Linear discrimination for three-level multivariate data with structured mean vectors and doubly exchangeable covariance structure

Ricardo Leiva[1] **and Anuradha Roy**[2]

[1] *Universidad Nacional de Cuyo, Mendoza, Argentina*
[2] *The University of Texas at San Antonio, USA*

Abstract

Under the assumption of multivariate normality we study discrimination among k populations, where m-variate observations are taken on each individual at u-sites and over v-time points. We do this by using a doubly exchangeable covariance structure consisting of three unstructured covariance matrices for three multivariate levels. More accurately, let t and s stand for a given point in time and a given site, respectively. Let $\boldsymbol{x}_{ts}^{(p)} : (\Omega, P) \to \Re^m,\ 1 \leq t \leq v,\ 1 \leq s \leq u$, be the $m-$dimensional normally distributed random vector underlying the p^{th} population. Then the random families $(\boldsymbol{x}_{1s}^{(p)})_{s\in 1\ldots u}, \ldots (\boldsymbol{x}_{vs}^{(p)})_{s\in 1\ldots u}$ are assumed to be exchangeable. Furthermore, for fixed t, the family of random variables $(\boldsymbol{x}_{ts}^{(p)})_{s\in 1\ldots u}$ is exchangeable.

Let $\boldsymbol{x}_{r,ts}^{(p)}$ be an $m-$variate vector of measurements of the r^{th} individual in the p^{th} population at the s^{th} site (space) and at the t^{th} time point, $r = 1, \ldots, n^{(p)},\ p = 1, \ldots, k,\ s = 1, \ldots, u,\ t = 1, \ldots, v$. Let $\boldsymbol{x}_{r,t}^{(p)}$ be the mu-variate vector of measurements corresponding to the r^{th} individual in the p^{th} population and at the t^{th} time point, that is, for each r, p and t the vector $\boldsymbol{x}_{r,t}^{(p)}$ is obtained by stacking all m responses of the r^{th} individual in the p^{th} population at the t^{th} time point on the first site, then stacking all its m responses on the second site and so on. Let $\boldsymbol{x}_{r}^{(p)} = (\boldsymbol{x}_{r,1}^{(p)\prime}, \ldots, \boldsymbol{x}_{r,v}^{(p)\prime})'$ be the muv-variate vector of all measurements corresponding to the r^{th} individual in the p^{th} population. Let $\boldsymbol{x}_{1}^{(p)}, \ldots, \boldsymbol{x}_{n^{(p)}}^{(p)}$ be a random sample of size $n^{(p)}$ from the p^{th} population with distribution $N_{muv}\left(\boldsymbol{\mu}_{\boldsymbol{x}^{(p)}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}\right)$, where $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ has a doubly exchangeable covariance structure as follows:

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

That is,

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

where $\mathbf{I}_a$ is the $a \times a$ identity matrix and $\mathbf{J}_a = \mathbf{1}_a \mathbf{1}_a'$. The $m \times m$ block diagonals $\mathbf{U}_0$ represent the variance-covariance matrix of the m response variables at any given site and at any given time point, whereas the $m \times m$ block off diagonals $\mathbf{U}_1$ represent the covariance matrix of the m response variables between any two sites and at any given time point. We assume $\mathbf{U}_0$ is constant for all sites and time points, and $\mathbf{U}_1$ is same between any two sites and for all time points. The $m \times m$ block off diagonals $\mathbf{W}$ represent the covariance matrix of the m response variables between any two time points. It is assumed to be the same for any pair of time points, irrespective of the same site or between any two sites.

We develop linear discriminant functions with a doubly exchangeable covariance structure in addition to the multiplicative, additive and unstructured mean vectors. The new discriminant functions with structured mean vectors are very effective in discriminating individuals in small sample scenario. No closed-form expression exists for the maximum likelihood estimates (MLEs) of the unknown population parameters. Iterative algorithms are proposed to calculate the MLEs of these unknown parameters. We compare the structured multiplicative and additive mean models in general. The proposed classification rules are demonstrated on a real data set. In order to exhibit the effectiveness of our new classification rules a number of pairs of training sample sizes, from very small to moderate pairs, are chosen from the populations. The error rates of the proposed classification rules are found to be much less than the error rate of the traditional linear classification rule, when in fact the traditional linear classification rule fails most of the time owing to the small sample sizes.

Keywords

Additive and multiplicative mean structures, Doubly exchangeable covariance structure, Linear discriminant function, Maximum likelihood estimates.

References

Leiva, R. and A. Roy (2008). Classification rules for triply multivariate data with an AR(1) correlation structure on the repeated measures over time. *J. Statist. Plann. Inference*. doi:10.1016/j.jspi.2008.11.013.

Leiva, R. (2007). Linear discrimination with equicorrelated training vectors. *J. Multivariate Anal. 98(2)*, 384–409.

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