Volume 21 (2), pp. 111–124
http://www.orssa.org.za

ORiON
ISSN 0529-191-X
©2005

Extending a scatterplot for displaying group structure in multivariate data: A case study

S Gardner[*] NJ Le Roux[†] T Rypstra[‡] JPJ Swart[‡]

Received: 3 March 2005; Revised: 31 May 2005; Accepted: 6 June 2005

Abstract

The power of canonical variate analysis (CVA) biplots, when regarded as extensions of ordinary scatterplots to describe variation and group structure in multivariate observations, is demonstrated by presenting a case study from the South African wood pulp industry. It is shown how multidimensional standards specified by users of a product may be added to the biplot in the form of acceptance regions such that the roles of the respective variables that influence the product can be ascertained. The case study considers an alternative to CVA and multivariate analysis of variance (MANOVA) when the application of these procedures becomes questionable as a result of dealing with small sample sizes and heterogeneity of covariance matrices. It is explained how analysis of distance (AOD) analogous to analysis of variance may be performed in such cases. Biplots to accompany AOD are provided. The biplots and AOD illustrated in the case study from the wood pulp industry have the potential to be used widely where a primary product, influenced by several variables, is produced and where this product is of importance to various secondary manufacturers depending on which set of multidimensional specifications are met.

Key words: Analysis of distance, biplot, canonical variate analysis, multidimensional scaling, multivariate analysis of variance, permutation tests, wood pulp.

1 Introduction

The identification, description and separation of different groups are of primary importance in diverse fields of human activity. In risk management, for example, a financial institution is awarded a credibility rating of A, B, C or D according to various financial indicators; in car insurance clients are divided into different risk groups on the basis of criteria such as age, sex, residential area and type of car; archaeological artefacts are ascribed to different

[*] Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland, 7602, South Africa

[†] Corresponding author: Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland, 7602, South Africa, email: `njlr@sun.ac.za`

[‡] Department of Forest and Wood Science, University of Stellenbosch, Stellenbosch, Private Bag X1, Matieland, 7602, South Africa

time periods according to various properties measured; a manufacturer of a consumer product seeks advice on differences in personal characteristics of clients buying its different brands. It follows from the above examples that several properties of any product or process can simultaneously determine group membership.

Several multivariate statistical techniques have been developed to address the problem of finding groups in data (see for example Kaufman and Rousseeuw (1990), McLachlan (1992), Hastie *et al.* (2001) and Johnson and Wichern (2002)). In addition to providing the necessary algebraic derivation for optimally separating the groups according to statistically defined criteria these techniques consist of accompanying graphics for displaying the results of the analysis. Canonical variate analysis (CVA), closely related to linear discriminant analysis, is a popular technique for separating groups by finding those linear combinations of the variables that optimise the ratio of the between group variation to the within group variation.

The aims of this paper are twofold: firstly, we illustrate how graphical displays accompanying a CVA may be enhanced by extending the well-known scatterplot of two variables to any number of variables. These enhancements include the display of predefined groups together with information of all variables contributing to group separation, as well as product specifications, as stipulated by a user. Secondly, the aim is to point out how to perform a statistical analysis together with the necessary graphical displays when the assumptions underlying a CVA are not met. In order to pursue these aims a case study from the South African wood pulp industry is presented. In the next section a CVA biplot is introduced for the simultaneous display of eight characteristics measured for five pine species. This biplot is related to a multivariate analysis of variance (MANOVA) performed for the pine species data set. In section 3 we show how to add specifications for two different paper products to a CVA biplot. Our second aim is addressed in section 4 where we consider the analysis of distances (AOD) as an alternative to a MANOVA when the assumptions underlying the latter procedure are violated. Finally, some concluding remarks are offered regarding the use of CVA and related AOD biplots in practice.

2 A case study from the Southern African pulp industry

Product quality is influenced by the homogeneity of the raw material(s) and/or the repeatability of the processing steps involved during manufacturing. The importance of this principle has to be emphasised strongly when the conversion of biomass material into finished products is considered. One of the largest problems for a wood pulp mill is the variability introduced by the raw material, wood. In plantation forestry there is some control, *i.e.* the optimal selection of tree species, age and environmental factors such as growth site, climate, altitude, *etc.* However, this is not the case with natural forests. Pulp mills and their downstream clients, the papermakers, have to control and manage raw material variability in wood obtained from different sources, so as to optimise pulp strength properties and yield, and to ensure pulp uniformity.

