_{0[j]} - \boldsymbol{U}_{1[j]}, & \text{(10a)}\\
\mathbf{\Delta}_{2[j]} &= \boldsymbol{U}_{0[j]} + (u-1)\,\boldsymbol{U}_{1[j]} - u\boldsymbol{W}_{[j]} = \left(\boldsymbol{U}_{0[j]} - \boldsymbol{U}_{1[j]}\right) + u\left(\boldsymbol{U}_{1[j]} - \boldsymbol{W}_{[j]}\right), \quad \textit{and} & \text{(10b)}\\
\mathbf{\Delta}_{3[j]} &= \boldsymbol{U}_{0[j]} + (u-1)\,\boldsymbol{U}_{1[j]} + u\,(v-1)\,\boldsymbol{W}_{[j]} = \left(\boldsymbol{U}_{0[j]} - \boldsymbol{U}_{1[j]}\right) + u\left(\boldsymbol{U}_{1[j]} - \boldsymbol{W}_{[j]}\right) + uv\boldsymbol{W}_{[j]}, & \text{(10c)}
\end{aligned}$$

are non singular matrices, the matrix $\mathbf{\Gamma}_{[j]}$ is non singular, and its inverse is given by

$$\mathbf{\Gamma}_{[j]}^{-1} = \boldsymbol{I}_{vu} \otimes \mathbf{\Delta}_{1[j]}^{-1} + \boldsymbol{I}_{v} \otimes \boldsymbol{J}_{u} \otimes \frac{1}{u}\left(\mathbf{\Delta}_{2[j]}^{-1} - \mathbf{\Delta}_{1[j]}^{-1}\right) + \boldsymbol{J}_{vu} \otimes \frac{1}{vu}\left(\mathbf{\Delta}_{3[j]}^{-1} - \mathbf{\Delta}_{2[j]}^{-1}\right). \tag{11}$$

Thus, we see that $\mathbf{\Gamma}_{[j]}^{-1}$ has the same structure as $\mathbf{\Gamma}_{[j]}$. Therefore, the doubly exchangeable covariance structure is invariant with respect to inverse, and so is local doubly exchangeable covariance structure.

2. *The determinant of $\mathbf{\Gamma}_{[j]}$ is given by*

$$\left|\mathbf{\Gamma}_{[j]}\right| = \left|\mathbf{\Delta}_{1[j]}\right|^{v(u-1)} \left|\mathbf{\Delta}_{2[j]}\right|^{(v-1)} \left|\mathbf{\Delta}_{3[j]}\right|.$$

See Roy and Leiva (2007) for the proof of this lemma.

Lemma 2. *If $\mathbf{\Gamma}_{\boldsymbol{x}}$ be a locally jointly equicorrelated with b blocks as follows*

$$\mathbf{\Gamma}_{\boldsymbol{x}} = diag\left(\mathbf{\Gamma}_{[1]}, \mathbf{\Gamma}_{[2]}, \ldots, \mathbf{\Gamma}_{[b]}\right),$$

then its inverse is given by

1.
$$\mathbf{\Gamma}_{\boldsymbol{x}}^{-1} = diag\left(\mathbf{\Gamma}_{[1]}^{-1}, \mathbf{\Gamma}_{[2]}^{-1}, \ldots, \mathbf{\Gamma}_{[b]}^{-1}\right), \tag{12}$$

and its determinant is given by

2.
$$\left|\mathbf{\Gamma}_{\boldsymbol{x}}\right| = \left|\mathbf{\Gamma}_{[1]}\right| \left|\mathbf{\Gamma}_{[2]}\right| \cdots \left|\mathbf{\Gamma}_{[b]}\right|. \tag{13}$$

These results are used in Section 4 to obtain the maximum likelihood estimate (MLE) of the doubly exchangeable covariance matrix $\mathbf{\Gamma}_{\boldsymbol{x}}$. Now, let $\boldsymbol{x}^{(c)}$ represent the $puv-$variate vector of all measurements corresponding to one individual in the c^{th} class where we assume a distribution

$N_{puv}\left(\boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}\right)$, and let $\boldsymbol{x}_1^{(c)}, \ldots, \boldsymbol{x}_{n^{(c)}}^{(c)}$ be a random sample of size $n^{(c)}$ of $\boldsymbol{x}^{(c)}$. The unstructured variance-covariance matrix Cov $\left[\boldsymbol{x}^{(c)}\right] = \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}$ has $q = puv\,(puv + 1)\,/2$ unknown parameters, which can be large for arbitrary values of p, u or v. However, to obtain a suitable covariance structure one needs to take into account the characteristics of the experimental design. One may assume a "jointly equicorrelated covariance" structure in the situation, where the data is multivariate with three orders. The biggest advantage of using this jointly equicorrelated covariance structure is that the double exchangeability in this structure considerably reduces the number of unknown parameters, and thus offers more reliable estimates. The resulting structure has only $\sum_{j=1}^{b} 3p_j\,(p_j + 1)\,/2$ unknown parameters, which is much much less than q. Moreover, this number does not even depend on u and v. That means we can get more information about the data that reduces the misclassification error rates of our new classifiers without increasing the number of unknown parameters.

4 Maximum likelihood estimates using local double exchangeability: single class case

As mentioned in the introduction we assume that $\mathrm{E}[\boldsymbol{x}_r] = \boldsymbol{\mu}_{\boldsymbol{x}} = \left(\boldsymbol{\mu}'_{\boldsymbol{x}[1]}, \ldots, \boldsymbol{\mu}'_{\boldsymbol{x}[b]}\right)$, where $\boldsymbol{\mu}_{\boldsymbol{x}[j]} = \left(\boldsymbol{\mu}'_{\boldsymbol{x}[j],1}, \ldots, \boldsymbol{\mu}'_{\boldsymbol{x}[j],v}\right)'$, with $\boldsymbol{\mu}_{\boldsymbol{x}[j],t} = \mathbf{1}_u \otimes \boldsymbol{\mu}_{[j],t}$ and $\boldsymbol{\mu}_{[j],t} \in \Re^{p_j}$ for $t = 1, \ldots, v$, $j = 1, \ldots, b$, and $\mathrm{Cov}[\boldsymbol{x}_r] = \boldsymbol{\Gamma}_{\boldsymbol{x}} = \mathrm{diag}\left(\boldsymbol{\Gamma}_{[1]}, \boldsymbol{\Gamma}_{[2]}, \ldots, \boldsymbol{\Gamma}_{[b]}\right)$, where the $p_j \times p_j$ blocks $\boldsymbol{\Gamma}_{[j]}, j = 1, \ldots, b$ are given in (9). The following theorem yields explicit expressions for the MLEs of $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{x}}$. Let $T = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ denote a random training sample of size n from a class with distribution $N_{puv}\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right)$.

Theorem 1. *Under the above assumptions, the maximum likelihood estimate of* $\boldsymbol{\mu}_{\boldsymbol{x}}$ *is*

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \overline{\boldsymbol{x}} = \left(\overline{\boldsymbol{x}}_{[1]}, \ldots, \overline{\boldsymbol{x}}_{[b]}\right),$$

where $\overline{\boldsymbol{x}}_{[j]} = \left(\mathbf{1}'_u \otimes \overline{\boldsymbol{x}}'_{[j],1}, \ldots, \mathbf{1}'_u \otimes \overline{\boldsymbol{x}}'_{[j],v}\right)'$ *and* $\overline{\boldsymbol{x}}_{[j],t}$ *is the sample mean vector at time* t*, that is,*

$$\overline{\boldsymbol{x}}_{[j],t} = \frac{1}{nu} \sum_{r=1}^{n} \sum_{s=1}^{u} \boldsymbol{x}_{r,[j],ts}, \quad \textit{for } t = 1, \ldots, v,$$

and the maximum likelihood estimate of $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ *is* $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}} = \textit{diag}\left(\widehat{\boldsymbol{\Gamma}}_{[1]}, \widehat{\boldsymbol{\Gamma}}_{[2]}, \ldots, \widehat{\boldsymbol{\Gamma}}_{[b]}\right)$*, where MLEs of the sets of equicorrelation parameters* $\{\boldsymbol{U}_{0[j]}\}_{j=1}^{b}$*,* $\{\boldsymbol{U}_{1[j]}\}_{j=1}^{b}$ *and* $\{\boldsymbol{W}_{[j]}\}_{j=1}^{b}$ *are given by*

$$\begin{aligned}
\widehat{\boldsymbol{U}}_{0[j]} &= \frac{1}{nuv} \sum_{r=1}^{n} \sum_{t=1}^{v} \sum_{s=1}^{u} \left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)\left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)', \\
\widehat{\boldsymbol{U}}_{1[j]} &= \frac{1}{nuv\,(u-1)} \sum_{r=1}^{n} \sum_{t=1}^{v} \sum_{s=1}^{u} \sum_{s \neq s^*=1}^{u} \left(\boldsymbol{x}_{r,[j],ts^*} - \overline{\boldsymbol{x}}_{[j],t}\right)\left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)', \\
\widehat{\boldsymbol{W}}_{[j]} &= \frac{1}{nu^2 v\,(v-1)} \sum_{r=1}^{n} \sum_{t=1}^{v} \sum_{t \neq t^*=1}^{v} \sum_{s=1}^{u} \sum_{s^*=1}^{u} \left(\boldsymbol{x}_{r,[j],t^*s^*} - \overline{\boldsymbol{x}}_{[j],t^*}\right)\left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)'.
\end{aligned}$$

The proof of this theorem is given in Appendix A.

