Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-024-00593-7

REGULAR ARTICLE

A two-group canonical variate analysis biplot for an optimal display of both means and cases

Niel le Roux[1] · **Sugnet Gardner-Lubbe**[1]

Received: 30 July 2022 / Revised: 22 March 2024 / Accepted: 27 March 2024

Abstract
Canonical variate analysis (CVA) entails a two-sided eigenvalue decomposition. When the number of groups, J, is less than the number of variables, p, at most $J - 1$ eigenvalues are not exactly zero. A CVA biplot is the simultaneous display of the two entities: group means as points and variables as calibrated biplot axes. It follows that with two groups the group means can be exactly represented in a one-dimensional biplot but the individual samples are approximated. We define a criterion to measure the quality of representing the individual samples in a CVA biplot. Then, for the two-group case we propose an additional dimension for constructing an optimal two-dimensional CVA biplot. The proposed novel CVA biplot maintains the exact display of group means and biplot axes, but the individual sample points satisfy the optimality criterion in a unique simultaneous display of group means, calibrated biplot axes for the variables, and within group samples. Although our primary aim is to address two-group CVA, our proposal extends immediately to an optimal three-dimensional biplot when encountering the equally important case of comparing three groups in practice.

Keywords Biplot · Canonical variate analysis · Classification · Data visualization · Discriminant analysis

Mathematics Subject Classification 62H30 · 15A21 · 93B60 · 91C20 · 97K80

1 Introduction

It is difficult to overrate the value of graphical displays to accompany formal statistical classification procedures (see e.g., Tukey 1975). Indeed, it is an open question if a complete assessment of group structure, including the overlap and separation of

✉ Niel le Roux
njlr@sun.ac.za

[1] Stellenbosch University, Stellenbosch, South Africa

Published online: 06 May 2024

groups and the within groups sample behaviour is possible without the aid of suitable graphics, relying exclusively on computed statistical measures and tables. In particular, graphical procedures that are capable of displaying simultaneously the means and the individuals of different groups together with information on the relative contributions of a chosen set of classification variables are highly in demand.

Gabriel (1971) proposed the biplot for displaying simultaneously the rows and columns of a data matrix $\boldsymbol{X} : n \times p$ in a single graph. A year later Gabriel showed how to construct a canonical variate analysis (CVA) biplot (Gabriel 1972) that provides a two-dimensional graphical approximation of various groups optimally separated according to a multidimensional CVA criterion. This CVA biplot became popular among statisticians and practitioners applying linear discriminant analysis (LDA) and CVA in various fields. Gittins (1985) can be consulted for an overview of CVA. Distances in this classical Gabriel CVA biplot are interpreted in terms of inner products between vectors. Such interpretations are not as straightforward as distances in an ordinary scatter plot. The unified biplot methodology proposed by Gower (1995) and extensively discussed in the monograph by Gower and Hand (1996) allows the CVA biplot to be regarded as a multivariate extension of an ordinary scatter plot. The concept of inter-sample distance is central in this approach, while information on the classifier variables is added to the graph of means and individual sample points in the form of biplot axes–an axis calibrated in its original units for each (classifier) variable. The perspective of Gower and Hand (1996) inspires the biplot methodology discussed in this paper. Gower et al. (2011) discuss the one-dimensional CVA biplot for use in two-group studies in some detail. Although the latter authors provide several enhancements to this one-dimensional CVA biplot they failed to address the challenge of constructing an optimal two-dimensional biplot for the two-group case.

Classification procedures in practice are often confronted with the problem of optimally distinguishing between two groups. The aim of several procedures used in multidimensional classification is to separate the group means optimally. This is the aim of CVA and the closely related procedure of LDA. These techniques involve the transformation of the original observations into a so-called canonical space. Flury (1997), among others, proves that in the case of J groups, only the first $K = min(p, J-1)$ elements of the group mean vectors differ in the canonical space. This induces some pitfalls when routinely constructing CVA biplots in the case where $p > J-1$ with J equaling two or three. Since in the case of two groups the group means can be exactly represented on a line, the associated CVA biplot becomes a one-dimensional plot with all p biplot axes representing the different variables on top of each other on the line extending through the representations of the two group means. All n sample points also fall on this line. The theoretical basis for constructing this line is provided by the eigenanalysis of a two-sided eigenvalue problem (see e.g., Gower and Hand (1996) and Gower et al. (2011)). In the two-sample case this eigenanalysis involves only one non-zero (positive) eigenvalue together with $p-1$ zero eigenvalues. Therefore, only the eigenvector associated with the single non-zero eigenvalue is uniquely defined. This eigenvector provides the scaffolding for constructing a one-dimensional CVA biplot. Although a two-dimensional biplot can be constructed using two eigenvectors, the second eigenvector will not be uniquely defined, as is the case when $J > 2$, unless some precautions are taken. A similar situation arises in the case

of three groups: the first two eigenvectors are then uniquely defined but not the third. Why do we want extra scaffolding dimensions for constructing CVA biplots when $p > J - 1$? There is a real advantage when we have a two-group or three-group classification problem: not only are the group means then represented exactly but also an improvement in the approximations of the individual sample points is accomplished. Therefore, in this paper, we first define a measure of how well any mean or individual sample point is represented in a CVA biplot. Then we show, in the case of two groups, how to construct a uniquely defined two-dimensional CVA biplot such that both the means and individual samples are optimally represented together with the variables in the form of calibrated biplot axes. In addition, we will show how this process can be extended directly for constructing a uniquely defined optimal three-dimensional biplot when three groups are considered.

The paper is organized as follows: in the next section, we begin with some historical background of discriminant analysis. After that, we briefly review the basic concepts and theory of CVA biplot methodology according to the perspective of Gower and Hand (1996). This is followed by a section describing the geometry of the graphical representation of the class means and the sample points about them. In section 4 it is discussed why the two-group CVA biplot deserves special attention. We then put forward proposals for one-dimensional CVA biplots as well as a two-dimensional CVA biplot such that the group means in a two-group classification problem are exactly represented together with an optimal representation of the individual sample points. We also show how to generalize this procedure to construct a three-dimensional biplot with similar properties for use when $J = 3 < p$. In section 6 we briefly discussed an alternative unique 2D biplot, which is based on the Bhattacharyya distance. The theoretical results are illustrated in section 7 where we provide examples, covering one- and two-dimensional CVA biplots. Some conclusions are considered in section 8.

2 A brief review of canonical variate analysis

2.1 Two-group discriminant analysis: Fisher's LDA

Two-group discriminant analysis considers two populations (groups) G_1 and G_2. An observation $\boldsymbol{x}$ of $\boldsymbol{X}^T = (X_1, X_2, \ldots, X_p)$ is to be allocated to one of these populations. It is assumed that the density $f_i(\boldsymbol{x})$ of $\boldsymbol{X}$ for $i = (1; 2)$ is known with expected value $\boldsymbol{\mu}_i : p \times 1$ and covariance matrix $\boldsymbol{\Sigma}_i : p \times p$, respectively. Let the prior probability of an unknown $\boldsymbol{x}$ belonging to G_i be given by

$p_1 = P(G_1)$ and $p_2 = P(G_2)$, respectively, with $p_i > 0$ and $p_1 + p_2 = 1$.

Define

$$
\begin{aligned}
\boldsymbol{\mu} &= E(\boldsymbol{X}) = p_1\boldsymbol{\mu}_1 + p_2\boldsymbol{\mu}_2,\\
\boldsymbol{T} &= E\Big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\Big],\\
\boldsymbol{B} &= p_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^T + p_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^T, \text{ and}\\
\boldsymbol{W} &= p_1\boldsymbol{\Sigma}_1 + p_2\boldsymbol{\Sigma}_2.
\end{aligned}
$$

Under the assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, (say) it follows that $\boldsymbol{W} = \boldsymbol{\Sigma}$. The Fisher LDA (Fisher 1936) searches for the linear function

$$Y = \boldsymbol{m}^T \boldsymbol{X}$$

with $E(Y) = \boldsymbol{m}^T E(\boldsymbol{X}) = \boldsymbol{m}^T \boldsymbol{\mu}$, $E(Y|G_i) = \boldsymbol{m}^T \boldsymbol{\mu}_i$ and $var(Y) = \boldsymbol{m}^T \boldsymbol{W} \boldsymbol{m}$ to maximize

$$\frac{p_1\left(\boldsymbol{m}^T \boldsymbol{\mu}_1 - \boldsymbol{m}^T \boldsymbol{\mu}\right)^2 + p_2\left(\boldsymbol{m}^T \boldsymbol{\mu}_2 - \boldsymbol{m}^T \boldsymbol{\mu}\right)^2}{\boldsymbol{m}^T \boldsymbol{W} \boldsymbol{m}} = \frac{\boldsymbol{m}^T \boldsymbol{B} \boldsymbol{m}}{\boldsymbol{m}^T \boldsymbol{W} \boldsymbol{m}}. \tag{1}$$

The maximum is obtained from the eigenequation

$$(\boldsymbol{W}^{-1} \boldsymbol{B}) \boldsymbol{m} = \boldsymbol{m} \boldsymbol{\Lambda}.$$

Pre-multiplying the above equation with $\boldsymbol{W}^{1/2}$ leads to

$$\left(\boldsymbol{W}^{-1/2} \boldsymbol{B} \boldsymbol{W}^{-1/2}\right)\left(\boldsymbol{W}^{1/2} \boldsymbol{m}\right) = \left(\boldsymbol{W}^{1/2} \boldsymbol{m}\right) \boldsymbol{\Lambda},$$

so that $\boldsymbol{l} = \boldsymbol{W}^{1/2} \boldsymbol{m}$ is the eigenvector of $\boldsymbol{W}^{-1/2} \boldsymbol{B} \boldsymbol{W}^{-1/2}$ associated with the largest eigenvalue λ_1 and $\boldsymbol{m} = \boldsymbol{W}^{-1/2} \boldsymbol{l}$.

Since $rank(\boldsymbol{B}) = 1 = rank(\boldsymbol{W}^{-1} \boldsymbol{B}) = rank(\boldsymbol{W}^{-1/2} \boldsymbol{B} \boldsymbol{W}^{-1/2})$, it follows that

$$\boldsymbol{B} \boldsymbol{M} = \boldsymbol{W} \boldsymbol{M} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}. \tag{2}$$

If we choose $\boldsymbol{L}$ to be an orthogonal matrix in $(\boldsymbol{W}^{-1/2} \boldsymbol{B} \boldsymbol{W}^{-1/2}) \boldsymbol{L} = \boldsymbol{L} \boldsymbol{\Lambda}$, then

$$\boldsymbol{L}^T \boldsymbol{L} = \boldsymbol{I} = \boldsymbol{M}^T \boldsymbol{W} \boldsymbol{M}$$

and

$$\boldsymbol{M}^T \boldsymbol{B} \boldsymbol{M} = \boldsymbol{M}^T \boldsymbol{W} \boldsymbol{M} \boldsymbol{\Lambda} = \boldsymbol{\Lambda}.$$

The transformation

$$Y_k = \boldsymbol{m}_k^T \boldsymbol{X}$$

for $k = 1, 2, \dots, p$ is termed a transformation into the canonical space with Y_1 the first canonical variate, where

$E(Y_1|G_i) = \boldsymbol{m}_1^T \boldsymbol{\mu}_i$ and $Var(Y_1|G_i) = Var(Y_1|G_i) = 1$ for $i = 1;\ 2$.