Usutu Pulp Company Ltd, Swaziland, embarked in 1985 on a program to assess and manage wood quality variation at the mill and to further exploit or add value to the material, once the underlying causes of variation had been understood. Sampling of pine species in their plantations took place in 1989 and the principal findings of the project were published by Morris *et al.* (1997) and Barnes *et al.* (1999). All wood from the available forests was *fibre-typed*, so that the homogeneity of the furnish at the pulp mill could be controlled through the chip pile. The effects of major genetic (species), physiological (growth rate and age) and environmental (altitude) factors on wood and unbleached kraft pulp properties were examined. In addition to the above papers a report on this work was jointly published by the Oxford Forestry Institute, the University of Oxford, and Sappi Forest Research, South Africa (see Clarke, *et al.* (2003)). This report was presented with general interpretations and conclusions without exhaustive statistical analysis, but the data were placed in the public domain with the intention to invite researchers to conduct their own statistical analyses. This paper represents an effort in this regard.

Clarke *et al.* (2003) reported on a CVA performed on the five pine species: *Pinus elliottii, P. kesiya, P. maximinoi, P. patula* and *P. taeda*, using height growth (m) at 11 years (Growth), wood density (kg m^{-3}) (Density), total pulp yield (% of original mass) (TotYield), alkali consumption (%) (Alkali), tearing index (mN m^2 g^{-1}) (Tear), burst index (kPa m^2 g^{-1}) (Burst), tensile index (Tensile) and tensile energy absorption (mJ g^{-1}) (TEA). The CVA results in two linear combinations of the eight independent variables separating the five pine species optimally. These two canonical variates account for 80.7% and 10.9% of the between species variation respectively. According to Clarke *et al.* (2003) TotYield, Density and Burst are the most important variables for distinguishing between the selected five pine species. These authors provide a conventional scatterplot similar to Figure 1 of the first and second canonical variate scores of the five species. Figure 1 suggests that the five species form three distinct groups: *P. elliottii* on its own, *P. maximinoi* and *P. patula* grouped together and a third group formed by *P. kesiya* and *P. taeda*.

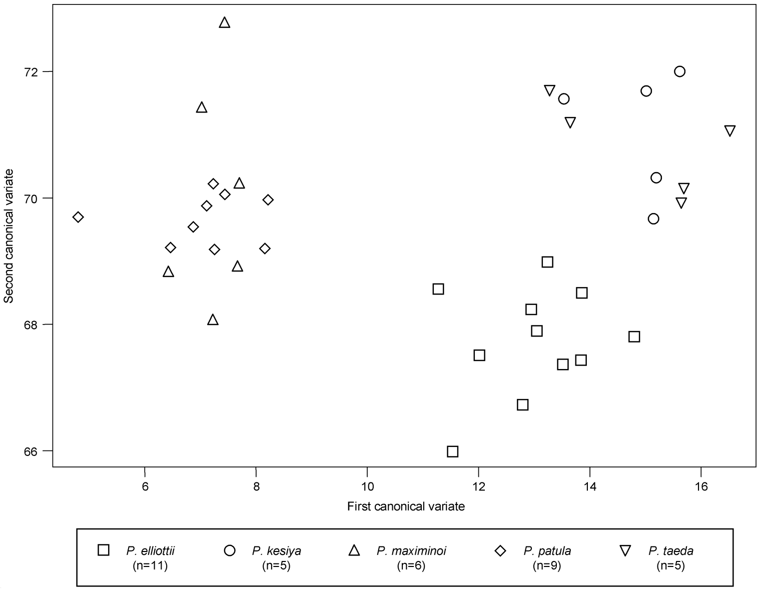

Figure 1: *Scatterplot of the first and second canonical variate scores of five pine species.*

Figure 1 provides striking evidence about the group structure of the five pine species, but it has a serious shortcoming: no information regarding the original physical variables is displayed. This limitation may be addressed by constructing a CVA biplot of the data. Gabriel (1971) introduced a biplot as the simultaneous display of the rows (the sampling units) and the columns (the variables) of a data matrix. Gabriel's biplot has been applied to many diverse fields, among them chemi-mechanical pulping trials by Broderick *et al.* (1995). These authors were able to utilise the Gabriel biplot for monitoring changes in a large number of pulp quality responses simultaneously. However, since the rows and the columns of the data matrix are represented by Gabriel as two sets of vectors in a biplot, distances between sample units are in terms of inner products. Consequently interpretation of this form of the biplot is not straight-forward. Gower and Hand (1996) presented a form of the biplot that is more accessible to non-statistical audiences by regarding it as a multivariate extension of an ordinary scatterplot: sample points appearing as points in a biplot that is equipped with p axes (instead of only two) representing all $p > 2$ variables simultaneously. Figure 2 is a CVA biplot according to the methodology provided by Gower and Hand (1996) of the data displayed in Figure 1.