5 Classification with locally doubly exchangeable covariance structure

In this section we derive the classification rule for $\mathcal{C}$ classes. Using the same notations as in the introduction, we assume that the vectors $\boldsymbol{x}_{r,11}^{(c)}, \ldots, \boldsymbol{x}_{r,1u}^{(c)}, \ldots, \boldsymbol{x}_{r,v1}^{(c)}, \ldots, \boldsymbol{x}_{r,vu}^{(c)}$ are locally jointly equicorrelated with sets of equicorrelation parameters $\{\boldsymbol{U}_{0[j]}^{(c)}\}_{j=1}^{b}$, $\{\boldsymbol{U}_{1[j]}^{(c)}\}_{j=1}^{b}$ and $\{\boldsymbol{W}_{[j]}^{(c)}\}_{j=1}^{b}$ such that $\sum_{j=1}^{b} p_j = p$, with $\mathrm{E}[\boldsymbol{x}_r^{(c)}] = \boldsymbol{\mu}_{\boldsymbol{x}^{(c)}} = \left(\boldsymbol{\mu}_{\boldsymbol{x}[1]}^{(c)\prime}, \ldots, \boldsymbol{\mu}_{\boldsymbol{x}[b]}^{(c)\prime}\right)'$, where $\boldsymbol{\mu}_{\boldsymbol{x}[j]}^{(c)} = \left(\boldsymbol{\mu}_{\boldsymbol{x}[j],1}^{(c)\prime}, \ldots, \boldsymbol{\mu}_{\boldsymbol{x}[j],v}^{(c)\prime}\right)'$, with $\boldsymbol{\mu}_{\boldsymbol{x}[j],t}^{(c)} = \mathbf{1}_u \otimes \boldsymbol{\mu}_{[j],t}^{(c)}$ and $\boldsymbol{\mu}_{[j],t}^{(c)} \in \Re^{p_j}$ for $t = 1, \ldots, v$, $j = 1, \ldots, b$, and $\mathrm{Cov}[\boldsymbol{x}_r^{(c)}] = \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}} = \mathrm{diag}\left(\boldsymbol{\Gamma}_{[1]}^{(c)}, \boldsymbol{\Gamma}_{[2]}^{(c)}, \ldots, \boldsymbol{\Gamma}_{[b]}^{(c)}\right)$. Let $T^{(c)} = \{\boldsymbol{x}_1^{(c)}, \ldots, \boldsymbol{x}_{n^{(c)}}^{(c)}\}$ be a random sample of size $n^{(c)}$ from the c^{th} class with distribution $N_{puv}\left(\boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}\right)$, for $c = 1, \ldots, \mathcal{C}$. These $\mathcal{C}$ random training samples are independent among each other. We will discuss the linear classifier case, that is, $\boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}} = \boldsymbol{\Gamma}_{\boldsymbol{x}}$, with equicorrelation parameters $\boldsymbol{U}_{0[j]}^{(c)} = \boldsymbol{U}_{0[j]}, \boldsymbol{U}_{1[j]}^{(c)} = \boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}^{(c)} = \boldsymbol{W}_{[j]}$, for $c = 1, \ldots, \mathcal{C}$ and $j = 1, \ldots, b$ in Section 5.1, and the quadratic and modified linear classifier cases in Section 5.2.

5.1 When sets of equicorrelation parameters among classes concur

In this case we assume that all populations have equal sets of equicorrelation parameters $\{\boldsymbol{U}_{0[j]}\}_{j=1}^{b}$, $\{\boldsymbol{U}_{1[j]}\}_{j=1}^{b}$ and $\{\boldsymbol{W}_{[j]}\}_{j=1}^{b}$, thus equal variance-covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$. Therefore, the likelihood function can be written as

$$L\left(\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}}, \ldots, \boldsymbol{\mu}_{\boldsymbol{x}^{(\mathcal{C})}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right) = \frac{\exp\left\{-\frac{1}{2}\sum_{c=1}^{\mathcal{C}}\sum_{r=1}^{n^{(c)}}\left(\boldsymbol{x}_r^{(c)} - \boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1}\left(\boldsymbol{x}_r^{(c)} - \boldsymbol{\mu}_{\boldsymbol{x}^{(c)}}\right)\right\}}{(2\pi)^{\frac{nvup}{2}} \left|\boldsymbol{\Gamma}_{\boldsymbol{x}}\right|^{\frac{n}{2}}},$$

where $n = \sum_{c=1}^{\mathcal{C}} n^{(c)}$, and $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{-1} = \mathrm{diag}\left(\boldsymbol{\Gamma}_{[1]}^{-1}, \boldsymbol{\Gamma}_{[2]}^{-1}, \ldots, \boldsymbol{\Gamma}_{[b]}^{-1}\right)$, and $p_j \times p_j$ blocks $\boldsymbol{\Gamma}_{[j]}^{-1}$ are given in (11). Along the lines presented in Appendix A, it can be proved that the MLEs of the mean vectors $\boldsymbol{\mu}_{[j],t}^{(c)} : j = 1, \ldots, b$, and $t = 1, \ldots, v$, are given by

$$\widehat{\boldsymbol{\mu}}_{[j],t}^{(c)} = \overline{\boldsymbol{x}}_{[j],t}^{(c)} = \frac{1}{n^{(c)}u}\sum_{r=1}^{n^{(c)}}\sum_{s=1}^{u}\boldsymbol{x}_{r,[j],ts}^{(c)}, \quad \text{for } c = 1, 2, \ldots, \mathcal{C} \text{ and } j = 1, 2, \ldots, b,$$

where MLEs of the sets of equicorrelation parameters $\{\boldsymbol{U}_{0[j]}\}_{j=1}^{b}$, $\{\boldsymbol{U}_{1[j]}\}_{j=1}^{b}$ and $\{\boldsymbol{W}_{[j]}\}_{j=1}^{b}$ are given by

$$\begin{aligned}
\widehat{\boldsymbol{U}}_{0[j]} &= \frac{1}{nuv}\sum_{c=1}^{\mathcal{C}}\sum_{r=1}^{n^{(c)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)', \\
\widehat{\boldsymbol{U}}_{1[j]} &= \frac{1}{nuv\,(u-1)}\sum_{c=1}^{\mathcal{C}}\sum_{r=1}^{n^{(c)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s\neq s^{*}=1}^{u}\left(\boldsymbol{x}_{r,[j],ts^{*}}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)', \\
\text{and}\quad \widehat{\boldsymbol{W}}_{[j]} &= \frac{1}{nu^{2}v\,(v-1)}\sum_{c=1}^{\mathcal{C}}\sum_{r=1}^{n^{(c)}}\sum_{t=1}^{v}\sum_{t\neq t^{*}=1}^{v}\sum_{s=1}^{u}\sum_{s^{*}=1}^{u}\left(\boldsymbol{x}_{r,[j],t^{*}s^{*}}^{(c)}-\overline{\boldsymbol{x}}_{[j],t^{*}}^{(c)}\right)\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)',
\end{aligned}$$

where $c = 1, \ldots, \mathcal{C}$. It is to be noted here that the MLEs of $\{\boldsymbol{U}_{0[j]}\}_{j=1}^{b}$, $\{\boldsymbol{U}_{1[j]}\}_{j=1}^{b}$ and $\{\boldsymbol{W}_{[j]}\}_{j=1}^{b}$ are derived for one population case, i.e., for $c = 1$ in Appendix A. The computation of the MLEs $\widehat{\boldsymbol{\mu}}_{[j],t}^{(c)}$ for $j = 1, \ldots, b,\ t = 1, \ldots, v$, and $\widehat{\boldsymbol{U}}_{0[j]}$, $\widehat{\boldsymbol{U}}_{1[j]}$ and $\widehat{\boldsymbol{W}}_{[j]}$ is straightforward, as all of them have closed form solutions.

We now consider the problem of assigning a new individual with $puv-$variate partitioned measurement vector $\boldsymbol{x}_0 = \left(\boldsymbol{x}_{0[1]}', \ldots, \boldsymbol{x}_{0[b]}'\right)'$ to one of the $\mathcal{C}$ classes. The previous set-up leads to a linear discriminant function as follows:
Under the assumptions of equal prior probabilities and equal costs of misclassification, we define the linear score $\ell^{(c)}$ as

$$\begin{aligned}
\widehat{\ell}^{(c)}\ (\boldsymbol{x}_0) &:= \boldsymbol{x}_0'\cdot\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1}\cdot\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(c)}}-\frac{1}{2}\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(c)}}'\cdot\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1}\cdot\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(c)}}, \\
&= \sum_{j=1}^{b}\left[\boldsymbol{x}_{0[j]}'\cdot\widehat{\boldsymbol{\Gamma}}_{[j]}^{-1}\cdot\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)\prime}-\frac{1}{2}\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)\prime}\cdot\widehat{\boldsymbol{\Gamma}}_{[j]}^{-1}\cdot\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)}\right],
\end{aligned}$$

as $\widehat{\ell}^{(c)}\ (\boldsymbol{x}_0) = \sum_{j=1}^{b}\widehat{\ell}_j^{(c)}(\boldsymbol{x}_{0[j]})$, where $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(c)}} = \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[1]}^{(c)\prime}, \ldots, \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[b]}^{(c)\prime}\right)'$, where $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)} = \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j],1}^{(c)\prime}, \ldots, \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j],v}^{(c)\prime}\right)'$, with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j],t}^{(c)} = \mathbf{1}_u \otimes \widehat{\boldsymbol{\mu}}_{[j],t}^{(c)}$ and $\widehat{\boldsymbol{\mu}}_{[j],t}^{(c)} = \overline{\boldsymbol{x}}_{[j],t}^{(c)} = \dfrac{1}{n^{(c)}u}\displaystyle\sum_{r=1}^{n^{(c)}}\sum_{s=1}^{u}\boldsymbol{x}_{r,[j],ts}$, for $t = 1, \ldots, v,\ \ j = 1, \ldots, b$ and $c = 1, \ldots, \mathcal{C}$. The estimate of inverse covariance matrix $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}^{-1}$ is obtained from (12) by replacing $\boldsymbol{U}_{0[j]}$, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ by $\widehat{\boldsymbol{U}}_{0[j]}$, $\widehat{\boldsymbol{U}}_{1[j]}$ and $\widehat{\boldsymbol{W}}_{[j]}$ respectively. Then the resulting classification rule is given as follows:

- Assign a new observation $\boldsymbol{x}_0 = \left(\boldsymbol{x}_{0[1]}', \ldots, \boldsymbol{x}_{0[b]}'\right)'$ to Class Π_i if $y(\boldsymbol{x}_0) = i$, i.e.

$$\ell^{(i)}\ (\boldsymbol{x}_0) = \text{largest of }\left\{\ell^{(c)}\ (\boldsymbol{x}_0) : c = 1, \ldots, \mathcal{C}\right\},\ \text{ for } i = 1, \ldots, \mathcal{C}.$$

This linear rule has been extensively studied by McLachlan (1992).

In the special case of $\mathcal{C} = 2$ populations we obtain a linear classifier, which by partitioning the observed vector $\boldsymbol{x}_0$ becomes:

Assign $\boldsymbol{x}_0$ to class Π_1 if

$$\sum_{j=1}^{b}\left[\boldsymbol{x}_{0[j]}^{\prime}\cdot\widehat{\boldsymbol{\Gamma}}_{[j]}^{-1}\cdot\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)}-\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)\right]>\frac{1}{2}\sum_{j=1}^{b}\left[\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)}+\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)^{\prime}\cdot\widehat{\boldsymbol{\Gamma}}_{[j]}^{-1}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)}-\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)\right],$$

and to the class 2 otherwise.

5.2 When sets of equicorrelation parameters among classes differ

In this case we assume that all populations have different sets of equicorrelation parameters $\{\boldsymbol{U}_{0[j]}^{(c)}\}_{j=1}^{b}$, $\{\boldsymbol{U}_{1[j]}^{(c)}\}_{j=1}^{b}$ and $\{\boldsymbol{W}_{[j]}^{(c)}\}_{j=1}^{b}$, thus different variance-covariance matrices $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{(c)}$. Following the same technique as before in this case also it can be proved that the maximum likelihood estimates of the means $\boldsymbol{\mu}_{[j],t}^{(c)}: t=1,\ldots,v$, and $j=1,\ldots,b$, are

$$\widehat{\boldsymbol{\mu}}_{[j],t}^{(c)}=\overline{\boldsymbol{x}}_{[j],t}^{(c)}=\frac{1}{n^{(c)}u}\sum_{r=1}^{n^{(c)}}\sum_{s=1}^{u}\boldsymbol{x}_{r,[j],ts}^{(c)},\quad\text{for } c=1,2,\ldots,\mathcal{C},$$

where MLEs of the sets of equicorrelation parameters $\{\boldsymbol{U}_{0[j]}^{(c)}\}_{j=1}^{b}$, $\{\boldsymbol{U}_{1[j]}^{(c)}\}_{j=1}^{b}$ and $\{\boldsymbol{W}_{[j]}^{(c)}\}_{j=1}^{b}$ are given by

$$\begin{aligned}
\widehat{\boldsymbol{U}}_{0[j]}^{(c)} &= \frac{1}{n^{(c)}uv}\sum_{r=1}^{n^{(c)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)^{\prime},\\
\widehat{\boldsymbol{U}}_{1[j]}^{(c)} &= \frac{1}{n^{(c)}uv\,(u-1)}\sum_{r=1}^{n^{(c)}}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s\neq s^{*}=1}^{u}\left(\boldsymbol{x}_{r,[j],ts^{*}}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)^{\prime},\\
\text{and}\quad\widehat{\boldsymbol{W}}_{[j]}^{(c)} &= \frac{1}{n^{(c)}u^{2}v\,(v-1)}\sum_{r=1}^{n^{(c)}}\sum_{t=1}^{v}\sum_{t\neq t^{*}=1}^{v}\sum_{s=1}^{u}\sum_{s^{*}=1}^{u}\left(\boldsymbol{x}_{r,[j],t^{*}s^{*}}^{(c)}-\overline{\boldsymbol{x}}_{[j],t^{*}}^{(c)}\right)\left(\boldsymbol{x}_{r,[j],ts}^{(c)}-\overline{\boldsymbol{x}}_{[j],t}^{(c)}\right)^{\prime}.
\end{aligned}$$

Here also the computation of the maximum likelihood estimates $\widehat{\boldsymbol{\mu}}_{[j],t}^{(c)}$, $\widehat{\boldsymbol{U}}_{0[j]}^{(c)}$, $\widehat{\boldsymbol{U}}_{1[j]}^{(c)}$ and $\widehat{\boldsymbol{W}}_{[j]}^{(c)}$ for $j=1,\ldots,b$, $t=1,\ldots,v$, are easy, as they all have closed form solutions.

The optimal classification rule for the $puv-$variate vector of observations, $\boldsymbol{x}$ is quadratic as the covariance matrices are not equal in $\mathcal{C}$ classes. More precisely, under the assumption of equal prior probabilities and equal costs of misclassification we define the estimation of the quadratic score $q^{(c)}$ for the observed vector $\boldsymbol{x}_0=\left(\boldsymbol{x}_{0[1]}^{\prime},\ldots,\boldsymbol{x}_{0[b]}^{\prime}\right)^{\prime}$ as

$$\begin{aligned}
\widehat{q}^{(c)}\;(\boldsymbol{x}_0) &:= -\frac{1}{2}\ln\left|\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}^{(c)}}\right|-\frac{1}{2}\left(\boldsymbol{x}_0-\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(c)}}\right)^{\prime}\cdot\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}^{(c)}}^{-1}\cdot\left(\boldsymbol{x}_0-\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(c)}}\right)\\
&= -\frac{1}{2}\sum_{j=1}^{b}\left[\ln|\hat{\Gamma}_{[j]}^{(c)}|+\left(\boldsymbol{x}_{0[j]}-\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)}\right)^{\prime}\cdot\left(\hat{\Gamma}_{[j]}^{(c)}\right)^{-1}\cdot\left(\boldsymbol{x}_{0[j]}-\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)}\right)\right],
\end{aligned}$$

where $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}^{(c)}} = \left(\widehat{\boldsymbol{\mu}}^{(c)\prime}_{\boldsymbol{x}[1]}, \ldots, \widehat{\boldsymbol{\mu}}^{(c)\prime}_{\boldsymbol{x}[b]}\right)'$, $\widehat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{x}[j]} = \left(\widehat{\boldsymbol{\mu}}^{(c)\prime}_{\boldsymbol{x}[j],1}, \ldots, \widehat{\boldsymbol{\mu}}^{(c)\prime}_{\boldsymbol{x}[j],v}\right)'$, with $\widehat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{x}[j],t} = \mathbf{1}_u \otimes \widehat{\boldsymbol{\mu}}^{(c)}_{[j],t}$ and $\widehat{\boldsymbol{\mu}}^{(c)}_{[j],t} = \overline{\boldsymbol{x}}^{(c)}_{[j],t} = \dfrac{1}{n^{(c)}u} \displaystyle\sum_{r=1}^{n} \sum_{s=1}^{u} \boldsymbol{x}_{r,[j],ts}$, for $t = 1, \ldots, v$, $j = 1, \ldots, b$ and $c = 1, \ldots, \mathcal{C}$. Thus the classification rule is:

Assign a new observation $\boldsymbol{x}_0$ to Class Π_i if $y(\boldsymbol{x}) = i$, i.e.,

$$\widehat{q}^{(i)}\ (\boldsymbol{x}_0) = \text{largest of } \left\{\widehat{q}^{(c)}\ (\boldsymbol{x}_0) : c = 1, \ldots, \mathcal{C}\right\}, \text{ for } i = 1, \ldots, \mathcal{C}.$$

Since the distribution theory associated with this quadratic rule is exceedingly difficult, we like many former authors $\big($Chaudhuri et al., 1991; Park and Kshirsagar, 1994; Leiva and Herrera, 1999$\big)$ propose linear solutions to this problem in this article. Linear rule is desirable than quadratic rule because of its simplicity, and we like to evaluate its theoretical misclassification probabilities. Under the normality assumption, Chaudhuri et al. (1991) found an asymptotic optimal linear classification function $\ell(\boldsymbol{x}) = \boldsymbol{\alpha}'\boldsymbol{x} + \boldsymbol{\beta}$ to classify an unknown individual with response variable $\boldsymbol{x}$ into one of the two populations Π_1 and Π_2. They assumed that the two classes have normal densities with mean vectors $\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}}$ and $\boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}$, and positive definite covariance matrices $\boldsymbol{\Gamma}_{\boldsymbol{x}^{(1)}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{x}^{(2)}}$ respectively and proved that the optimal vector $\boldsymbol{\alpha} = \left(\boldsymbol{\Gamma}_{\boldsymbol{x}^{(1)}} + \boldsymbol{\Gamma}_{\boldsymbol{x}^{(2)}}\right)^{-1} \left(\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}} - \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}\right)$, under certain regularity conditions. Chaudhuri et al. (1991) considered the optimality in the sense of maximizing the Bhattacharyya (1943) distance asymptotically. An equivalent linear solution can be found based on a definition of separation between two density functions as suggested by Park and Kshirsagar in 1994. They modified the distance between two normal classes that was originally proposed by Paranjpe and Gore (1994). Park and Kshirsagar's modified distance $\delta^2\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}\right)$ between two vectors, $\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}}$ and $\boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}$ as follows

$$\delta^2\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}\right) = \boldsymbol{\mu}'_{\boldsymbol{x}} \cdot \left(\frac{\boldsymbol{\Gamma}_{\boldsymbol{x}^{(1)}} + \boldsymbol{\Gamma}_{\boldsymbol{x}^{(2)}}}{2}\right)^{-1} \cdot \boldsymbol{\mu}_{\boldsymbol{x}} \tag{14}$$

where $\boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{\mu}_{\boldsymbol{x}^{(1)}} - \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}$ is a shift vector, $c = 1, 2$. The distance $\delta^2\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}\right)$ is indeed a generalization of the square of the Mahalanobis distance $\delta^2\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right)$ given in Section 2.