After properly scaled, the solution $\boldsymbol{m}_1$ maximizing (1) can be written as

$$\boldsymbol{W}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{3}$$

which is known as Fisher's linear discriminant function (LDF). The maximum of (1) is given by the squared Mahalanobis distance, namely

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{W}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

For $k = 2, 3, \ldots, p$ we have $Var(Y_k|G_i) = 1$ and

$$\boldsymbol{m}_k^T \boldsymbol{B} \boldsymbol{m}_k = 0. \tag{4}$$

Since

$$\begin{aligned}\boldsymbol{B} =& p_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^T + p_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^T \\ =& \left(p_1 p_2^2 + p_2 p_1^2\right)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T,\end{aligned}$$

it follows from (4) that $\boldsymbol{m}_k^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$ and all differences vanish between the groups for the second and further canonical variates.

Rao (1948) extends the Fisher's LDA procedure to $J > 2$ groups by deriving $J - 1$ linear discriminant functions of the form (3), but in this paper we are primarily interested in the case $J = 2$.

2.2 Bayes linear and quadratic classifiers

Under the assumption of multivariate normal distributions with different covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, the Bayes quadratic classifier (see e.g., Hastie et al. 2001) is given by

$$\begin{aligned}&\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_1) - \frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_2) \\ &\quad + \frac{1}{2} log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) \gtrless log\left(\frac{p_1}{p_2}\right).\end{aligned}$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, (say) we have the Bayes linear classifier

$$\frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} + \frac{1}{2}\left(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2\right) \gtrless log\left(\frac{p_1}{p_2}\right).$$

It is clear that if $p_1 = p_2$, the Bayes linear classifier is equivalent to Fisher's LDF.

In general, the Bhattacharyya distance measures the similarity of two probability distributions. When the distributions concerned are $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ it is

given by (see e.g., Fukunaga 1990; Hennig 2004):

$$D_{Bhat} = \frac{1}{8}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2} log \left(\frac{\left| \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}} \right). \quad (5)$$

It can be shown that the sample form of (5) provides an upper bound for the Bayes error (see e.g., McLachlan 1992). Furthermore, when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ then (5) is proportional to a squared Mahalanobis distance. For this case, Fukunaga (1990) shows that D_{Bhat} is maximized by $Y = \boldsymbol{m}^T \boldsymbol{X}$ where $\boldsymbol{Bm} = \lambda_1 \boldsymbol{Wm}$, so that maximization is achieved when $\boldsymbol{m}$ is taken as (3).

2.3 Two-group discrimination where group means and or group covariance matrices may differ

While LDA discussed above allows only for groups to differ with respect to the means, Fukunaga (1990) looks for a linear transformation $\boldsymbol{Y} = \boldsymbol{XM}$ to separate groups with respect to their means or covariance matrices. With the notation of the subsection above, write

$$\begin{aligned} \boldsymbol{W} =& p_1 \boldsymbol{\Sigma}_1 + p_2 \boldsymbol{\Sigma}_2 = p_1 E\Big[(\boldsymbol{X} - \boldsymbol{\mu}_1)(\boldsymbol{X} - \boldsymbol{\mu}_1)^T | G_1 \Big] \\ &+ p_2 E\Big[(\boldsymbol{X} - \boldsymbol{\mu}_2)(\boldsymbol{X} - \boldsymbol{\mu}_2)^T | G_2 \Big]. \end{aligned}$$

Fukunaga (1990) provides four criteria for class separability:

$J_1 = tr(\boldsymbol{S}_2^{-1}\boldsymbol{S}_1)$; $J_2 = log|\boldsymbol{S}_2^{-1}\boldsymbol{S}_1|$; $J_3 = tr(\boldsymbol{S}_1) - \lambda(tr(\boldsymbol{S}_2) - c)$, where λ is a Lagrange multiplier, c is a constant, and $J_4 = tr(\boldsymbol{S}_1)/tr(\boldsymbol{S}_2)$, where $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, respectively, are one of $\boldsymbol{T}$, $\boldsymbol{B}$, or, $\boldsymbol{W}$.

Let $J_i(k)$ indicate the criterion with $\boldsymbol{S}_1 = \boldsymbol{B}$ and $\boldsymbol{S}_2 = \boldsymbol{W}$, and where $\boldsymbol{Y} = \boldsymbol{XM}$ with $\boldsymbol{M} : p \times k$. Since we are restricted to linear transformations, optimization of $J_1(1)$ is equivalent to Fisher discriminant analysis with (1) and $J_1(1) = \lambda_1$ with no other dimension contributing to the value of J_1. Fukunaga (1990) further shows that criterion J_1 gives the same optimum transformation for other combinations of $\boldsymbol{B}$, $\boldsymbol{W}$, and $\boldsymbol{T}$ for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ and also for optimizing J_2.

When $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$, (5) becomes

$$\begin{aligned} D_{Bhat} =& \frac{1}{2} log \left(\frac{\left| \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}} \right) \\ =& \frac{1}{4} \Big[log\Big(|\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 + 2\boldsymbol{I}_p| \Big) - plog(4) \Big]. \end{aligned} \quad (6)$$

If more than a single dimension is needed we have that

$$J_1 = log|(\boldsymbol{M}^T \boldsymbol{\Sigma}_2 \boldsymbol{M})^{-1}(\boldsymbol{M}^T \boldsymbol{\Sigma}_1 \boldsymbol{M}) + (\boldsymbol{M}^T \boldsymbol{\Sigma}_1 \boldsymbol{M})^{-1}(\boldsymbol{M}^T \boldsymbol{\Sigma}_2 \boldsymbol{M}) + 2\boldsymbol{I}_k|$$

is maximized by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{M}$ with $\boldsymbol{M} : p \times k$, where

$$(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)\boldsymbol{M} = \boldsymbol{M}\Big(\Big(\boldsymbol{M}^T\boldsymbol{\Sigma}_2\boldsymbol{M}\Big)^{-1}\Big(\boldsymbol{M}^T\boldsymbol{\Sigma}_1\boldsymbol{M}\Big)\Big)$$

and

$$(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2)\boldsymbol{M} = \boldsymbol{M}\Big((\boldsymbol{M}^T\boldsymbol{\Sigma}_1\boldsymbol{M})^{-1}(\boldsymbol{M}^T\boldsymbol{\Sigma}_2\boldsymbol{M})\Big).$$

Thus, $\boldsymbol{M}$ must contain the eigenvectors of both $\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$. However, they have the same eigenvectors since $\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1 = (\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2)^{-1}$ and we have that

$$(\boldsymbol{M}^T\boldsymbol{\Sigma}_2\boldsymbol{M})^{-1}(\boldsymbol{M}^T\boldsymbol{\Sigma}_1\boldsymbol{M}) = \boldsymbol{\Lambda}$$

and

$$(\boldsymbol{M}^T\boldsymbol{\Sigma}_1\boldsymbol{M})^{-1}(\boldsymbol{M}^T\boldsymbol{\Sigma}_2\boldsymbol{M}) = \boldsymbol{\Lambda}^{-1},$$

so that $(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)\boldsymbol{M} = \boldsymbol{M}\boldsymbol{\Lambda}$ and $(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2)\boldsymbol{M} = \boldsymbol{M}\boldsymbol{\Lambda}^{-1}$.

Therefore, if $k > 1$ dimensions are needed the k eigenvectors are chosen, which correspond to the k largest values for J, i.e., corresponding to the largest values $\lambda_i + \frac{1}{\lambda_i} + 2$.

3 Discriminant analysis with sampled data

In practice, discriminant analysis is usually performed using sampled data. This necessitates the substitution of population parameters with sample estimates in the formulas introduced in section 2. The plug-in principle and maximum likelihood method are popular methods for this.

It is well known that LDA often outperforms quadratic discriminant analysis (QDA) even when the assumption of equal covariance matrices is violated (see e.g., Flury et al. 1997; McLachlan 1992). This can be attributed to the large number of parameters that have to be estimated in QDA with over-parameterization inducing a loss of power. This contributes to the popularity of LDA among practitioners and so in the rest of the paper, our focus will be on LDA.

Consider the data matrix $\boldsymbol{X} : n \times p$ centered such that $\mathbf{1}^T\boldsymbol{X} = \mathbf{0}^T$. The data contained in $\boldsymbol{X}$ consists of p measurements made for each of the J groups. The group sizes are $n_1, n_2, \ldots, n_J$, respectively, such that $\sum_{i=1}^{J} n_i = n$. Let $\boldsymbol{N}_g = diag(n_1, n_2, \ldots, n_J)$, so that a matrix of group means can be calculated as

$$\overline{\boldsymbol{X}} : J \times p = \boldsymbol{N}_g^{-1}\boldsymbol{G}^T\boldsymbol{X} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{X}, \tag{7}$$

where $\boldsymbol{G} : n \times J$ denotes an indicator matrix defining the J groups.

Let $\mathcal{V}(\boldsymbol{X}^T)$ denote the vector space generated by the columns of $\boldsymbol{X}^T$. We assume this vector space of p-vectors to be of dimension p. Since each row of $\overline{\boldsymbol{X}}$ is a linear combination of the rows of $\boldsymbol{X}$ it follows that $\overline{\boldsymbol{X}}^T \in \mathcal{V}(\boldsymbol{X}^T)$.

Define:

1. $\boldsymbol{S}_B : p \times p$, as the between-group matrix of squares and products: $\boldsymbol{S}_B = \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} = \boldsymbol{X}^T \boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{X}$ and
2. $\boldsymbol{S}_W : p \times p$ as the within-group matrix of squares and products: $\boldsymbol{S}_W = \boldsymbol{X}^T\boldsymbol{X} - \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} = \boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T)\boldsymbol{X}$.

Generally, $rank(\boldsymbol{S}_W) = p$ while $rank(\boldsymbol{S}_B) = min(J-1, p)$.

The two-sided eigenvalue problem

$$(\boldsymbol{S}_B)\boldsymbol{B} = (\boldsymbol{S}_W)\boldsymbol{B}\boldsymbol{\Lambda} \tag{8}$$

provides the solution $\boldsymbol{b}_1$ to the CVA criterion

$$\underset{\boldsymbol{b}}{\text{maximize}} \left(\frac{\boldsymbol{b}^T(\boldsymbol{S}_B)\boldsymbol{b}}{\boldsymbol{b}^T(\boldsymbol{S}_W)\boldsymbol{b}} \right), \text{ subject to } \boldsymbol{b}^T(\boldsymbol{S}_W)\boldsymbol{b} = 1. \tag{9}$$

In the above, the diagonal matrix $\boldsymbol{\Lambda} : p \times p$ contains the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, where $\lambda_J = \lambda_{J+1} = \ldots = \lambda_p = 0$ if $J < p+1$.

The matrix $\boldsymbol{B} = \left[\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p\right]$ contains all p solutions to the two-sided eigenvalue problem. Only the solution $\boldsymbol{b}_1$ is optimal for the CVA criterion (9). The matrix $\boldsymbol{B} : p \times p$ is non-singular with $\boldsymbol{B}^{-1} = \boldsymbol{B}^T(\boldsymbol{S}_W)$, while the columns of $\boldsymbol{B}$ are orthogonal in the metric $\boldsymbol{S}_W$ because of the constraints $\boldsymbol{B}^T(\boldsymbol{S}_W)\boldsymbol{B} = \boldsymbol{I}$.