Comparing Figure 2 to Figure 1 we draw attention to the following:

- The biplot in Figure 2 aims to separate the five groups optimally and to preserve inter data point distances. Unlike in Figure 1 the distances between all points including the means in Figure 2 can be evaluated in terms of ordinary Euclidean distances, since special care has been taken in the construction of Figure 2 to ensure that a change of one unit in a horisontal direction is geometrically the same as a similar change in a vertical direction.

- The canonical variates are not shown in Figure 2, but form the scaffolding for constructing a graph in which the sample points are shown together with eight axes representing the variables.

- Each axis in Figure 2 is calibrated in the original units of measurement; therefore it allows — similar to an ordinary scatterplot — for reading-off the value of the respective variable for a given point. This process is illustrated for the mean Tensile value of *P. kesiya.* All other mean values may be obtained similarly. These graphically obtained values compare favourably with the true mean values given in Table 1.

	TotYield	Alkali	Density	TEA	Tensile	Tear	Burst	Growth
P. ell	43.30	74.57	543.86	91.85	1799.09	11.90	6.15	14.31
P. kes	44.36	74.28	529.97	96.16	1810.00	9.71	6.57	16.30
P. max	45.63	75.68	456.58	93.58	1761.67	10.85	6.55	13.73
P. pat	46.44	74.39	497.68	93.63	1624.44	13.20	6.19	15.24
P. tae	44.24	74.64	567.50	96.00	1808.00	11.78	6.55	15.58

Table 1: *Means of five species of pines on eight variables.*

- Unlike an ordinary scatterplot, the axes in Figure 2 cannot be used for placing a new point on the biplot. The axes provided are called prediction axes and are only to be used for reading-off values of existing points for the different variables. Instead of prediction axes we could have equipped the biplot with axes that allow placement of new points. Such axes are called interpolation axes. Gower and Hand

(1996) or Aldrich *et al.* (2004) may be consulted for the algebraic derivation of prediction axes, interpolation axes and calibrating biplot axes. Since it is natural to use scatterplot axes for reading-off values, we prefer to equip biplots with prediction axes and execute placement of new points programmatically, based on the algebraic derivation of interpolating such points onto the biplot scaffolding (see Aldrich *et al.* (2004) or Gower and Hand (1996)).