Observe that provided appropriate grounds, the distance $\delta^2\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}\right)$ can be related to the misclassification probability defined in (4). By the strict monotonicity of $\Phi(\cdot)$, one can see from (4) that $\delta\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right) = -2\Phi^{-1}\left(\mathcal{E}_{\text{opt}}\right)$, and hence for the normal class-conditional distributions, $\mathcal{E}_{\text{opt}}$ and $\delta\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right)$ provide equivalent information about the classification performance. This in turn implies that $\delta^2\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}\right)$ can also be used as a measure of separation between Π_1 and Π_2. In particular, Leiva and Herrera (1999) have studied the asymptotic properties of this separation measere and showed that it is related to Hellinger's similarity measure (Rao and Varadarajan, 1963) and Matushita's closeness measure (Matushita, 1966).

Leiva and Herrera (1999) also obtained the optimal discriminant solution by maximizing the separation measure (14) on a transformed variable $\boldsymbol{z} = \boldsymbol{\alpha}'\boldsymbol{x} + \boldsymbol{\beta}$, where the maximization is done with respect to $\boldsymbol{\alpha}$. They showed that $\boldsymbol{z}$ can be chosen as

$$\boldsymbol{z} = T\left(\boldsymbol{x}\right) = \left(\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}} - \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}\right)' \left(\frac{\boldsymbol{\Gamma}_{\boldsymbol{x}^{(1)}} + \boldsymbol{\Gamma}_{\boldsymbol{x}^{(2)}}}{2}\right)^{-1} \boldsymbol{x},$$

and thus, $\boldsymbol{\alpha}$ can be chosen as

$$\boldsymbol{\alpha}' = \left(\boldsymbol{\mu}_{\boldsymbol{x}^{(1)}} - \boldsymbol{\mu}_{\boldsymbol{x}^{(2)}}\right)' \left(\frac{\boldsymbol{\Gamma}_{\boldsymbol{x}^{(1)}} + \boldsymbol{\Gamma}_{\boldsymbol{x}^{(2)}}}{2}\right)^{-1}.$$

Under the assumptions of equal prior probabilities, equal costs of misclassification and given the locally jointly equiocorrelated structure of $\boldsymbol{\Gamma}_{\boldsymbol{x}^{(c)}}$ with the same block-sized components for $c = 1$ and 2, the M-linear (modified linear) sample classification rule with the corresponding partition of the observed vector, $\boldsymbol{x}_0 = \left(\boldsymbol{x}_{0[1]}, \ldots, \boldsymbol{x}_{0[b]}\right)'$ is defined as follows:

Assign $\boldsymbol{x}_0$ to class Π_1 if

$$\sum_{j=1}^{b} \left[\boldsymbol{x}_{0[j]} \cdot \left(\frac{\widehat{\boldsymbol{\Gamma}}_{[j]}^{(1)} + \widehat{\boldsymbol{\Gamma}}_{[j]}^{(2)}}{2}\right)^{-1} \cdot \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)\right]$$

$$> \frac{1}{2} \sum_{j=1}^{b} \left[\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right) \cdot \left(\frac{\widehat{\boldsymbol{\Gamma}}_{[j]}^{(1)} + \widehat{\boldsymbol{\Gamma}}_{[j]}^{(2)}}{2}\right)^{-1} \cdot \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)\right]$$

and to class Π_2 otherwise. Here,

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)} = \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j],1}^{(c)\prime}, \ldots, \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j],v}^{(c)\prime}\right)', \quad \text{with} \quad \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j],t}^{(c)} = \mathbf{1}_u \otimes \widehat{\boldsymbol{\mu}}_{[j],t}^{(c)} \quad \text{and} \quad \widehat{\boldsymbol{\mu}}_{[j],t}^{(c)} \in \Re^{p_j},$$

for $t = 1, \ldots, v$, $j = 1, \ldots, b$, and $c = 1, 2$, and $\widehat{\boldsymbol{\Gamma}}_{[j]}^{(c)}$ is given in (9).

Observe that $\dfrac{\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}^{(1)}} + \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}^{(2)}}}{2}$ has also the form (8) as each of $\widehat{\boldsymbol{\Gamma}}_{[j]}^{(c)}$, $j = 1, \ldots, b$, $c = 1, 2$ is invariant with respect to addition, and therefore its inverse can be obtained using (12). The difference between the linear and M-linear procedures is that the former uses a weighted average of sample covariance matrices while the later uses an unweighted average. Thus, the two procedures will be exactly same when sample sizes are equal. The generalization of this rule for $\mathcal{C} > 2$ is simple.

6 Class of asymptotically equivalent block structure approximations of $\boldsymbol{\Gamma}_{\boldsymbol{x}}$

For a reasonable choice of the covariance structure approximation in the classification framework, we need to investigate the effect of the block size, $p_j uv$, $j = 1, \ldots, b$, on the probability of misclassification.

The quantity $\mathcal{E}$ for $\mathcal{C} = 2$ is defined in (3). Recall from (7) that under constraints $\max\limits_{j=1,\ldots,b_\lambda} p_j < n/uv$, imposed on the the block size in Algorithm 1 of Pavlenko et al. (2012), each choice of the penalty parameter λ in (gLasso section), provides an approximate block structure of $\Gamma_{\boldsymbol{x}}$, so that for a general choice of u and v the resulting estimator of the class covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is given by $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x},\lambda} = \operatorname{diag}\left[\widehat{\boldsymbol{\Gamma}}_{[1],\lambda}, \ldots, \widehat{\boldsymbol{\Gamma}}_{[b],\lambda}\right]$, where $\widehat{\boldsymbol{\Gamma}}_{[j],\lambda}$ for $j = 1, \ldots, b$ are the ML estimators of the corresponding block diagonal entry. Hence, to optimize the selection of the block structure we need to minimize $\mathcal{E}$ with respect to λ. However, the solution of Algorithm 2 in Pavlenko et al. (2012) is not unique due to the sensitivity of the Cuthill-McKee reordering transform to the choice of the initial node; see Algorithm 2 in Pavlenko et al. (2012) for details. To avoid this problem, we instead express $\mathcal{E}$ in terms of the block size, $p_j uv$ for a given λ, and show that the effect of $p_j uv$ on the classification accuracy is negligible in growing dimensions asymptotics. Since λ is assumed to be fixed for a specific block structure, we drop this index in the rest of the article.

First, by using the block partitioned mean vector $\boldsymbol{\mu}_{\boldsymbol{x}}$ (defined in Section 3.2), which is constant over sites (u), and by the conditional independence of $\boldsymbol{x}_{r,[j]}$'s for $j = 1, \ldots, b$, given class variable $y(\boldsymbol{x})$, we find the empirical counterpart of the theoretical linear score $\ell^{(c)}(\boldsymbol{x})$ in (1) as

$$\widehat{\ell}^{(c)}(\boldsymbol{x}_0) = \sum_{j=1}^{b} \widehat{\ell}_j^{(c)}(\boldsymbol{x}_{0[j]}) = \sum_{j=1}^{b} \left[\boldsymbol{x}'_{0[j]} \cdot \widehat{\boldsymbol{\Gamma}}_{[j]}^{(c)^{-1}} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)} - \frac{1}{2} \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)\prime} \cdot \widehat{\boldsymbol{\Gamma}}_{[j]}^{(c)^{-1}} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)}\right] + \ln \pi_c,$$

where $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(c)}$ are ML estimators of $\boldsymbol{\mu}_{\boldsymbol{x}[j]}^{(c)}$ for all $c = 1, \ldots, \mathcal{C}$ as specified in Section 4, and $\widehat{\boldsymbol{\Gamma}}_{[j]}^{(c)^{-1}}$ for the class c is obtained from (11) by plugging-in the ML estimators of equicorrelation parameters $\boldsymbol{U}_{0[j]}$, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ that are also obtained in Section 4.