Canonical variates are defined by the transformation $\boldsymbol{y}^T = \boldsymbol{x}^T\boldsymbol{B}$, where $\boldsymbol{x}$ is any p-vector belonging to $\mathcal{V}(\boldsymbol{X}^T)$. The centered data matrix itself is transformed to the canonical variate values matrix $\boldsymbol{Y} : n \times p$ through the one-to-one (canonical) transformation

$$\boldsymbol{Y} : n \times p = \boldsymbol{X}\boldsymbol{B}. \tag{10}$$

The transformation (10) implies a transformation of $\mathcal{V}(\boldsymbol{X}^T)$ to $\mathcal{V}(\boldsymbol{Y}^T)$, the canonical space, of dimension p since $rank(\boldsymbol{Y}) = rank(\boldsymbol{X})$. Furthermore, (10) implies that

$$\overline{\boldsymbol{X}}\boldsymbol{B} = \overline{\boldsymbol{Y}} : J \times p. \tag{11}$$

We will call $\overline{\boldsymbol{Y}}$ the canonical means matrix. It follows that the columns of $\overline{\boldsymbol{Y}}^T$ generate a subspace of dimension $min(J-1, p)$ of $\mathcal{V}(\boldsymbol{Y}^T)$. This subspace is denoted by $\mathcal{V}(\overline{\boldsymbol{Y}}^T)$.

4 Biplot display of the canonical means matrix $\overline{Y}$ and the canonical variates values matrix Y

An r-dimensional canonical variate analysis (CVA) plot is constructed by taking the first r canonical variates, associated with

$$\boldsymbol{B}_r = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r], \text{ where } \boldsymbol{B} = \left[\boldsymbol{B}_r, \boldsymbol{b}_{r+1}, \ldots, \boldsymbol{b}_p\right] \tag{12}$$

to provide the coordinates (or scaffolding) for representing the J canonical group means contained in (11) as points in r dimensions. If this plot is equipped with p linear axes to represent the original p variables, a CVA biplot is obtained. Each of these biplot axes is determined by a vector, which induces also a graduation on it. Gower and Hand (1996) consider two types of CVA biplots, each one characterized by its system of p linear axes, its aim and its corresponding geometry:

- The *interpolation* biplot, which has the aim of placing on the plot the image $(y_1, y_2, ..., y_r)$ of any new point $\boldsymbol{x} \in \mathcal{V}(\boldsymbol{X}^T)$.
- The *prediction* biplot, which has the aim of estimating the point $\boldsymbol{x}$ (i.e., the set of variable values) having as an image a given point $(y_1, y_2, ..., y_r)$ in the plot.

Once the CVA biplot for representing the group means is constructed, all transformed samples contained in $\boldsymbol{Y}$ can be interpolated into the biplot. Thus both the canonical means and the transformed samples $\boldsymbol{Y}$ can be displayed in a CVA biplot in an r-dimensional subspace of $\mathcal{V}(\overline{\boldsymbol{Y}}^T)$, (with $r \leq min(J-1, p)$). Typically, an r of two or three will be chosen to construct this subspace that we will call the biplot space.

Gower and Hand (1996) show the above processes of prediction and interpolation to be based on the following: A sample $\boldsymbol{x} : p \times 1$ can be interpolated into $\mathcal{V}(\boldsymbol{Y}^T)$ by

$$\boldsymbol{y} : p \times 1 = \boldsymbol{B}^T \boldsymbol{x}, \text{ i.e., } \boldsymbol{y}^T = \boldsymbol{x}^T \boldsymbol{B} = \sum_{k=1}^{p} (x_k \boldsymbol{e}_k^T) \boldsymbol{B}.$$

The representation of $\boldsymbol{x}$ in the biplot space is given by

$$\boldsymbol{z}^T : 1 \times r = \boldsymbol{x}^T \boldsymbol{B}_r = \sum_{k=1}^{p} (x_k \boldsymbol{e}_k^T) \boldsymbol{B}_r, \tag{13}$$

where $\boldsymbol{B}_r$ is defined in (12).

Prediction is the inverse of interpolation and since $\boldsymbol{B}$ is non-singular it follows by inverting the above formula for interpolation that

$\boldsymbol{x}^T = \boldsymbol{y}^T \boldsymbol{B}^{-1}$. The matrix $\boldsymbol{B}^{-1}$ can be partitioned into

$$\boldsymbol{B}^{-1} = \begin{bmatrix} \boldsymbol{B}^{(r)} : r \times p \\ \boldsymbol{B}^{(2)} : p - r \times p \end{bmatrix}. \tag{14}$$

The predicted value for the kth variable can be written as $\hat{x}_k = \boldsymbol{z}^T \boldsymbol{B}^{(r)} \boldsymbol{e}_k$ and therefore, the predicted value for $\boldsymbol{x}^T$ is

$$\begin{aligned}\hat{\boldsymbol{x}}^T =& \boldsymbol{z}^T \boldsymbol{B}^{(r)} \left[\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\right] \\ =& \boldsymbol{z}^T \boldsymbol{B}^{(r)} \\ =& \boldsymbol{x}^T \boldsymbol{B}_r \boldsymbol{B}^{(r)}. \end{aligned} \tag{15}$$

It follows from (15) that

$$\hat{\boldsymbol{X}} = \boldsymbol{X} \boldsymbol{B}_r \boldsymbol{B}^{(r)}, \tag{16}$$

and in addition

$$\hat{\overline{\boldsymbol{X}}} = \overline{\boldsymbol{X}} \boldsymbol{B}_r \boldsymbol{B}^{(r)}. \tag{17}$$

Since in the CVA biplot described above, the samples are interpolated into the biplot constructed for the canonical means, it is expected that the class means will be better represented than the canonical variate values i.e., the rows of $\boldsymbol{Y}$. What is needed then, are measures of fit for use in CVA.

4.1 Measures of fit for use in CVA

From the identity

$$\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{G}^T \boldsymbol{G})^{-1} \boldsymbol{G}^T \boldsymbol{X} + \left(\boldsymbol{I} - \boldsymbol{G}(\boldsymbol{G}^T \boldsymbol{G})^{-1} \boldsymbol{G}^T\right) \boldsymbol{X},$$

we have that

$$\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{X}^T \boldsymbol{Q} \boldsymbol{X} + \boldsymbol{X}^T (\boldsymbol{I} - \boldsymbol{Q}) \boldsymbol{X},$$

where the matrix $\boldsymbol{Q} : n \times n = \boldsymbol{G}(\boldsymbol{G}^T \boldsymbol{G})^{-1} \boldsymbol{G}^T$ is positive semi-definite, symmetric and idempotent.

This between-and-within group structure is of interest for:

- Assigning a given sample to its most appropriate group.
- Relating the groups to one another.

A study of the relationships among the groups encourages the use of low-dimensional approximations for visualization, including CVA biplots.

4.2 Measures of fit for CVA biplots: recovering the canonical group means

The orthogonal partitioning

$$\boldsymbol{B}^T \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} \boldsymbol{B} = \boldsymbol{B}^T (\hat{\overline{\boldsymbol{X}}})^T \boldsymbol{N}_g \hat{\overline{\boldsymbol{X}}} \boldsymbol{B} + \boldsymbol{B}^T \left(\overline{\boldsymbol{X}} - \hat{\overline{\boldsymbol{X}}}\right)^T \boldsymbol{N}_g \left(\overline{\boldsymbol{X}} - \hat{\overline{\boldsymbol{X}}}\right) \boldsymbol{B},$$

(see Gardner-Lubbe et al. 2008) allows an overall measure of how well the group means are represented in the CVA biplot, namely

$$Overall\ \ quality = \frac{tr\left(\boldsymbol{B}^T(\hat{\overline{\boldsymbol{X}}})^T\boldsymbol{N}_g\hat{\overline{\boldsymbol{X}}}\boldsymbol{B}\right)}{tr\left(\boldsymbol{B}^T\overline{\boldsymbol{X}}^T\boldsymbol{N}_g\overline{\boldsymbol{X}}\boldsymbol{B}\right)} = \frac{\sum_{k=1}^{r}\lambda_k}{\sum_{k=1}^{p}\lambda_k}.$$

In the orthogonal partitioning above, the matrix $\boldsymbol{B}$ can be eliminated to define *axis predictivities* as the diagonal elements of the matrix

$$\boldsymbol{\Pi} : p \times p = diag\left((\hat{\overline{\boldsymbol{X}}})^T\boldsymbol{N}_g\hat{\overline{\boldsymbol{X}}}\right)\left[diag(\overline{\boldsymbol{X}}^T\boldsymbol{N}_g\overline{\boldsymbol{X}})\right]^{-1}.$$

Each axis predictivity is a measure of how well the values for the group means can be determined from the CVA biplot for the variable associated with that particular biplot axis. We note that the overall quality is a weighted mean of the individual axis predictivities.

4.3 Measures of fit for CVA biplots: recovering the individual samples

We also need predictivities for individual samples corrected for class means i.e., for $(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}$. So, our starting point becomes the decomposition

$$(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}\boldsymbol{B} = (\boldsymbol{I}-\boldsymbol{Q})\hat{\boldsymbol{X}}\boldsymbol{B} + (\boldsymbol{I}-\boldsymbol{Q})\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)\boldsymbol{B},$$

where $\hat{\boldsymbol{X}}$ is defined in (16). Then the following orthogonal decompositions (see Gower et al. 2011) hold:

1. *Type A*

$$\begin{aligned}\boldsymbol{B}^T\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}\boldsymbol{B} =& \boldsymbol{B}^T\hat{\boldsymbol{X}}^T(\boldsymbol{I}-\boldsymbol{Q})\hat{\boldsymbol{X}}\boldsymbol{B}\\ &+\boldsymbol{B}^T\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)^T(\boldsymbol{I}-\boldsymbol{Q})\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)\boldsymbol{B}.\end{aligned}$$

2. *Type B*

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{Q}) =& (\boldsymbol{I}-\boldsymbol{Q})\hat{\boldsymbol{X}}\boldsymbol{B}\boldsymbol{B}^T\hat{\boldsymbol{X}}^T(\boldsymbol{I}-\boldsymbol{Q})\\ &+(\boldsymbol{I}-\boldsymbol{Q})\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)\boldsymbol{B}\boldsymbol{B}^T\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)^T(\boldsymbol{I}-\boldsymbol{Q}).\end{aligned}$$

While the class means are exactly represented in a subspace of dimension $min(J-1, p)$ of the canonical space, this is generally not true for the individual samples. From Type B orthogonality within-group sample predictivities can be defined as the diagonal elements of

$$\boldsymbol{\Psi}_W : n \times n \;=\; diag\Big((\boldsymbol{I}-\boldsymbol{Q})\hat{\boldsymbol{X}}\boldsymbol{S}_W^{-1}\hat{\boldsymbol{X}}^T(\boldsymbol{I}-\boldsymbol{Q})\Big)\Big[diag\Big((\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}\boldsymbol{S}_W^{-1}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{Q})\Big)\Big]^{-1}$$

4.4 CVA biplots for $J = 2$ groups

When $J = 2$ it now follows that:

- The underlying two-sided eigenequation has one non-zero eigenvalue and $p - 1$ zero eigenvalues.
- All p class means are exactly represented in a single dimension.
- This single dimension contains the p biplot axes (each with predictivity 100% for recovering the group means) on top of each other.
- Overall quality of representing group means is 100%.
- The one-dimensional CVA biplot is optimal for representing groups irrespective of the number of variables p.
- The samples are not exactly represented in the one-dimensional CVA biplot.

Our challenge is now to add another dimension for improving recovering of sample information without changing optimality properties already available for the group means. We address this challenge by considering the orthonormal complement in the canonical space of the subspace containing the two group means. Therefore, we consider eigenvectors associated with the zero eigenvalues. These eigenvectors have no natural ordering associated with them. So, any one of these eigenvectors or even any linear combination of them has the same claim to be used as a second scaffolding axis. Hence, our aim is to find the linear combination that satisfies some optimality criterion using Type A and Type B orthogonality for natural candidates.

4.5 Optimal two-dimensional CVA biplot for $J = 2$ groups: optimality criterion based on *Type B* orthogonality

Consider minimizing

$$sum\Big\{diag\Big((\boldsymbol{I}-\boldsymbol{Q})(\boldsymbol{X}-\widehat{\boldsymbol{X}})\boldsymbol{B}\boldsymbol{B}^T(\boldsymbol{X}-\widehat{\boldsymbol{X}})^T(\boldsymbol{I}-\boldsymbol{Q})\Big)\Big\}.$$

Now,

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{Q})\Big(\boldsymbol{X}-\hat{\boldsymbol{X}}\Big) &= \Big(\boldsymbol{I}-\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\Big)\Big(\boldsymbol{X}-\hat{\boldsymbol{X}}\Big)\\ &= \Big(\boldsymbol{X}-\hat{\boldsymbol{X}}\Big)-\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\Big(\boldsymbol{X}-\hat{\boldsymbol{X}}\Big)\\ &= \Big(\boldsymbol{X}-\hat{\boldsymbol{X}}\Big)-\boldsymbol{G}\overline{\boldsymbol{X}}+\boldsymbol{G}\hat{\overline{\boldsymbol{X}}}.\end{aligned}$$

We have $\boldsymbol{G}\hat{\overline{\boldsymbol{X}}} = \boldsymbol{G}\overline{\boldsymbol{X}}$ if the eigenvector associated with $\lambda > 0$ is chosen. Therefore, maximizing the sum of within-group sample predictivities becomes equivalent

to minimizing $sum\left\{diag\left(\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)\boldsymbol{B}\boldsymbol{B}^{T}\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)^{T}\right)\right\}$. However, this sum remains constant for any additional eigenvector. This is not surprising because in the canonical space the constraint $\boldsymbol{B}^{T}\boldsymbol{S}_{W}\boldsymbol{B}=\boldsymbol{I}$, implies constant variation in all dimensions, resulting in Type B orthogonality not useful to define a criterion for an optimal two-dimensional biplot.

4.6 Optimal two-dimensional CVA biplot for $J = 2$ groups: optimality criterion based on *Type A* orthogonality

Since the matrix $\boldsymbol{B}$ is non-singular it can be eliminated from the equation defining Type A orthogonality. As before, $(\boldsymbol{I}-\boldsymbol{Q})\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)=\left(\boldsymbol{X}-\hat{\boldsymbol{X}}\right)$. Therefore, the proposed optimality criterion for constructing an optimal two-dimensional CVA biplot for two groups is the total squared reconstruction error for samples:

$$TSRES = tr\left\{\left(X-\widehat{X}\right)\left(X-\widehat{X}\right)^{T}\right\}. \tag{18}$$

A similar measure of the goodness of approximations of the means can be defined as the total squared reconstruction error for means:

$$TSREM = tr\left\{\left(\overline{X}-\widehat{\overline{X}}\right)\left(\overline{X}-\widehat{\overline{X}}\right)^{T}\right\}. \tag{19}$$

5 Optimal CVA biplots when the number of groups is less than or equal to the number of variables

From now on we consider the case where $J < p+1$. Then it follows that the canonical means matrix $\overline{\boldsymbol{Y}}$ is of the form

$$\begin{bmatrix}\overline{y}_{1}^{T}\\ \cdots\\ \overline{y}_{J}^{T}\end{bmatrix}=\begin{bmatrix}k_{11} & \cdots & k_{1(J-1)} & 0 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ k_{J1} & \cdots & k_{J(J-1)1} & 0 & \cdots & 0\end{bmatrix}, \tag{20}$$

(see e.g., Flury, 1997, p. 491).

Hence, the canonical transformation optimally separates the p-vectors $\overline{y}_{1}^{T},\overline{y}_{2}^{T},\ldots,\overline{y}_{J}^{T}$ in $J-1$ dimensions while all differences among them vanish in dimensions $J, J+1,\ldots,p$. It follows that $\mathcal{V}(\overline{\boldsymbol{Y}}^{T})$ is now of dimension $J-1$ and it is thus possible to consider a biplot space of dimension $r > J-1$. The resulting r-dimensional CVA biplot is not uniquely defined because the first $J-1$ columns of $\boldsymbol{B}: p\times p$ appended with any set of $r-J+1$ of its remaining columns will result in a biplot where the canonical means are exactly represented. Therefore, $\overline{\boldsymbol{x}}_{j}^{T}\boldsymbol{B}=\left[k_{j1}\ldots k_{j(J-1)}0\ldots 0\right]$ for $j=1,2,\ldots,J$ with the predicted value for $\overline{\boldsymbol{x}}_{j}$

given by the jth row, $\hat{\overline{\boldsymbol{x}}}_j^T$, of (17). It follows that

$$\begin{aligned}
\hat{\overline{\boldsymbol{x}}}_j^T &= \overline{\boldsymbol{x}}_j^T \boldsymbol{B}_r \boldsymbol{B}^{(r)} \\
&= \overline{\boldsymbol{x}}_j^T \left[\boldsymbol{b}_1\ \boldsymbol{b}_2\ \ldots\ \boldsymbol{b}_{J-1}\ \boldsymbol{b}_J\ \ldots \boldsymbol{b}_r\right] \begin{bmatrix} \boldsymbol{b}^{(1)} \\ \boldsymbol{b}^{(2)} \\ \ldots \\ \boldsymbol{b}^{(r)} \end{bmatrix}, \text{ where } \begin{bmatrix} \boldsymbol{b}^{(1)} \\ \boldsymbol{b}^{(2)} \\ \ldots \\ \boldsymbol{b}^{(r)} \end{bmatrix} = \boldsymbol{B}^{(r)} \\
&= \overline{\boldsymbol{x}}_j^T \left[\boldsymbol{b}_1\ \boldsymbol{b}_2\ \ldots\ \boldsymbol{b}_{J-1}\ \boldsymbol{b}_J\ \ldots \boldsymbol{0}\right] \begin{bmatrix} \boldsymbol{b}^{(1)} \\ \boldsymbol{b}^{(2)} \\ \ldots \\ \boldsymbol{b}^{(p)} \end{bmatrix}, \text{ with } \begin{bmatrix} \boldsymbol{b}^{(1)} \\ \boldsymbol{b}^{(2)} \\ \ldots \\ \boldsymbol{b}^{(p)} \end{bmatrix} = \boldsymbol{B}^{-1} \\
&= \overline{\boldsymbol{x}}_j^T \text{ for } j = 1, 2, \ldots, J.
\end{aligned}$$

Therefore, in the above r-dimensional CVA biplot the canonical means are exactly represented, resulting in $TSREM = 0$. For a sample $\boldsymbol{x}_i^T$, we have the ith row of $\boldsymbol{X}$,

$$\boldsymbol{x}_i^T = \boldsymbol{x}_i^T \boldsymbol{B}_r \boldsymbol{B}^{(r)} \neq \boldsymbol{x}_i^T \boldsymbol{B}\boldsymbol{B}^{-1},$$

since in general $\boldsymbol{x}_i^T \boldsymbol{b}_j$ is not zero for all $j = J, J+1, \ldots, p$ so that $TSRES > 0$.

This leaves us with the challenge to construct scaffolding axes $j = J, J+1, \ldots, r$ in addition to those contained in $\boldsymbol{B}_{J-1}$ so that *TSRES* is minimized without sacrificing what we already have for the class means in $J-1$ dimensions.

Possible candidates for the additional scaffolding axes are any $r - J + 1$ of the vectors $\boldsymbol{b}_J, \boldsymbol{b}_{J+1}, \ldots, \boldsymbol{b}_p$. All these vectors are associated with the zero eigenvalues (diagonal elements of $\boldsymbol{\Lambda}$). Therefore, there is no natural ordering of them as is the case with the $J-1$ eigenvectors associated with the non-zero eigenvalues. Furthermore, since $\overline{\boldsymbol{X}}\boldsymbol{b}_i = \boldsymbol{0}$ for $i = J, J+1, \ldots, p$ it follows that $\overline{\boldsymbol{X}}\boldsymbol{d} = \boldsymbol{0}$ where $\boldsymbol{d}$ is any linear combination of the vectors $\boldsymbol{b}_J, \boldsymbol{b}_{J+1}, \ldots, \boldsymbol{b}_p$. A similar result will hold for any set of basis vectors of the vector space generated by the columns of the matrix

$$\boldsymbol{B}^* = \left[\boldsymbol{b}_J, \boldsymbol{b}_{J+1}, \ldots, \boldsymbol{b}_p\right], \tag{21}$$

so that $\overline{\boldsymbol{X}}\boldsymbol{B}^* = \boldsymbol{0}$.

Therefore, a set of $r - J + 1$ linear independent vectors of the form $\boldsymbol{d}$ where $\boldsymbol{d}$ is a linear combination of any basis of $\mathcal{V}(\boldsymbol{B}^*)$ is needed such that the scaffolding vectors consisting of the first $J-1$ columns of $\boldsymbol{B}$ together with the $r-J+1$ $\boldsymbol{d}$ vectors minimize *TSRES* for all legitimate choices of the $\{\boldsymbol{d}\}$. Write these r scaffolding vectors as the columns of the matrix $\boldsymbol{D}_r$ i.e.,

$$\boldsymbol{D}_r = \left[\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{J-1}, \boldsymbol{d}_J, \boldsymbol{d}_{J+1}, \ldots, \boldsymbol{d}_r\right] \tag{22}$$

and let the columns of $\boldsymbol{D} : p \times p = \left[\boldsymbol{D}_r, \boldsymbol{d}_{r+1}, \ldots, \boldsymbol{d}_p\right]$ represent a basis of $\mathcal{V}(\boldsymbol{B})$. It follows that any column of $\boldsymbol{B}$ can be written as a linear combination of the columns

of $\boldsymbol{D}$. Therefore, there exists a non-singular matrix $\boldsymbol{C} : p \times p$ such that $\boldsymbol{B} = \boldsymbol{DC}$ i.e., $\boldsymbol{D} = \boldsymbol{BF}$ with $\boldsymbol{F} = \boldsymbol{C}^{-1}$.