Further, as in the previous consideration, without loss of generality we focus on the analysis of the misclassification error for the case of $\mathcal{C} = 2$ and assume that $\pi_1 = \pi_2 = 1/2$ in (2). The resulting estimated classifier is then expressed as

$$\widehat{\ell}(\boldsymbol{x}_0) = \sum_{j=1}^{b} \widehat{\ell}_j(\boldsymbol{x}_{0[j]}) = \sum_{j=1}^{b} \left[\left(\boldsymbol{x}_{0[j]} - \frac{1}{2}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)\right)' \cdot \widehat{\boldsymbol{\Gamma}}_{[j]}^{-1} \cdot \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)\right]. \qquad (15)$$

To find the explicit expressions for the misclassification probability of (15) in terms of $p_j uv$, we assume that $\boldsymbol{x}_0 = \left(\boldsymbol{x}'_{0[1]}, \ldots, \boldsymbol{x}'_{0[b]}\right)' \in \Pi_1$ and recall that $p_j uv < n$ for all $j = 1, \ldots, b$. Then the conditional distribution of the jth block linear score given $\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}, \widehat{\boldsymbol{\Gamma}}_{[j]}\right)$ is normal,

$$\widehat{\ell}_j(\boldsymbol{x}_{0[j]}) | \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}, \widehat{\boldsymbol{\Gamma}}_{[j]} \sim N_{p_j vu}\left(E_{[j]}, V_{[j]}\right),$$

where

$$E_{[j]} = \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)' \cdot \widehat{\boldsymbol{\Gamma}}_{[j]}^{-1} \cdot \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \boldsymbol{\mu}_{\boldsymbol{x}[j]}^{(1)}\right) - \frac{1}{2}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)' \cdot \widehat{\boldsymbol{\Gamma}}_{[j]}^{-1} \cdot \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right) \qquad (16)$$

and

$$V_{[j]} = \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right)' \cdot \widehat{\boldsymbol{\Gamma}}_{[j]}^{-1} \cdot \boldsymbol{\Gamma}_{[j]}^{-1} \cdot \widehat{\boldsymbol{\Gamma}}_{[j]}^{-1} \cdot \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(1)} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}^{(2)}\right). \qquad (17)$$

In asymptotic study of ultra high dimensional data with block dependence structure, we turn to the growing dimensions asymptotic framework $\big($see e.g. Pavlenko (2003), Pavlenko et al. (2012)$\big)$, that allows $b \to \infty$ together with n so that $b/n \to \beta$ as $n \to \infty$, where $\beta > 0$. The advantage of this approach is that we can find a closed form expression for $\mathcal{E}$ assuming that the block size $p_j uv$ is fixed for each given λ, as shown below.

For fixed $p_j uv$, as $b \to \infty$ the classifier (15) can be considered as the sum of growing number of independent random variables $\widehat{\ell}_j(\boldsymbol{x}_{0[j]})$. Using the results of Pavlenko (2003) we further see that the distribution of $\widehat{\ell}(\boldsymbol{x}_0)$ is asymptotically normal. Due to normality assumption on the class conditional distributions, this result is a special case of Pavlenko (2003), where the more general consideration of the distributional properties of the supervised classifier under the assumption of block-wise dependence structure was given. In particular, in Pavlenko (2003), the asymptotic distributional properties of the classifier were studied under more relaxed conditions, which impose a set of regularity conditions on the class conditional distributions and constrained convergence rates of parameter estimators.

The asymptotic normality of $\widehat{\ell}(\boldsymbol{x}_0)$ suggest that the misclassification probability, $\mathcal{E}_{opt}$ can be approximated by

$$\Phi\left(-\frac{\sum_{j=1}^{b} \mathrm{E}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})\big|\boldsymbol{x}_0 \in \Pi_1\right]}{\sum_{j=1}^{b} \mathrm{Var}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})|\boldsymbol{x}_0 \in \Pi_1\right]}\right), \tag{18}$$

where $\mathrm{E}_T[\cdot\big|\boldsymbol{x}_0 \in \Pi_1]$ and $\mathrm{Var}_T[\cdot\big|\boldsymbol{x}_0 \in \Pi_1]$ denote the expectation and variance with respect to the training data T respectively, and $\boldsymbol{x}_0$ is independent of $\widehat{\boldsymbol{\mu}}^{(1)}_{\boldsymbol{x}[j]}$, $\widehat{\boldsymbol{\mu}}^{(2)}_{\boldsymbol{x}[j]}$ and $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{x}}$. Further, by applying the results of Davis (1987) to the moments of $\widehat{\ell}_j(\boldsymbol{x}_{0[j]})$, as expressed in (16) and (17) we obtain

$$\mathrm{E}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})\big|\boldsymbol{x}_0 \in \Pi_1\right] = \frac{1}{2}\frac{m}{\nu_j}\left[\delta_j^2 + p_j uv\left(\frac{1}{n^{(1)}} - \frac{1}{n^{(2)}}\right)\right] \tag{19}$$

and

$$\begin{aligned}
&\left(\frac{m}{\nu_j}\right)^2 (\nu_j - 2)(\nu_j + 1)\mathrm{Var}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})\big|\boldsymbol{x} \in \Pi_1\right] \\
&= \frac{1}{2}\delta_j^4(\nu_j + 1) + \delta_j^2(\nu_j + 1)\left[\nu_j\left(1 + \frac{1}{n^{(2)}}\right) + \left(\frac{1}{n^{(1)}} - \frac{1}{n^{(2)}}\right)\right] \\
&+p_j uv(\nu_j + p_j uv)\left[\nu_j\left(\frac{1}{n^{(1)}} - \frac{1}{n^{(2)}}\right) + \frac{1}{2}(\nu_j + 1)\left(\frac{1}{(n^{(1)})^2} + \frac{1}{(n^{(2)})^2}\right) - \frac{1}{n^{(1)}n^{(2)}}\right],
\end{aligned} \tag{20}$$

where $m = n^{(1)} + n^{(2)} - 2$, $\nu_j = m - p_j uv - 1$. The square of the Mahalanobis distance δ_j for the jth block can be expressed as

$$\begin{aligned}
\delta_j^2 &= \boldsymbol{\mu}'_{\boldsymbol{x}[j]} \cdot \boldsymbol{\Gamma}^{-1}_{[j]} \cdot \boldsymbol{\mu}_{\boldsymbol{x}[j]} \\
&= \boldsymbol{\mu}'_{\boldsymbol{x}[j]} \cdot \left(\boldsymbol{I}_{vu} \otimes \boldsymbol{\Delta}^{-1}_{1[j]} + \boldsymbol{I}_v \otimes \boldsymbol{J}_u \otimes \frac{1}{u}\left(\boldsymbol{\Delta}^{-1}_{2[j]} - \boldsymbol{\Delta}^{-1}_{1[j]}\right) + \boldsymbol{J}_{vu} \otimes \frac{1}{vu}\left(\boldsymbol{\Delta}^{-1}_{3[j]} - \boldsymbol{\Delta}^{-1}_{2[j]}\right)\right) \cdot \boldsymbol{\mu}_{\boldsymbol{x}[j]},
\end{aligned} \tag{21}$$

where $\boldsymbol{\mu}_{\boldsymbol{x}[j]} = \boldsymbol{\mu}^{(1)}_{\boldsymbol{x}[j]} - \boldsymbol{\mu}^{(2)}_{\boldsymbol{x}[j]}$ is the shift vector for the j^{th} block, and $\boldsymbol{\Delta}_{1[j]}$, $\boldsymbol{\Delta}_{2[j]}$ and $\boldsymbol{\Delta}_{3[j]}$ are expressed in (10a), (10b) and (10c) respectively in terms of equicorrelation parameters $\boldsymbol{U}_{0[j]}$, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ for $j = 1, \ldots, b$.

Now, the main result of this section is stated in the following proposition.

Proposition 1. *Assume that maximal possible growth rate of the block size identified by Algorithm 1 in Pavlenko et al. (2012) is constrained by $p_j < n/uv$ for all $j = 1, \ldots, b$, and let $\overline{p} = \frac{1}{b}\sum_{j=1}^{b} p_j$ be the average block size where averaging is performed for a specific λ. Then, for a third-order data with locally doubly exchangeable covariance structure, the asymptotic effect of $\overline{p}uv$ on $\mathcal{E}_{opt}$ is of the order $\mathcal{O}(\frac{\overline{p}vu}{n^2})$.*

The proof this proposition is given in Appendix B.

This proposition says, in words, that by considering a path of solutions in the Algorithm 1 in Pavlenko et al. (2012) for a range of penalization parameter such that $\lambda = \mathcal{O}(\sqrt{\frac{\ln(p)}{n}})$, and by constraining the block size to $p_j < n/uv$, we derive a structure approximation that leads to the classifier with asymptotically equivalent performance for each value of λ. Hence, asymptotically the average block size, $\bar{p}uv$ does not effect the classification accuracy. This means that the block structures having such property form a *class of asymptotically equivalent structure approximations for UHDHOD*, nonetheless for finite sample the effect could be investigated for each particular choice of p_j, u, v and n. Numerical studies of these effects are presented in the next section.