Straightforward algebraic manipulation shows that $\boldsymbol{F}$ is of the form

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{F}^* \end{bmatrix}, \tag{23}$$

where $\boldsymbol{F}^*$ is an $(p - J + 1) \times (p - J + 1)$ orthogonal matrix. We provide a detailed derivation as supplementary material. Write

$$\boldsymbol{F}^* = \left[\boldsymbol{f}_1^*, \boldsymbol{f}_2^*, \ldots, \boldsymbol{f}_{p-J+1}^* \right]. \tag{24}$$

Then, $\boldsymbol{F}^{-1} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{F}^*)^T \end{bmatrix}$ and our scaffolding vectors for constructing the r-dimensional CVA biplot are the columns of

$$\begin{aligned} \boldsymbol{D}_r =& \left[\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{J-1}, \boldsymbol{d}_J, \boldsymbol{d}_{J+1}, \ldots, \boldsymbol{d}_r \right] \\ =& \boldsymbol{B} \left[\begin{bmatrix} \boldsymbol{I}_{J-1} \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{f}_1^* \end{bmatrix} \cdots \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{f}_{r-J+1}^* \end{bmatrix} \right]. \end{aligned} \tag{25}$$

Furthermore,

$$\boldsymbol{D}^{(r)} = \begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & (\boldsymbol{f}_1^*)^T \\ \cdots & \cdots \\ \boldsymbol{0}^T & (\boldsymbol{f}_{r-J+1}^*)^T \end{bmatrix} \boldsymbol{B}^{-1}. \tag{26}$$

The approximations of the rows of $\boldsymbol{X}$, i.e., the original samples, in the biplot constructed on the scaffolding provided by the columns of $\boldsymbol{D}_r$ follow from (16) and using (25) and (26) as

$$\begin{aligned} \hat{\boldsymbol{X}} =& \boldsymbol{X} \boldsymbol{D}_r \boldsymbol{D}^{(r)} \\ =& \boldsymbol{XB} \left[\begin{bmatrix} \boldsymbol{I}_{J-1} \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{f}_1^* \end{bmatrix} \cdots \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{f}_{r-J+1}^* \end{bmatrix} \right] \begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & (\boldsymbol{f}_1^*)^T \\ \cdots & \cdots \\ \boldsymbol{0}^T & (\boldsymbol{f}_{r-J+1}^*)^T \end{bmatrix} \boldsymbol{B}^{-1} \\ =& \boldsymbol{XB} \begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{f}_1^* (\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^* (\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^* (\boldsymbol{f}_{r-J+1}^*)^T \end{bmatrix} \boldsymbol{B}^{-1}, \end{aligned} \tag{27}$$

where, from the orthogonality of $\boldsymbol{F}^*$, it follows that $(\boldsymbol{f}_1^*)^T \boldsymbol{f}_1^* = (\boldsymbol{f}_2^*)^T \boldsymbol{f}_2^* = \cdots = (\boldsymbol{f}_{r-J+1}^*)^T \boldsymbol{f}_{r-J+1}^* = 1$.

The criterion *TSRES* now becomes

$$\begin{aligned}
&\left\|\boldsymbol{X}-\hat{\boldsymbol{X}}\right\|^2 \\
&\quad = tr\left\{(\boldsymbol{X}-\hat{\boldsymbol{X}})(\boldsymbol{X}-\hat{\boldsymbol{X}})^T\right\} \\
&\quad = \left\|\boldsymbol{X}\left(\boldsymbol{I}_p-\boldsymbol{B}\begin{bmatrix}\boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{f}_1^*(\boldsymbol{f}_1^*)^T+\boldsymbol{f}_2^*(\boldsymbol{f}_2^*)^T+\cdots+\boldsymbol{f}_{r-J+1}^*(\boldsymbol{f}_{r-J+1}^*)^T\end{bmatrix}\boldsymbol{B}^{-1}\right)\right\|^2.
\end{aligned} \tag{28}$$

To construct an r-dimensional CVA biplot satisfying our aim of minimizing *TSRES* while simultaneously providing 100% accurate predictions for the J sample means when $J < p$, we propose the following:

– Find the solution of

$$argmin\left\|\boldsymbol{X}\left(\boldsymbol{I}_p-\boldsymbol{B}\boldsymbol{L}\boldsymbol{B}^{-1}\right)\right\|^2, \tag{29}$$

where $\boldsymbol{L}=\begin{bmatrix}\boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{f}_1^*(\boldsymbol{f}_1^*)^T+\boldsymbol{f}_2^*(\boldsymbol{f}_2^*)^T+\cdots+\boldsymbol{f}_{r-J+1}^*(\boldsymbol{f}_{r-J+1}^*)^T\end{bmatrix}$ and the minimum is taken with respect to the $\boldsymbol{f}_j^*$ such that $(\boldsymbol{f}_j^*)^T\boldsymbol{f}_j^*=1$ for $j=1,2,\ldots,r-J+1$.

– Use the optimum $\{\boldsymbol{f}_1^*,\boldsymbol{f}_2^*,\ldots,\boldsymbol{f}_{r-J+1}^*\}$ to construct $(\boldsymbol{B}_{opt})_r=\left[\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{J-1},\boldsymbol{d}_J,\ldots,\boldsymbol{d}_r\right]$ where

$$\begin{aligned}
\boldsymbol{d}_j &= \boldsymbol{B}^*(\boldsymbol{f}_{j-J+1}^*)_{opt} \\
&= \boldsymbol{B}\begin{bmatrix}\boldsymbol{0} \\ (\boldsymbol{f}_{j-J+1}^*)_{opt}\end{bmatrix} \\
&= f_{J(j-J+1)}^{opt}\boldsymbol{b}_J+f_{(J+1)(j-J+1)}^{opt}\boldsymbol{b}_{J+1}+\cdots+f_{p(j-J+1)}^{opt}\boldsymbol{b}_p,
\end{aligned}$$

for $j=J,J+1,\ldots,r$.

– Next,$(\boldsymbol{B}_{opt})_r$ is used for constructing the r-dimensional CVA biplot with calibrated prediction (or interpolation) axes.

– Finally, calculate a standardised form of $min(TSRES)$: $\frac{min(TSRES(\boldsymbol{X},\hat{\boldsymbol{X}}))}{tr(\boldsymbol{X}\boldsymbol{X}^T)}$, as a measure of the accuracy of the approximations of the individual sample points in the r-dimensional biplot.

The solution for (29) can be found from (28) as follows: Since $\boldsymbol{B}^T\boldsymbol{S}_W\boldsymbol{B}=\boldsymbol{I}_p$ it follows that

$$\begin{aligned}
\boldsymbol{I}_p &= \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}-\boldsymbol{B}^T\overline{\boldsymbol{X}}^T\boldsymbol{N}_g\overline{\boldsymbol{X}}\boldsymbol{B} \\
&= \begin{bmatrix}(\boldsymbol{B}_{J-1})^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}_{J-1} & (\boldsymbol{B}_{J-1})^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}^* \\ (\boldsymbol{B}^*)^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}_{J-1} & (\boldsymbol{B}^*)^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}^*\end{bmatrix}
\end{aligned}$$

$$
\begin{aligned}
&- \begin{bmatrix} (\boldsymbol{B}_{J-1})^T \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} \boldsymbol{B}_{J-1} & (\boldsymbol{B}_{J-1})^T \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} \boldsymbol{B}^* \\ (\boldsymbol{B}^*)^T \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} \boldsymbol{B}_{J-1} & (\boldsymbol{B}^*)^T \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} \boldsymbol{B}^* \end{bmatrix} \\
&= \begin{bmatrix} (\boldsymbol{B}_{J-1})^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{B}_{J-1} & (\boldsymbol{B}_{J-1})^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{B}^* \\ (\boldsymbol{B}^*)^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{B}_{J-1} & (\boldsymbol{B}^*)^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{B}^* \end{bmatrix} - \begin{bmatrix} (\boldsymbol{B}_{J-1})^T \overline{\boldsymbol{X}}^T \boldsymbol{N}_g \overline{\boldsymbol{X}} \boldsymbol{B}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},
\end{aligned}
$$

establishing that

$$
(\boldsymbol{B}^*)^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{B}^* = \boldsymbol{I}_{p-J+1}. \tag{30}
$$

Since $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{J-1} & \boldsymbol{B}^* \end{bmatrix}$ and non-singular, we set

$$
\boldsymbol{B}^{-1} = \begin{bmatrix} \boldsymbol{B}^{(J-1)} : (J-1) \times p \\ \boldsymbol{B}^{(2)} : (p-J+1) \times p \end{bmatrix}.
$$

Write $\boldsymbol{U} : (p-J+1) \times (p-J+1) = \boldsymbol{f}_1^* (\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^* (\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^* (\boldsymbol{f}_{r-J+1}^*)^T$; then it follows that

$$
\begin{aligned}
&\boldsymbol{X}\left(\boldsymbol{I}_p - \boldsymbol{B} \begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{f}_1^* (\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^* (\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^* (\boldsymbol{f}_{r-J+1}^*)^T \end{bmatrix} \boldsymbol{B}^{-1} \right) \\
&= \boldsymbol{X}\left(\boldsymbol{I}_p - \left(\boldsymbol{B}_{J-1} \boldsymbol{B}^{(J-1)} + \boldsymbol{B}^* \boldsymbol{U} \boldsymbol{B}^{(2)} \right) \right).
\end{aligned} \tag{31}
$$

From $\begin{bmatrix} \boldsymbol{B}_{J-1} & \boldsymbol{B}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^{(J-1)} \\ \boldsymbol{B}^{(2)} \end{bmatrix} = \boldsymbol{I}_p$ it follows that $\boldsymbol{B}_{J-1} \boldsymbol{B}^{(J-1)} = \boldsymbol{I}_p - \boldsymbol{B}^* \boldsymbol{B}^{(2)}$ so that (31) can be written as

$$
\boldsymbol{X}\left(\boldsymbol{I}_p - \left(\boldsymbol{B}_{J-1} \boldsymbol{B}^{(J-1)} + \boldsymbol{B}^* \boldsymbol{U} \boldsymbol{B}^{(2)} \right) \right) = \boldsymbol{X}\left(\boldsymbol{B}^* \boldsymbol{B}^{(2)} - \boldsymbol{B}^* \boldsymbol{U} \boldsymbol{B}^{(2)} \right). \tag{32}
$$

Therefore,

$$
\begin{aligned}
&\left\| \boldsymbol{X}\left(\boldsymbol{I}_p - \boldsymbol{B} \begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{f}_1^* (\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^* (\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^* (\boldsymbol{f}_{r-J+1}^*)^T \end{bmatrix} \boldsymbol{B}^{-1} \right) \right\|^2 \\
&= \left\| \boldsymbol{X} \boldsymbol{B}^* (\boldsymbol{I}_{p-J+1} - \boldsymbol{U}) \boldsymbol{B}^{(2)} \right\|^2 \\
&= tr\left\{ \boldsymbol{X} \boldsymbol{B}^* (\boldsymbol{I}_{p-J+1} - \boldsymbol{U}) \boldsymbol{B}^{(2)} (\boldsymbol{B}^{(2)})^T (\boldsymbol{I}_{p-J+1} - \boldsymbol{U}) (\boldsymbol{B}^*)^T \boldsymbol{X}^T \right\} \\
&= tr\left\{ \boldsymbol{X} \boldsymbol{B}^* (\boldsymbol{B}^{(2)})^T (\boldsymbol{B}^*)^T \boldsymbol{X}^T \right\} - 2tr\left\{ \boldsymbol{X} \boldsymbol{B}^* \boldsymbol{U} \boldsymbol{B}^{(2)} (\boldsymbol{B}^{(2)})^T (\boldsymbol{B}^*)^T \boldsymbol{X}^T \right\} \\
&\quad + tr\left\{ \boldsymbol{X} \boldsymbol{B}^* \boldsymbol{U} (\boldsymbol{B}^{(2)})^T \boldsymbol{U} (\boldsymbol{B}^*)^T \boldsymbol{X}^T \right\}
\end{aligned} \tag{33}
$$

$$
= tr\left\{ \boldsymbol{B}^{(2)} (\boldsymbol{B}^{(2)})^T \right\} - 2tr\left\{ \boldsymbol{U} \boldsymbol{B}^{(2)} (\boldsymbol{B}^{(2)})^T \right\} + tr\left\{ \boldsymbol{U} \boldsymbol{B}^{(2)} (\boldsymbol{B}^{(2)})^T \boldsymbol{U} \right\} \tag{34}
$$

by substituting (30) in (33). Write $\boldsymbol{H} = \boldsymbol{B}^{(2)}(\boldsymbol{B}^{(2)})^T : (p-J+1)\times(p-J+1)$ then it follows that $\boldsymbol{H}$ has rank $(p-J+1)$ and is thus positive definite. Therefore,

$$\left\| \boldsymbol{X}\left(\boldsymbol{I}_p - \boldsymbol{B}\begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{f}_1^*(\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^*(\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^*(\boldsymbol{f}_{r-J+1}^*)^T \end{bmatrix}\boldsymbol{B}^{-1}\right)\right\|^2$$
$$= tr\{\boldsymbol{H}\} - 2tr\{\boldsymbol{U}\boldsymbol{H}\} + tr\{\boldsymbol{U}\boldsymbol{H}\boldsymbol{U}\}, \tag{35}$$

where

$$tr\{\boldsymbol{U}\boldsymbol{H}\} = (\boldsymbol{f}_1^*)^T\boldsymbol{H}\boldsymbol{f}_1^* + (\boldsymbol{f}_2^*)^T\boldsymbol{H}\boldsymbol{f}_2^* + \cdots + (\boldsymbol{f}_{r-J+1}^*)^T\boldsymbol{H}\boldsymbol{f}_{r-J+1}^* \tag{36}$$

and

$$\begin{aligned} tr\{\boldsymbol{U}\boldsymbol{H}\boldsymbol{U}\} &= tr\{\boldsymbol{U}\boldsymbol{U}\boldsymbol{H}\} \\ &= tr\Big\{\Big(\boldsymbol{f}_1^*(\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^*(\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^*(\boldsymbol{f}_{r-J+1}^*)^T\Big) \\ &\quad \times \Big(\boldsymbol{f}_1^*(\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^*(\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^*(\boldsymbol{f}_{r-J+1}^*)^T\Big)\boldsymbol{H}\Big\}. \end{aligned}$$

Since $(\boldsymbol{f}_i^*)^T\boldsymbol{f}_j^* = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases}$ it follows that

$$\begin{aligned} tr\{\boldsymbol{U}\boldsymbol{H}\boldsymbol{U}\} &= tr\Big\{\Big(\boldsymbol{f}_1^*(\boldsymbol{f}_1^*)^T\boldsymbol{H} + \boldsymbol{f}_2^*(\boldsymbol{f}_2^*)^T\boldsymbol{H} + \cdots + \boldsymbol{f}_{r-J+1}^*(\boldsymbol{f}_{r-J+1}^*)^T\boldsymbol{H}\Big)\Big\} \\ &= (\boldsymbol{f}_1^*)^T\boldsymbol{H}\boldsymbol{f}_1^* + (\boldsymbol{f}_2^*)^T\boldsymbol{H}\boldsymbol{f}_2^* + \cdots + (\boldsymbol{f}_{r-J+1}^*)^T\boldsymbol{H}\boldsymbol{f}_{r-J+1}^*. \end{aligned} \tag{37}$$

Substituting (36) and (37) into (35) leads to

$$\left\| \boldsymbol{X}\left(\boldsymbol{I}_p - \boldsymbol{B}\begin{bmatrix} \boldsymbol{I}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{f}_1^*(\boldsymbol{f}_1^*)^T + \boldsymbol{f}_2^*(\boldsymbol{f}_2^*)^T + \cdots + \boldsymbol{f}_{r-J+1}^*(\boldsymbol{f}_{r-J+1}^*)^T \end{bmatrix}\boldsymbol{B}^{-1}\right)\right\|^2$$
$$= tr\{\boldsymbol{H}\} - \big((\boldsymbol{f}_1^*)^T\boldsymbol{H}\boldsymbol{f}_1^* + (\boldsymbol{f}_2^*)^T\boldsymbol{H}\boldsymbol{f}_2^* + \cdots + (\boldsymbol{f}_{r-J+1}^*)^T\boldsymbol{H}\boldsymbol{f}_{r-J+1}^*\big). \tag{38}$$

Remembering that $\boldsymbol{H}$ is positive definite, the criterion (38) can be minimized by maximizing each of the terms $(\boldsymbol{f}_j^*)^T\boldsymbol{H}\boldsymbol{f}_j^*$ with respect to $\boldsymbol{f}_j^*$ for $j = 1, 2, \ldots, r-J+1$ under the constraint $(\boldsymbol{f}_j^*)^T\boldsymbol{f}_j^* = 1$. This is readily accomplished by introducing the Lagrange multiplier λ_j to form

$$(\boldsymbol{f}_j^*)^T\boldsymbol{H}\boldsymbol{f}_j^* - \lambda_j\big((\boldsymbol{f}_j^*)^T\boldsymbol{f}_j^* - 1\big). \tag{39}$$

Differentiating (39) with respect to λ_j and equating to zero leads to $\boldsymbol{H}\boldsymbol{f}_j^* = \lambda_j\boldsymbol{f}_j^*$ i.e., $\lambda_j = (\boldsymbol{f}_j^*)^T\boldsymbol{H}\boldsymbol{f}_j^*$. Thus $(\boldsymbol{f}_j^*)^T\boldsymbol{H}\boldsymbol{f}_j^*$ is maximized when $\boldsymbol{f}_j^*$ is a normalized eigenvector associated with the jth largest eigenvalue of $\boldsymbol{H}$ and hence the criterion (39) attains its minimum when the $\{\boldsymbol{f}_j^*\}$ are set to the normalized eigenvectors associated with the largest $r-J+1$ eigenvalues of $\boldsymbol{H} = \boldsymbol{B}^{(2)}(\boldsymbol{B}^{(2)})^T$, respectively. Denote

these eigenvectors by $\boldsymbol{f}_1^{opt}, \boldsymbol{f}_2^{opt}, \cdots, \boldsymbol{f}_{r-J+1}^{opt}$, respectively, where $r \leq p$ then a $p \times r$ matrix $(\boldsymbol{B}_{opt})_r$ can be constructed as

$$(\boldsymbol{B}_{opt})_r = \left[\boldsymbol{B}_{J-1}, \boldsymbol{B}^*\left[\boldsymbol{f}_1^{opt}, \boldsymbol{f}_2^{opt}, \cdots, \boldsymbol{f}_{r-J+1}^{opt}\right]\right]. \tag{40}$$

Setting $r = p$ in the above leads to a matrix $\boldsymbol{B}_{opt}$ of size $p \times p$ which is non-singular, allowing for the computation of the matrices $(\boldsymbol{B}_{opt})_r$, consisting of the first r columns of $\boldsymbol{B}_{opt}$, and $(\boldsymbol{B}_{opt})^{(r)}$, consisting of the first r rows of $(\boldsymbol{B}_{opt})^{-1}$. Therefore

$$\hat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{D}_r\boldsymbol{D}^{(r)} = \boldsymbol{X}(\boldsymbol{B}_{opt})_r(\boldsymbol{B}_{opt})^{(r)} \tag{41}$$

will minimize *TSRES*.

6 An alternative biplot based on the Bhattacharyya distance for the two sample case

If, analogous to section 3 the population parameters in (5) are replaced with their sample estimates the sample version of the Bhattacharyya distance consists of two terms. The first measures the dissimilarity between the two sample means and the second measures the dissimilarity between the two sample covariance matrices. Fukunaga (1990) uses this property to construct a second dimension for a visual display of the rows of a data matrix $\boldsymbol{X} : n \times p$ in two dimensions. Furthermore, Hennig (2004) utilizes the Bhattacharyya distance, among other methods to construct two-dimensional visualizations of the dispersions of two asymmetric samples – asymmetric in the sense that one is known to be more homogeneous and the other to be more heterogeneous. However, it should be noted that none of the visualizations proposed by Fukunaga (1990) and Hennig (2004) are biplots because no information regarding the columns of $\boldsymbol{X}$ is displayed. The optimal two-group CVA biplot proposed in section 5 assumes equality of covariance matrices as is usual for CVA. This implies that the second term of the sample version of (5) will vanish and optimization of (5) becomes equivalent to maximizing (9).

Denote the transformation to the canonical space $\mathcal{C}$ by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{M}$ with $\boldsymbol{M} : p \times p$. We can write

$$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{V},$$

where $\mathcal{C}_1$ is one-dimensional based on $\boldsymbol{m}_1$ and $\mathcal{V}$ is $(p-1)$-dimensional with basis $\boldsymbol{m}_2, \boldsymbol{m}_3, \ldots, \boldsymbol{m}_p$. In $\mathcal{V}$ we have

$$\begin{aligned}
&\boldsymbol{Y}^{*T} = [Y_2, Y_3, \ldots, Y_p],\\
&\boldsymbol{Y}^*|G_i : (p-1) \times 1 \sim (\boldsymbol{\mu}^*, \boldsymbol{\Sigma}_i^*), i = 1; 2,
\end{aligned}$$

with $\boldsymbol{\mu}^{*T} = [\mu_2^{(Y)}, \mu_3^{(Y)}, \ldots, \mu_p^{(Y)}]$, $\boldsymbol{\Sigma}_1^* = \boldsymbol{M}^{*T}\boldsymbol{\Sigma}_1\boldsymbol{M}^*$, $\boldsymbol{\Sigma}_2^* = \boldsymbol{M}^{*T}\boldsymbol{\Sigma}_2\boldsymbol{M}^*$ and $\boldsymbol{M}^* : p \times (p-1) = [\boldsymbol{m}_2, \boldsymbol{m}_3, \ldots, \boldsymbol{m}_p]$.

Since $\boldsymbol{\mu}_1^* = \boldsymbol{\mu}_2^* = \boldsymbol{\mu}^*$, (say) we have from (5)

$$D_{Bhat} = \frac{1}{2} log \left(\frac{\left| \frac{\boldsymbol{\Sigma}_1^* + \boldsymbol{\Sigma}_2^*}{2} \right|}{\sqrt{|\boldsymbol{\Sigma}_1^*||\boldsymbol{\Sigma}_2^*|}} \right). \tag{42}$$

Maximization of the sample version of (42) proceeds parallel to the process described in section 2.3. The outcome is the matrix $\boldsymbol{A} : (p-1) \times (p-1)$ containing as columns the required eigenvectors arranged in decreasing order of the values of $\lambda_i^* + \frac{1}{\lambda_i^*} + 2$.

A 2D biplot can now be constructed using the methods described in sections 3 and 4 by first noting that the matrix $\boldsymbol{M}$ is available as the matrix $\boldsymbol{B}$ of section 3. Next, we calculate the matrix

$$\boldsymbol{K} : p \times p = [\boldsymbol{m}_1 \quad \boldsymbol{M}^* \boldsymbol{A}]$$

and its inverse $\boldsymbol{K}^{-1}$. Let $\boldsymbol{a} : (p-1) \times 1$ denote the first column of $\boldsymbol{A}$, then the 2D biplot can be constructed as described in section 4 using the rows of $\boldsymbol{Z} : n \times 2$, where

$$\boldsymbol{Z} = [\boldsymbol{X}\boldsymbol{m}_1 \quad \boldsymbol{Y}^* \boldsymbol{a}] = \boldsymbol{X}[\boldsymbol{m}_1 \quad \boldsymbol{M}^* \boldsymbol{a}]$$

for plotting the samples. The variables are represented by p calibrated biplot axes constructed from

$$\frac{marker}{\boldsymbol{e}_k^T \boldsymbol{K}^{(2)T} \boldsymbol{K}^{(2)} \boldsymbol{e}_k} \boldsymbol{K}^{(2)} \boldsymbol{e}_k,$$

where $\boldsymbol{K}^{(2)}$ denotes the first 2 rows of $\boldsymbol{K}^{-1}$.

7 Examples

As an example, we consider the Vertebral Column Data Set from the UCI Machine Learning Repository (Barreto et al. 2011) and discussed in detail by da Rocha Neto et al. (2011). The full data set contains measurements on six continuous/numeric variables relating to the shape and orientation of the pelvis and lumbar spine for each of the 310 individuals (samples). These samples are classified as normal, disk hernia, or spondylolisthesis patients. As an example of a two-group CVA, we study the subset of 60 disk hernia and 150 spondylolisthesis patients. The six numeric variables are Pelvic incidence, Pelvic tilt, Lumbar lordosis angle, Sacral slope, Pelvic radius, and Degree spondylolisthesis.