7 A simulation study

We conduct a simulation study to see the the effect of the block size p_j and the effect of number of repeated measurements v over time on the performance accuracy of our new linear classifier. To see the effect of the block size on the performance accuracy we consider three partitioning scenarios of the $(p \times p)$ dimensional variance-covariance matrix, where $p = p_1 + \cdots + p_b$. In our simulation study we keep the block sizes (p_j) to be constant p_0 (say) in each scenario. To vary the block sizes for a fixed number of variables $p = 12$ we consider three choices of block sizes $p_0 = 4$, $p_0 = 3$ and $p_0 = 2$ resulting in $b = 3, 4$ and 6 blocks respectively. For each p_0 we set the values of v, the number of repeated measurements over time as 3 and 5, and the number of sites as u=2 for all scenarios. We assume that the two populations have the same locally jointly equicorrelated covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ as defined in (8), where the diagonal blocks $\boldsymbol{\Gamma}_{[j]}$, for $j = 1, \ldots, b$ are jointly equicorrelated covariance matrix with equicorrelation parameters $\boldsymbol{U}_{0[j]}$, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$. Further, for each choice of p_0, v and u we assume the same jointly equicorrelated covariance matrix with equicorrelation parameters $\boldsymbol{U}_{0[j]}$, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$. Our choice for $p_0 = 4$ and $b = 3$ are given below:

$$\boldsymbol{U}_{0[1]} = \begin{bmatrix} 2 & 1 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 2 & 3 & 5 & 4 \\ 1 & 2 & 4 & 5 \end{bmatrix}, \quad \boldsymbol{U}_{0[2]} = \begin{bmatrix} 7 & 1 & 2 & 1 \\ 1 & 5 & 3 & 2 \\ 2 & 3 & 6 & 5 \\ 1 & 2 & 5 & 5 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{U}_{0[3]} = \begin{bmatrix} 5 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 6 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix},$$

for three blocks of sizes (4×4) for $j = 1, 2$ and 3. For all three blocks The covariance matrices $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ are chosen as follows

$$\boldsymbol{U}_{1[j]} = \begin{bmatrix} 0.40 & 0.11 & 0.40 & 0.30 \\ 0.11 & 0.60 & 0.15 & 0.30 \\ 0.40 & 0.15 & 0.60 & 0.20 \\ 0.30 & 0.30 & 0.20 & 0.10 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{W}_{[j]} = \begin{bmatrix} 0.30 & 0.20 & 0.10 & 0.10 \\ 0.20 & 0.30 & 0.10 & 0.07 \\ 0.10 & 0.10 & 0.30 & 0.20 \\ 0.10 & 0.07 & 0.20 & 0.20 \end{bmatrix}.$$

The equicorrelation parameters $\boldsymbol{U}_{0[j]}$, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ for $b = 4$ are chosen as the main majors from the above equicorrelation parameters for $b = 3$. This strategy provides only three sets of $\boldsymbol{U}_{0[j]}$s for $b = 4$ so the last $\boldsymbol{U}_{0[4]}$ is chosen as

$$\boldsymbol{U}_{0[4]} = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix}$$

Using the similar technique we set the (2×2) equicorrelation parameters $\boldsymbol{U}_{0[j]}$, $\boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ for $b = 6$. We choose the mean vectors $\boldsymbol{\mu}_{\boldsymbol{x}}^{(c)}$ for class $c = 1, 2$ as defined in Section 3.2 with $\boldsymbol{\mu}_{[j],1}^{(1)} = (2, 1, 1, 0.5)'$, $\boldsymbol{\mu}_{[j],2}^{(1)} = (2.5, 1.5, 1.5, 1)'$ and $\boldsymbol{\mu}_{[j],3}^{(1)} = (3, 1.5, 1.5, 1)'$ for all $j = 1, 2$ and 3. And, $\boldsymbol{\mu}_{[j],1}^{(2)} = (0, 1, 0, 0.5)'$, $\boldsymbol{\mu}_{[j],2}^{(2)} = (2.0, 1.5, 1.0, 1.0)'$ and $\boldsymbol{\mu}_{[j],3}^{(2)} = (2.0, 1.5, 0.5, 0.5)'$ for all $j = 1, 2$ and 3.

Training samples of sizes $(\mathrm{n}^{(1)}, \mathrm{n}^{(2)})$ =(3, 3), (4, 4), (5, 5) (very small), (8, 8), (12, 12) (small), (15, 15), (25, 25) (moderate) and $(50, 50), (100, 100), (500, 500)$ (large), and a pair of test samples (2000, 2000) are generated from the puv-variate normal populations $N_{puv}(\boldsymbol{\mu}_{\boldsymbol{x}}^{(c)}, \boldsymbol{\Gamma}_{\boldsymbol{x}})$, $c = 1, 2$, where $\boldsymbol{\mu}_{\boldsymbol{x}}^{(c)}$ 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 $(\boldsymbol{U}_{0[j]}, \boldsymbol{U}_{1[j]}, \boldsymbol{W}_{[j]})$ for all partition scenarios using the ML method as discussed in Section 4.

Table 1 shows the misclassification error rates (MERs) of the test set of (2000, 2000). We see that smaller the block size p_0, the better the classification performance, i.e., lower the MERs. This pattern is observed for all pairs of training samples and within that for each $v = 3$ and 5, except for a small bump at sample sizes (8, 8) and $v = 3$. These results indicate that for smaller sample sizes, the classifiers with smaller blocks outperform those with larger blocks over all sets of u and v; this result is indeed expected as with smaller p_0 there are less parameters to be estimated. We also see that MER always decreases with the increase of v with the exception of small bumps sometimes. This is due to the fact that more repeated measurements means more information and it leads to less MERs. And, this is true for all pairs of sample sizes. We also see that with the increase of training sample sizes MER decreases as we get reliable parameter estimates; more samples, more information, thus much reliable estimates of the unknown parameters.

This result could also be interpreted in terms of bandable covariance structure; the wider the band, the more parameters to be estimated, this in turn yields more pronounced effect on the classification performance. Observe however that our empirical results indicate that the performance accuracy is better when approximating the band structure by smaller blocks, i.e., when the blocks are inside the band, than approximating by larger blocks, i.e., when blocks cover the band. Note that in the later case the approximation involves spurious non-zero covariances naturally affecting the misclassification.

For completeness reason we also intend to see whether the performance pattern of our new classifiers remain the same for the large sample consideration $((50, 50), (100, 100), (500, 500))$; see Table 2. It turns out that the pattern of Table 1 is preserved in Table 2. However, the effect of block size seems to be less pronounced with the increase of the sample size, i.e., the resulting MERs become very close to each other, demonstrating more stable behavior for all the block sizes; this supports our limit results of Proposition 1 stating that the choice of the block size at the model selection stage is asymptotically small.

Table 2: MERs (%) for the simulated data for different training sample sizes and with different block sizes

$\mathrm{n}^{(1)}, \mathrm{n}^{(2)}$	(50,50)		(100,100)		(500,500)	
$v \rightarrow$ Block Size (p_0) $\downarrow$	3	5	3	5	3	5
4	6.200	4.100	5.250	3.475	5.175	2.625
3	4.625	2.675	3.400	2.350	3.150	1.800
2	3.325	2.175	2.350	1.900	2.250	1.650

8 Conclusions and scope for the future

In this article we explore the block-wise sparsity, which leads to the additive structure of the resulting linear and quadratic classifiers, which in turn essentially simplifies covariance estimation in ultra high-dimensional settings. The technique of the covariance structure learning was extended to the higher order data with specific block-wise covariance structure. In particular, our approach allows for modeling higher order data with varying sets of equicorrelation parameters, reflecting many sites and time points. This type of simultaneous covariance structure learning and parameter estimation is of great advantage for supervised classification. One possibility to extend our approach is to turn to other graphical techniques for covariance model selection, such as l_1-penalized log determinant Bregman divergence, where the structure is specified by the graph of an associated Gaussian Markov random field, as described in Ravikumar et al. (2011). Another extension may be possible by using Bayesian perspective, namely turn to Bayesian predictive classification where priors are

Table 1: MERs (%) for the simulated data for different training sample sizes and with different block sizes

$n^{(1)}, n^{(2)}$	(3,3)		(4,4)		(5,5)		(8,8)		(12,12)		(15,15)		(25,25)	
$v \rightarrow$ Block Size (p_0) ↓	3	5	3	5	3	5	3	5	3	5	3	5	3	5
4	32.425	30.850	26.900	17.550	18.500	9.400	15.800	6.400	10.250	7.375	9.300	9.100	7.750	5.875
3	27.100	19.125	15.050	12.875	14.275	8.525	7.175	7.100	6.900	6.700	7.025	5.675	5.525	4.900
2	21.325	13.375	8.600	7.125	10.450	5.950	8.750	4.500	5.100	3.825	3.800	4.050	3.925	2.175

assigned to the set of equicorrelation parameters in the model (9); see Corander et al. (2012) for details. Observe however that the block diagonal covariance structure in Corander et al. (2012) was pre-specified; it would be interesting to impose proper priors on the covariance structure as well.

Acknowledgements

Tatjana Pavlenko gratefully acknowledges financial support in part from the Swedish Research Council under grant 421-2008-1966 and Anuradha Roy gratefully acknowledges financial s upport in part from the Royal Swedish Academy of Sciences, Göran Gustafsson Stiftelse, grant 2012-2013.