Table 1 contains the group means for each of the six variables - with the outlier (see Fig. 2) included and excluded. It is hard to see from the table the presence of the outlier but, as is evident from Figs. 1 and 2, the outlier will be clearly revealed in a CVA biplot.

The one-dimensional biplot of this two-group data set is shown in Fig. 1. In this biplot, the two group means as well as all six variables (in the form of six calibrated axes) lie on the one scaffolding line defined by the singular vector associated with

Table 1 Group means for the six measurements V1 = Pelvic incidence, V2 = Pelvic tilt, V3 = Lumbar lordosis angle, V4 = Sacral slope, V5 = Pelvic radius and V6 = Degree spondylolisthesis

	Disk Hernia	Spondylolisthesis	
		Outlier included	Outlier excluded
V1	47.6383	71.5137	71.1223
V2	17.3987	20.7480	20.8309
V3	35.4635	64.1099	64.2154
V4	30.2400	50.7661	50.2919
V5	116.4750	114.5183	114.5642
V6	2.4793	51.8969	49.4362

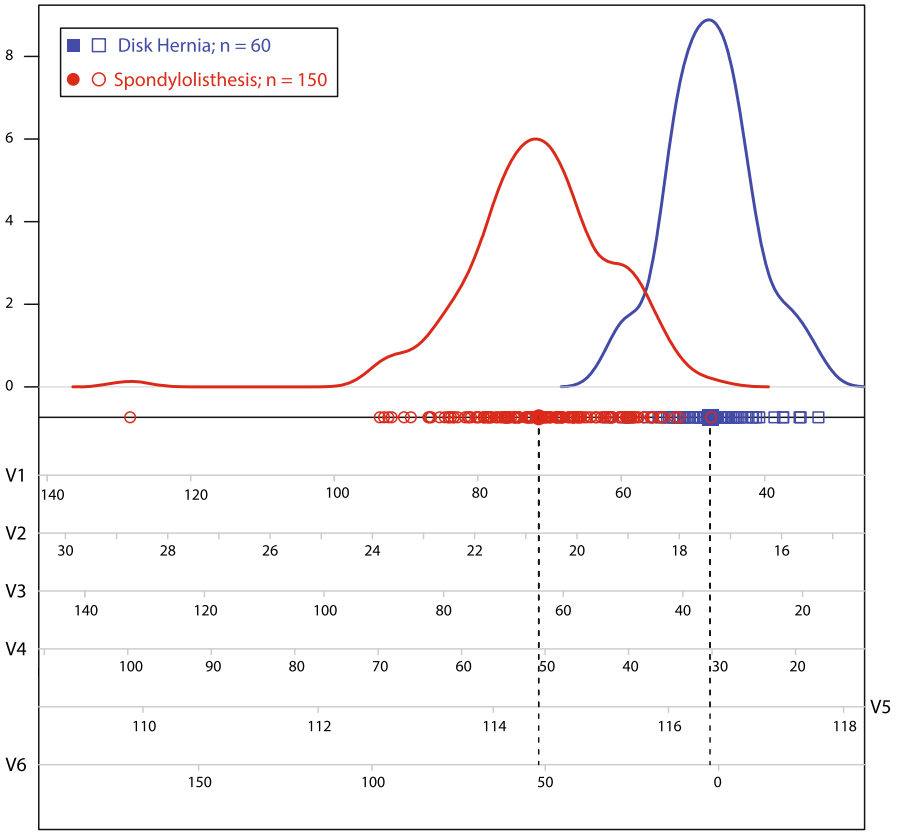

Fig. 1 1D biplot of the two-group Vertebral Column data. V1 = Pelvic incidence, V2 = Pelvic tilt, V3 = Lumbar lordosis angle, V4 = Sacral slope, V5 = Pelvic radius, V6 = Degree spondylolisthesis. The one-dimensional biplot is enhanced by superimposing density estimates for the interpolated sample points in the two groups

the single non-zero singular value of the underlying two-sided eigenequation. The six calibrated axes representing the variables have been vertically translated to aid the interpretation of the biplot. The values for each of the group means for all six variables can easily be read from the biplot axes and it can be verified that these values coincide exactly with the corresponding values in Table 1. As expected all variables have predictivities of 100% for determining the mean values with $TSREM = 0$. All the individual sample points have been interpolated onto the bipot and therefore they also lie on the single scaffolding axis defining the biplot. However, $TSRES > 0$ with a standardized version $TSRES/(tr(\boldsymbol{X}\boldsymbol{X}^T) = 0.4702$. To visualize the within groups variation the biplot has been enhanced by the addition of density estimates of the two sets of sample points interpolated onto the one-dimensional CVA biplot. These density estimates show graphically the separation/overlap of the two groups. Inspection of the six biplot axes suggests V5 to be negatively correlated with the other five variables; the latter are all pairwise positively correlated.

It is clear that much can be learned from the one-dimensional CVA biplot but some serious issues are calling for considering a second scaffolding axis:

- Is the conspicuous outlier an outlier on all variables?
- Can the approximation of the sample points be improved without sacrificing what has been achieved with the mean vectors?
- Is it possible to construct a more detailed visualization of the separation/overlap of the two groups?
- Is it possible to construct a more accurate and detailed visualization of the correlation structure of the six variables?

Figure 2 provides an answer to the above issues. The optimal two-dimensional CVA biplot in Fig. 2 demonstrates the following:

- The biplot is uniquely defined.
- Each of the six biplot axes has a predictivity of 100% for determining the values of the two groups for all variables.This results in *TSREM* to remain zero. Figure 2 provides a clearer picture of how the predictivities are determined.
- Introducing the second scaffolding axis improves the approximation of the samples appreciably: the standardized *TSRES* decreases to 0.1799 (a decrease of more than 60% of the corresponding value obtained in Fig. 2).
- The second scaffolding axis is optimal in the sense that no other scaffolding axis can be found, which will improve *TSRES* while restricting *TSREM* to zero.
- It is clear that Sample 116 is less of an outlier on V3 and V4 than on V2 and V6.
- There is a suggestion that while V5 appears to be negatively correlated with V3 and V4 it appears to be positively correlated with V2 and V6. We note that the addition of a second scaffolding axis provides angles between the biplot axes, which allow for visualizing the approximate correlational structure between the variables.

The biplot in Fig. 2 has been enhanced by superimposing 95%-bags onto the biplot. Alpha-bags are discussed in detail by Gower et al. (2011). The 95%-bag used here contains the innermost 95% of the bivariate sample points, where the innermost is relative to the Tukey median (Ruts and Rousseeuw 1996). Now, we are ready for a detailed appraisal of the overlap/separation of the two groups based on the two-dimensional

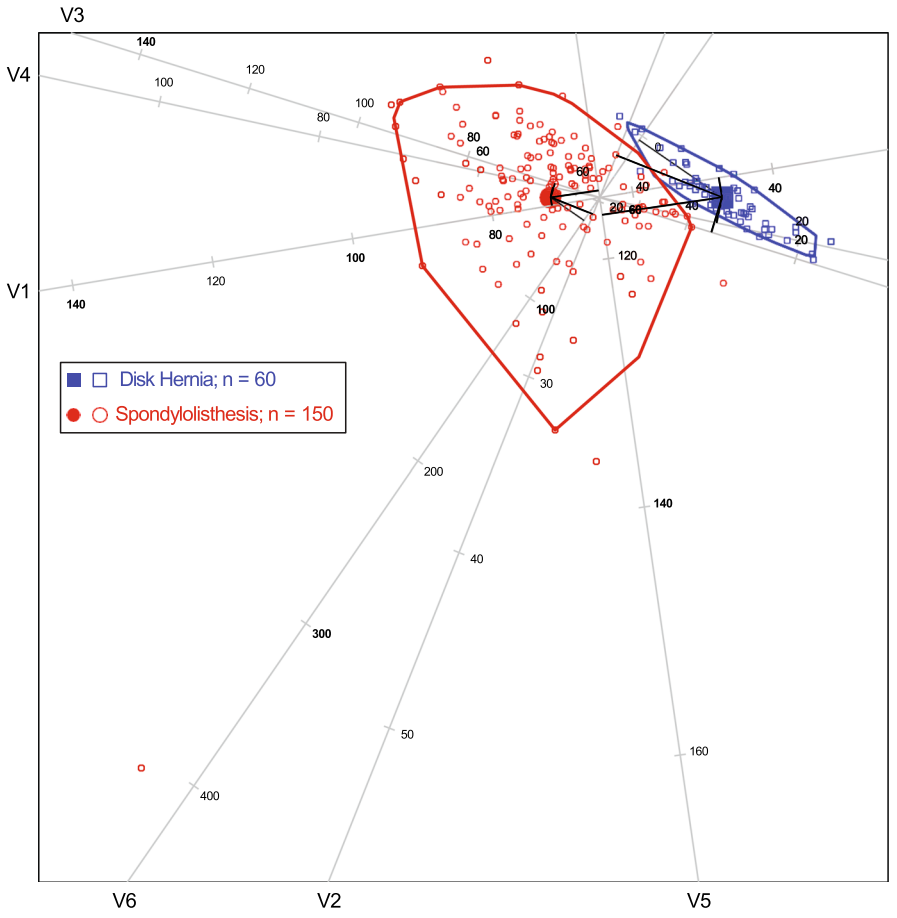

Fig. 2 Optimal 2D CVA biplot of the two-group Vertebral Column data. V1 = Pelvic incidence, V2 = Pelvic tilt, V3 = Lumbar lordosis angle, V4 = Sacral slope, V5 = Pelvic radius, V6 = Degree spondylolisthesis. The optimal two-dimensional biplot is enhanced by superimposing bags containing the innermost 95% samples of the two groups respectively

clouds of points visualizing the within groups sample variation. However, we first exclude Sample 116 from the analysis and consider the optimal two-dimensional CVA biplot given in Fig. 3 overlaid with 95%-bags. This figure shows:

- Clearly how the biplot axes are used to determine the group means exactly for each variable.
- The angles between the biplot axes allow for an approximate visual appraisal of the correlation structure.
- It is seen that the two 95%-bags almost touch each other but do not overlap giving us an overall quantitative measure of the degree of separation between the two groups.
- The standardized *TSRES* value is 0.2155.
- Although the two clouds of points have a high degree of separation it can also be seen that