A Maximum likelihood estimation of $\boldsymbol{\mu}_{[j]t}$, $\boldsymbol{U}_{0[j]}, \boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$, $j = 1, \ldots, b,\ t = 1, \ldots, v$

Proof of Theorem 1: Using the same notations as in Section 4, the likelihood function $L = L\left(\boldsymbol{\mu}_{\boldsymbol{x}}; \boldsymbol{\Gamma}_{\boldsymbol{x}}\right) = L\left(\boldsymbol{\mu}_{[j]1}, \ldots, \boldsymbol{\mu}_{[j]v}, \boldsymbol{U}_{0[j]}, \boldsymbol{U}_{1[j]}, \boldsymbol{W}_{[j]}\right)$ for $j = 1, \ldots, b$ can be written as

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

or equivalently,

$$L\left(\boldsymbol{\mu}_{\boldsymbol{x}_*}, \boldsymbol{\Gamma}_{\boldsymbol{x}_*}\right) = \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_* - \boldsymbol{\mu}_{\boldsymbol{x}_*}\right)' \boldsymbol{\Gamma}_{\boldsymbol{x}_*}^{-1}\left(\boldsymbol{x}_* - \boldsymbol{\mu}_{\boldsymbol{x}_*}\right)\right\}}{(2\pi)^{\overline{\frac{npuv}{2}}} \left|\boldsymbol{\Gamma}_{\boldsymbol{x}_*}\right|^{\overline{\frac{1}{2}}}},$$

where

$$\boldsymbol{x}_{[j]*} = \left(\boldsymbol{x}'_{1,[j]}, \ldots, \boldsymbol{x}'_{n,[j]}\right)',$$

$$\boldsymbol{\mu}'_{\boldsymbol{x}_{[j]*}} = \mathbf{1}'_n \otimes \boldsymbol{\mu}'_{\boldsymbol{x}[j]} = \mathbf{1}'_n \otimes \left(\mathbf{1}'_u \otimes \boldsymbol{\mu}'_{[j],1}, \ldots, \mathbf{1}'_u \otimes \boldsymbol{\mu}'_{[j],v}\right),$$

and

$$\boldsymbol{\Gamma}_{[j]*} = \boldsymbol{I}_n \otimes \boldsymbol{\Gamma}_{[j]},$$

which is a block diagonal matrix with n identical diagonal blocks $\boldsymbol{\Gamma}_{\boldsymbol{x}}$. Thus, using the results (12) and (13) the log likelihood function can be written as

$$\begin{aligned}
\ln(L\left(\boldsymbol{\mu}_{\boldsymbol{x}_*}, \boldsymbol{\Gamma}_{\boldsymbol{x}*}\right)) &= -\frac{npuv}{2}\ln\left(2\pi\right) - \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), \\
&= -\frac{npuv}{2}\ln\left(2\pi\right) - \frac{n}{2}\ln(\left|\boldsymbol{\Gamma}_{[1]}\right|\left|\boldsymbol{\Gamma}_{[2]}\right|\ldots\left|\boldsymbol{\Gamma}_{[b]}\right|) \\
&\quad -\frac{1}{2}\sum_{r=1}^{n}\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right)'\operatorname{diag}\left(\boldsymbol{\Gamma}_{[1]}^{-1}, \boldsymbol{\Gamma}_{[2]}^{-1}, \ldots, \boldsymbol{\Gamma}_{[b]}^{-1}\right)\left(\boldsymbol{x}_r - \boldsymbol{\mu}_{\boldsymbol{x}}\right), \\
&= -\frac{npuv}{2}\ln\left(2\pi\right) - \frac{n}{2}\ln(\left|\boldsymbol{\Gamma}_{[1]}\right|\left|\boldsymbol{\Gamma}_{[2]}\right|\ldots\left|\boldsymbol{\Gamma}_{[b]}\right|) \\
&\quad -\frac{1}{2}\sum_{r=1}^{n}\sum_{j=1}^{b}\left(\boldsymbol{x}_{r,[j]} - \boldsymbol{\mu}_{\boldsymbol{x}[j]}\right)'\boldsymbol{\Gamma}_{[j]}^{-1}\left(\boldsymbol{x}_{r,[j]} - \boldsymbol{\mu}_{\boldsymbol{x}[j]}\right), \\
&= -\frac{npuv}{2}\ln\left(2\pi\right) - \frac{n}{2}\sum_{j=1}^{b}\ln\left|\boldsymbol{\Gamma}_{[j]}\right| \\
&\quad -\frac{1}{2}\sum_{j=1}^{b}\sum_{r=1}^{n}\left(\boldsymbol{x}_{r,[j]} - \boldsymbol{\mu}_{\boldsymbol{x}[j]}\right)'\boldsymbol{\Gamma}_{[j]}^{-1}\left(\boldsymbol{x}_{r,[j]} - \boldsymbol{\mu}_{\boldsymbol{x}[j]}\right), \\
&= -\frac{npuv}{2}\ln\left(2\pi\right) - \frac{1}{2}\sum_{j=1}^{b}\ln\left|\boldsymbol{\Gamma}_{[j]_*}\right| \\
&\quad -\frac{1}{2}\sum_{j=1}^{b}\left(\boldsymbol{x}_{[j]_*} - \boldsymbol{\mu}_{\boldsymbol{x}[j]_*}\right)'\boldsymbol{\Gamma}_{[j]_*}^{-1}\left(\boldsymbol{x}_{[j]_*} - \boldsymbol{\mu}_{\boldsymbol{x}[j]_*}\right), \\
&= -\frac{npuv}{2}\ln\left(2\pi\right) - \frac{1}{2}\sum_{j=1}^{b}\ln\left|\boldsymbol{\Gamma}_{[j]_*}\right| \\
&\quad -\frac{1}{2}\sum_{j=1}^{b}\left(\boldsymbol{x}_{[j]_*} - \boldsymbol{\mu}_{\boldsymbol{x}[j]_*}\right)'\boldsymbol{\Gamma}_{[j]_*}^{-1}\left(\boldsymbol{x}_{[j]_*} - \boldsymbol{\mu}_{\boldsymbol{x}[j]_*}\right), \qquad \text{(A1)}
\end{aligned}$$

where the matrix $\boldsymbol{\Gamma}_{[j]}^{-1}$ is given in (11). Using the centered vectors $\overset{\bullet}{\boldsymbol{x}}_{r,[j],ts} = \boldsymbol{x}_{r,[j],ts} - \boldsymbol{\mu}_{[j],t}$, for $r = 1, \ldots, n$, $s = 1, \ldots, u$, and $t = 1, \ldots, v$, the sum of quadratic forms for the j^{th} block

$$\begin{aligned}
Q\left(\boldsymbol{x}_{[j]_*}\right) &= \sum_{r=1}^{n}\left(\boldsymbol{x}_{r,[j]} - \boldsymbol{\mu}_{\boldsymbol{x}[j]}\right)'\boldsymbol{\Gamma}_{[j]}^{-1}\left(\boldsymbol{x}_{r,[j]} - \boldsymbol{\mu}_{\boldsymbol{x}[j]}\right) \\
&= \left(\boldsymbol{x}_{[j]_*}{-}\boldsymbol{\mu}_{\boldsymbol{x}[j]_*}\right)'\boldsymbol{\Gamma}_{[j]_*}^{-1}\left(\boldsymbol{x}_{[j]_*} - \boldsymbol{\mu}_{\boldsymbol{x}[j]_*}\right).
\end{aligned}$$

Now, the log likelihood function (A1) can be written as

$$\ln(L\left(\boldsymbol{\mu}_{\boldsymbol{x}_*}, \boldsymbol{\Gamma}_{\boldsymbol{x}*}\right)) = -\frac{npuv}{2}\ln\left(2\pi\right) - \frac{1}{2}\sum_{j=1}^{b}\left[\ln\left|\boldsymbol{\Gamma}_{[j]_*}\right| + Q\left(\boldsymbol{x}_{[j]_*}\right)\right]. \qquad \text{(A2)}$$

Thus, we see from (A2) that the mean vector and the variance-covariance matrix $\boldsymbol{\mu}_{\boldsymbol{x}[j]*}$ and $\boldsymbol{\Gamma}_{[j]_*}$ for each of the j^{th} blocks, $j = 1, \ldots, b$ can be obtained separately. We first minimize $Q\left(\boldsymbol{x}_{[j]_*}\right)$ with respect to $\boldsymbol{\mu}_{\boldsymbol{x}[j]}$ for a fixed covariance matrix $\boldsymbol{\Gamma}_{[j]_*}$. By substituting this estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}[j]}$ into (A2), we minimize the log-likelihood equation with respect to $\boldsymbol{\Gamma}_{[j]_*}$ to get the MLE of $\boldsymbol{\Gamma}_{[j]_*}$.

Following the same techniques as Roy and Leiva (2007) we have the MLEs of $\boldsymbol{\mu}'_{\boldsymbol{x}[j]}$ for the j^{th} block as follows:

$$\widehat{\boldsymbol{\mu}}'_{\boldsymbol{x}[j]} = \left(\mathbf{1}'_u \otimes \widehat{\boldsymbol{\mu}}'_{[j],1}, \ldots, \mathbf{1}'_u \otimes \widehat{\boldsymbol{\mu}}'_{[j],v}\right),$$

where $\widehat{\boldsymbol{\mu}}_{[j],t} = \overline{\boldsymbol{x}}_{[j],t} = \dfrac{1}{nu}\displaystyle\sum_{r=1}^{n}\sum_{s=1}^{u}\boldsymbol{x}_{r,[j],ts}$, for $t = 1, \ldots, v$, and the MLEs of $\boldsymbol{U}_{0[j]}, \boldsymbol{U}_{1[j]}$ and $\boldsymbol{W}_{[j]}$ for the j^{th} block as follows:

$$\begin{aligned}
\widehat{\boldsymbol{U}}_{0[j]} &= \frac{1}{nuv}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)\left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)', \\
\widehat{\boldsymbol{U}}_{1[j]} &= \frac{1}{nuv\,(u-1)}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{s=1}^{u}\sum_{s\neq s^*=1}^{u}\left(\boldsymbol{x}_{r,[j],ts^*} - \overline{\boldsymbol{x}}_{[j],t}\right)\left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)', \\
\text{and}\quad \widehat{\boldsymbol{W}}_{[j]} &= \frac{1}{nu^2v\,(v-1)}\sum_{r=1}^{n}\sum_{t=1}^{v}\sum_{t\neq t^*=1}^{v}\sum_{s=1}^{u}\sum_{s^*=1}^{u}\left(\boldsymbol{x}_{r,[j],t^*s^*} - \overline{\boldsymbol{x}}_{[j],t^*}\right)\left(\boldsymbol{x}_{r,[j],ts} - \overline{\boldsymbol{x}}_{[j],t}\right)',
\end{aligned}$$

for $j = 1, \ldots, b$.