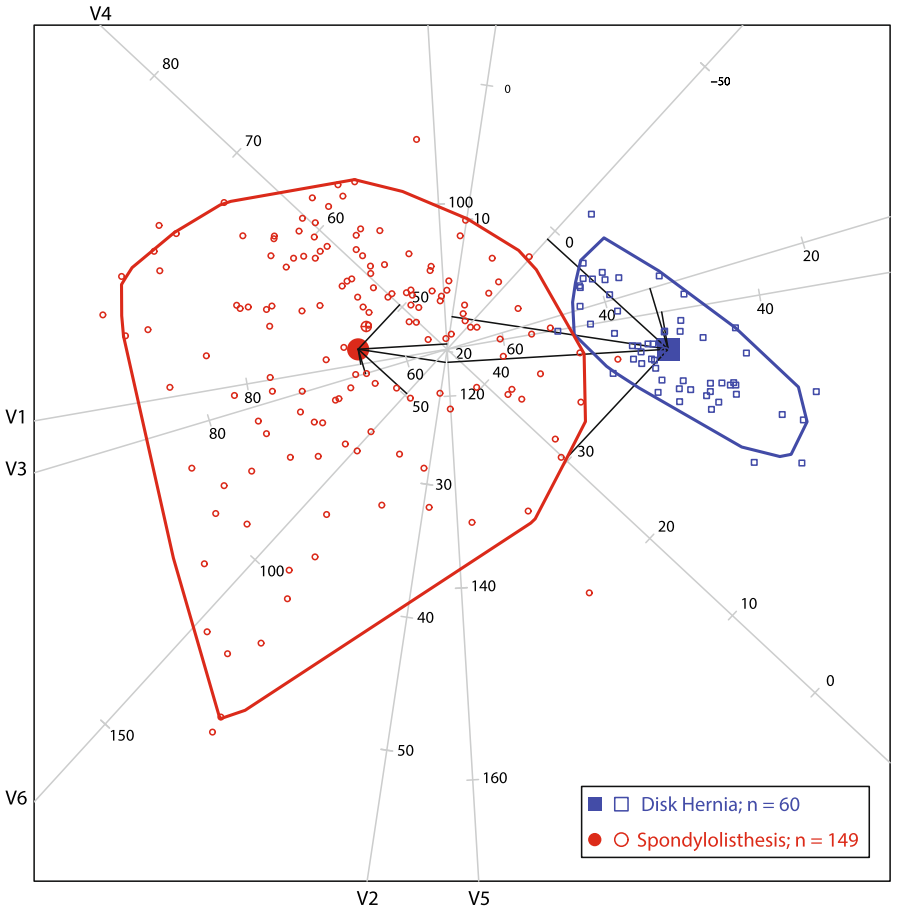

Fig. 3 Optimal 2D CVA biplot of the two-group Vertebral Column data excluding outlier sample 116. V1 = Pelvic incidence, V2 = Pelvic tilt, V3 = Lumbar lordosis angle, V4 = Sacral slope, V5 = Pelvic radius, V6 = Degree spondylolisthesis. The optimal two-dimensional biplot is enhanced by superimposing bags containing the innermost 95% samples of the two groups respectively

– there is a high degree of overlap between the two groups concerning V2 and V5;
– there is almost no overlap on V1, V3 and V6;
– on V4 large Disk Hernia values overlap with small Spondylolisthesis measurements, while small measurements of V4 almost exclusively occur in Disk Hernia.

The CVA biplots in Figs. 2 and 3 assume equal covariance matrices. We can now relax this assumption and construct in Fig. 4, the biplot based on the Bhattacharyya distance as discussed in section 6.

Although the appearance of this biplot is quite similar to that of the corresponding optimal CVA biplot shown in Fig. 3 its standardized *TSRES* value is approximately 10% higher, namely 0.2367.

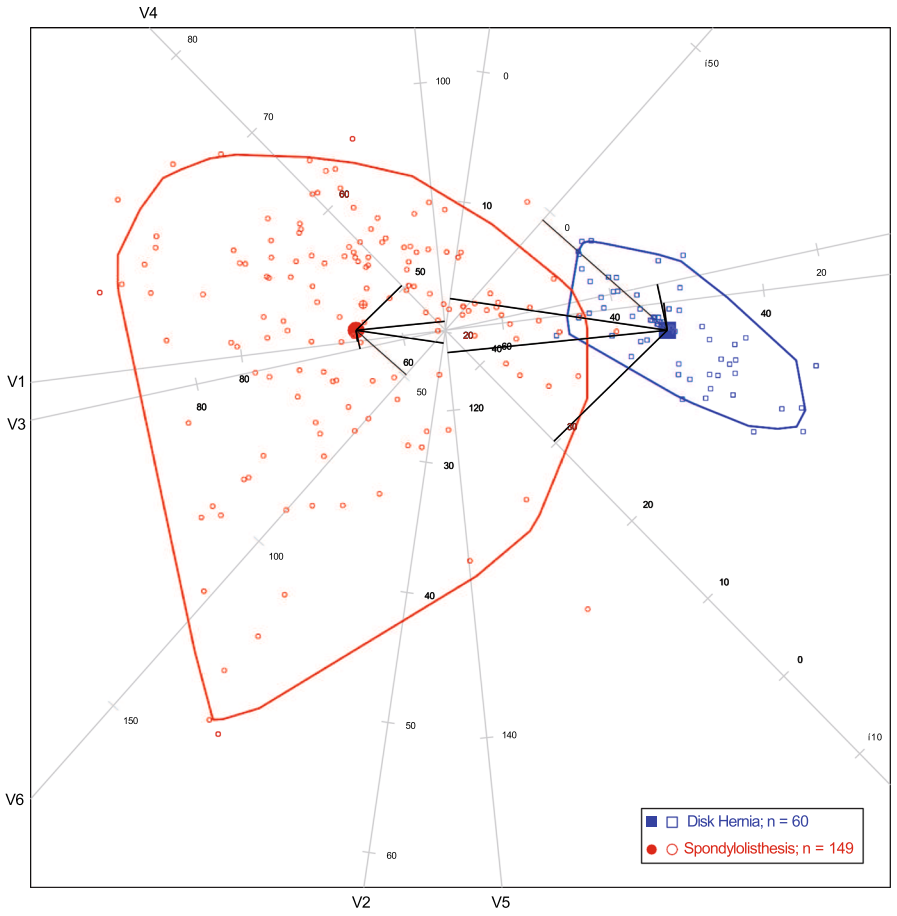

Fig. 4 The 2D biplot based on the Bhattacharyya distance of the two-group Vertebral Column data excluding outlier sample 116. V1 = Pelvic incidence, V2 = Pelvic tilt, V3 = Lumbar lordosis angle, V4 = Sacral slope, V5 = Pelvic radius, V6 = Degree spondylolisthesis. The optimal two-dimensional biplot is enhanced by superimposing bags containing the innermost 95% samples of the two groups respectively

8 Conclusions

CVA biplots are constructed to show in a single plot the group means as points and all the variables as calibrated linear biplot axes. In the case of two groups the CVA biplot becomes a line containing all these points and biplot axes. Furthermore, there is no approximation in the positions of the group means and their respective values, which can be exactly determined from the biplot axes. It is common practice to interpolate the individual sample points onto the CVA biplot as well, but then they appear as approximations in the one-dimensional CVA biplot space. Since all the biplot axes lie on top of each other it is difficult to use them for determining the values of the two means for the different variables. However, the vertical translation of one of these axes

does not affect the values of the two means for that particular variable. Therefore, as has been shown in Fig. 1, vertical translation of the biplot axes does not change the dimensionality of the CVA biplot but increases the usefulness of the different biplot axes for reading off values of the respective variables.

The fundamental question that is addressed in this paper is: What can be gained, if anything, by increasing the dimensionality of the above one-dimensional CVA biplot to two dimensions? This question can be rephrased as: Can we add a second dimension to our one-dimensional CVA biplot to improve the approximation of the individual sample points leading to a better understanding of the within groups variability while leaving unchanged the optimal representation of the two group means? As it turned out the addition of a second dimension is not a straightforward process since there are an infinite number of ways that this can be done. Furthermore, if existing software is used for constructing a two-dimensional CVA biplot in the two-group case the result can be highly misleading. This is because the two-sided eigenequation underlying the CVA procedure has only a single non-zero singular value with no natural ordering of the zero singular values resulting in the indeterminacy of singular vectors associated with zero singular values. Therefore, to guarantee a unique solution for finding a second dimension, we had to consider a criterion to optimize the approximation of the sample points while leaving the optimal representation of the group means unchanged. We suggested the *TSRES* criterion for meeting this goal. Minimizing *TSRES* results in a uniquely defined two-dimensional CVA biplot for the two-group case. It optimizes the approximations of the within-group samples while the two group means are exactly represented with the linear biplot axes providing exact values for both groups on all variables.

Figure 3 shows that our proposed optimal two-dimensional CVA biplot for two groups has the potential to provide the researcher with a tool that not only distinguishes the group means optimally but also where the within sample variation can be depicted graphically to gain deeper insight into the separation/overlap of the two groups. Moreover, it is unique and thus prevents any possibility of ambiguity when routinely using existing software. Thus we have attained our primary objective as is illustrated in the example discussed above.

The algebra underlying the optimal two-dimensional CVA biplot extends directly to an optimal three-dimensional biplot when dealing with a three-groups case having a CVA biplot space of dimension two.

An alternative suggestion based on the Bhattacharyya distance is available when relaxing the equal within-group covariance matrix assumption. As can be seen from Fig. 4, the biplot is slightly different, but the overall conclusion regarding overlap and separation in terms of the individual variables remains unchanged. However, the biplot based on Bhattacharyya distance is designed to optimize a different objective than the optimization of the sum of the squared approximations of the data matrix. Therefore, our preferred 2D CVA biplot to construct in the two-group case is the proposed optimal CVA biplot.

Finally, we note that our R code for constructing the biplots discussed in this paper is available in Lubbe et al. (2023).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-024-00593-7.

Acknowledgements We are grateful to the comments made by an anonymous reviewer and the review editor that have improved the quality of this manuscript.

Funding Open access funding provided by Stellenbosch University. This work is based upon research supported in part by the National Research Foundation (NRF) of South Africa (Grant Number 103310). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors, and therefore the NRF does not accept any liability in regard thereof.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Barreto GdA, da Rocha Neto AR, da Mota Filho HAF (2011) Vertebral column data set. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals of Eugencs 7:179–188

Flury B (1997) A first course in multivariate statistics. Springer-Verlag, New York

Flury L, Boukai B, Flury BD (1997) The discrimination subspace model. J American Stat Assoc 92:758–766

Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press, Boston

Gabriel KR (1971) The biplot graphical display of matrices with application to principal component analysis. Biometrika 58:451–467

Gabriel KR (1972) Analysis of meteorological mata by means of manonical decomposition and biplots. J Appl Meteorol 11:1071–1077

Gardner-Lubbe S, Le Roux NJ, Gower JC (2008) Measures of fit in principal component and canonical variate analyses. J Appl Stat 35:947–965

Gittins R (1985) Canonical analysis. Springer-Verlag, Berlin

Gower JC (1995) A general theory of biplots. In: Krzanowski WJ (ed) Recent advances in descriptive multivariate analysis. Clarendon Press, Oxford, pp 283–303

Gower JC, Hand DJ (1996) Biplots. Chapman & Hall, London

Gower JC, Le Roux NJ, Lubbe S (2011) Understanding biplots. Wiley, Chichester

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer-Verlag, New York

Hennig C (2004) Asymmetric linear dimension reduction for classification. J Comput Graph Stat 13:930–945

Lubbe S, Le Roux N, Nienkemper-Swanepoel J, Ganey R, Van der Merwe C (2023) biplotEZ: EZ-to-Use Biplots. https://CRAN.R-project.org/package=biplotEZ, r package version 1.2.0

McLachlan GJ (1992) Discriminant analysis and statistical pattern recognition. John Wiley, New York

Rao CR (1948) The utilization of multiple measurements in problems of biological classification. J Royal Stat Soc , Series B 10:159–193

da Rocha Neto AR, Sousa R, Barreto GdA, Cardoso JS (2011) Diagnostic of pathology on the vertebral column with embedded reject option. In: Iberian Conference on Pattern Recognition and Image Analysis, Springer, pp 588–595

Ruts I, Rousseeuw PJ (1996) Computing depth contours of bivariate point clouds. Comput Stat Data Anal 23:153–168
Tukey JW (1975) Mathematics and the picturing of data. Proc Int Congress Math 2:523–531