B Proof of Proposition 1

To simplify asymptotic consideration, we set $p_j = \overline{p}$ for all $j = 1, \ldots, b$, and $n^{(1)} = n^{(2)} = n_0$ (say). Thus, the equations (19) and (20) reduce to

$$\mathrm{E}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})|\boldsymbol{x}_0 \in \Pi_1\right] = \frac{1}{2}\frac{m}{\nu_j}\delta_j^2$$

and

$$\mathrm{Var}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})|\boldsymbol{x}_0 \in \Pi_1\right] = \frac{\nu^2\left[\frac{1}{2}\delta_j^4(\nu+1) + \delta_j^2(\nu+1) + \overline{p}uv(\nu+\overline{p}uv)/n_0^2\right]}{m^2(\nu-2)(\nu+1)} \tag{B1}$$

where $\nu = 2n_0 - \overline{p}uv - 3$. The proof proceeds now by summing $\mathrm{E}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})|\boldsymbol{x}_0 \in \Pi_1\right]$ over j, and by dividing the variance term in (B1) into three parts as follows

$$\mathrm{Var}_T\left[\widehat{\ell}_j(\boldsymbol{x}_{0[j]})|\boldsymbol{x}_0 \in \Pi_1\right] = \frac{\nu^2}{m^2}\cdot\frac{1}{(nu-2)(\nu+1)}\left[I_j + II_j + III_j\right],$$

so that

$$
\begin{aligned}
I &= \sum_{j=1}^{b} I_j = \frac{1}{2}\frac{\nu^2}{(\nu-2)m^2}\sum_{j=1}^{b}\delta_j^4, && \text{(B2a)}\\
II &= \sum_{j=1}^{b} II_j = \frac{\nu^3\delta^2}{m^2(\nu-2)}\left(1+\frac{1}{n_0}\right), && \text{(B2b)}\\
\text{and}\quad III &= \sum_{j=1}^{b} III_j = \frac{\nu^2 puv(\nu+\overline{p}uv)}{m^2 n_0^2(\nu-2)(\nu+1)}, && \text{(B2c)}
\end{aligned}
$$

where

$$
\begin{aligned}
\delta^2 &= \sum_{j=1}^{b}\delta_j^2\\
&= \sum_{j=1}^{b}\left[\boldsymbol{\mu}'_{\boldsymbol{x}[j]}\cdot\left(\boldsymbol{I}_{vu}\otimes\boldsymbol{\Delta}^{-1}_{1[j]}+\boldsymbol{I}_{v}\otimes\boldsymbol{J}_{u}\otimes\frac{1}{u}\left(\boldsymbol{\Delta}^{-1}_{2[j]}-\boldsymbol{\Delta}^{-1}_{1[j]}\right)+\boldsymbol{J}_{vu}\otimes\frac{1}{vu}\left(\boldsymbol{\Delta}^{-1}_{3[j]}-\boldsymbol{\Delta}^{-1}_{2[j]}\right)\right)\cdot\boldsymbol{\mu}_{\boldsymbol{x}[j]}\right],
\end{aligned}
$$

with $\boldsymbol{\mu}_{\boldsymbol{x}[j]} = \boldsymbol{\mu}^{(1)}_{\boldsymbol{x}[j]} - \boldsymbol{\mu}^{(2)}_{\boldsymbol{x}[j]}$ is the shift vector for the j^{th} block. Expressions $I - III$ in (B2a), (B2b) and (B2c) respectively in combination with some algebraic simplifications, (18) yields

$$
\Phi\Big(-\frac{\delta}{2}\Big[\Big(\frac{\nu}{m\delta}\Big)^2(I+II+III)\Big]^{-1/2}\Big).
$$

Now, by observing that $n_0 \sim \mathcal{O}(n)$ and by ignoring the terms of $\mathcal{O}(1/n)$ and $\mathcal{O}(p/n^2)$, we observe that the term I is $o(n^{-1})$, by the constraints on $\overline{p}$, $\overline{p} \ll n$ and by the boundary conditions on the Mahalanobis distance. As $n \to \infty$, by the assumption (7), II goes to δ^2, a constant which does not depend on $\overline{p}$. Next, by observing $\nu + \overline{p}uv = 2n_0 - 3$ and ν/m^2 is of order $\mathcal{O}(puv/n^2)$, it is then clear that III is bounded by $\mathcal{O}(\overline{p}uv/n^2)$ where $\overline{p}/n^2 \to 0$ as $n \to \infty$ by the assumption (7).

References

[1] 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.

[2] Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third Ed. Wiley, New Jersey.

[3] Bartlett, M.S., 1951. An Inverse Matrix Adjustment Arising in Discriminant Analysis, Annals of Mathematical Statistics 22(1), 107-111.

[4] Bhattacharyya, C., Grate, L.R., Rizki, A., Radisky, D., Molina, F.J., Jordan, M.I., Bissell M.J., Mian, I. S., 2003. Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data, Signal Processing 83(4), 729-743.

[5] 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.

[6] Corander, J., Koski, T., Pavlenko, T., Tillander, A., 2012. Bayesian block-diagonal predictive classifier for Gaussian data, Advances in Intelligent and Soft Computing, Springer Berlin/Heidelberg, forthcoming.

[7] Davis, A., 1987. Moments of linear discirminant functions and an asymptotic confidence interval for the log odds ratio, Biometrika 74, 829-840.

[8] Dudoit, S., Fridlyand J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97, 7787.

[9] Friedman, J., Hastie T., Tibshirani, R., 2008. Sparce inverse covariance estimation with the graphical lasso, Biostatistics 9(3), 432-441.

[10] Harville, D.A., 1997. Matrix algebra from a Statistician's perspective, Springer-Verlag NY.

[11] Kim, K.I., Simon, R., 2011. Probabilistic classifiers with high-dimensional data, Biostatistics 12(3), 399-412.

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

[13] Lai, C., Reinders, M.J., VanT Veer L.J., Wessels, L.F., 2006. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets, BMC Bioinformatics 7, 235.

[14] Leiva, R., 2007. Linear discrimination with equicorrelated training vectors, J. Multivariate Analysis. 98(2), 384-409.

[15] Leiva, R., 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.

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

[17] 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 (2011a).

[18] Leiva, R., Roy, A., 2012. Linear discrimination for three-level multivariate data with separable additive mean vector and doubly exchangeable covariance structure, Computational Statistics and Data Analysis 56(6), 1644-1661.

[19] Lauritzen, S.L., 1996. Graphical Models, Oxford University Press, New York.

[20] Matusita, K., 1966. A distance and related statistics in multivariate analysis. In: Krishnaiah P K (eds) Multivariate analysis. Academic Press, New York, pp 187-202

[21] McLachlan, G.J., 2004. Discriminant analysis and statistical pattern recognition. John Wiley & Sons, Hoboken, New Jersey.

[22] Okamoto, M., 1963. An asymptotic expansion for the distribution of the linear discriminant function. Annals of Mathematical Statistics 34, 1286-1301

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

[24] 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

[25] Pavlenko, T., Björkstrom A., 2010. Exploiting sparse dependence structure in model based classification, in combining soft computing and statistical methods in data analysis, C. Borgelt, G. Gonzlez-Rodrguez, W. Trutschnig, M. Lubiano, M. Gil, P. Grzegorzewski, and O. Hryniewicz, eds., Advances in Intelligent and Soft Computing, Vol. 77, Springer Berlin/Heidelberg, 509-517.

[26] Pavlenko, T., 2003. On feature selection, curse-of-dimensionality and error probability in discriminant analysis, J. of Statistical Planning and Inference 115, 565-584.

[27] Pavlenko, T., Björkstrom A., Tillander, A., 2012. Covariance Structure Approximation via gLasso in High-Dimensional Supervised Classification, Journal of Applied Statistics 1-24.

[28] Rao, C.R., Varadarajan, V.S., 1963. Discrimination of Gaussian processes. Sankhya A 25, 303-330

[29] Ravikumar, P., Wainwright, M.J., Raskutti, G., Yu, B., 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electronic Journal of Statistics 5, 935-980.

[30] Rothman, A., Bickel, P., Levina, E., and Zhu, J., 2008. Sparse permutation invariant covariance estiamtion. Electronic Journal of Statistics 2, 494-515.

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

[32] Seber, G.A.E., 1984. Multivariate observations. John Wiley & Sons, New York.

[33] Shutoh, N., 2011. Aymptotic expansions relating to discrimination based on two-step monotone missing samples. Journal of Statistical Planning and Inference 141, 1297-1306.

[34] Shutoh, N., Hyodo, M., Seo, T., 2011. An asymptotic approximation for EPMC in linear discriminant analysis based on two-step monotione missing samples. Journal of Multivariate Analysis 102, 252-263.

[35] Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Statist. Soc., B 58, 267-288.