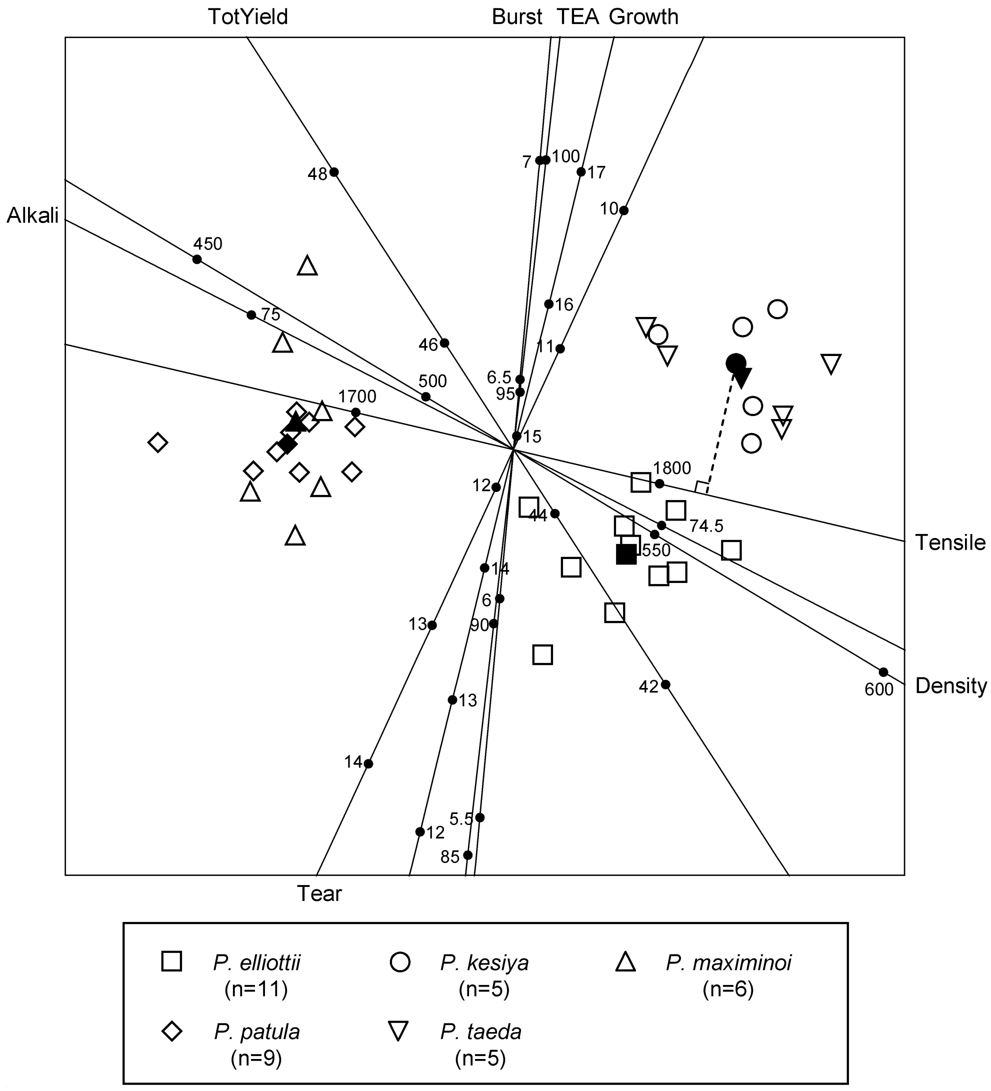

Figure 2: *CVA biplot of the pine species with group means shown as solid symbols respectively.*

- The biplot in Figure 2, unlike Figure 1, provides suggestions as to which of the variables contribute to separation between the five species: *e.g.* the variable Burst separates relatively high valued *P. kesiya* and *P. taeda* from relatively low valued *P. elliottii* with *P. patula* and *P. maximinoi* in between; the variable Density separates relatively high valued *P. kesiya*, *P. taeda* and *P. elliottii* from low valued *P. patula* and *P. maximinoi*; the variable TotYield separates high-valued *P. patula* and *P. maximinoi* from relatively low valued *P. kesiya*, *P. taeda* and *P. elliottii*. The conclusions of Clarke *et al.* (2003) are thus supported.

- The quality of the CVA biplot, as measured by the size of the eigenvalues associated with the first two canonical variables, is 91.6%.
- Gower and Hand (1996) proposed an adequacy measure of how well each variable is represented in the biplot. The adequacies for the eight variables in the pine species data set are given in Table 2.

Total Yield	Alkali	Density	TEA	Tensile	Tear	Burst	Growth
0.615	0.005	0.909	0.004	0.134	0.392	0.399	0.375

Table 2: *Adequacies associated with the axes (variables) in Figure 2.*

It is clear, from Table 2, that Alkali and TEA have very low adequacies and these variables are, therefore, not well represented in Figure 2. Reconstructing the CVA biplot by omitting these two variables results in only negligible changes in the positions of the sample points, the group means and the remaining axes (see Figure 3).

Contrast	TotYield		Alkali		Density		TEA	
P. ell – P. kes	–3.37	1.25	–4.11	4.69	–65.26	93.05	–10.59	1.96
P. ell – P. max	–4.51	–0.16	–5.25	3.03	12.80	161.76	–7.64	4.16
P. ell – P. pat	–5.07	–1.22	–3.48	3.85	–19.78	112.15	–7.02	3.44
P. ell – P. tae	–3.25	1.37	–4.47	4.33	–102.79	55.52	–10.43	2.12
P. kes – P. max	–3.87	1.32	–6.34	3.54	–15.49	162.25	–4.47	9.62
P. kes – P. pat	–4.48	0.31	–4.66	4.44	–49.57	114.15	–3.96	9.01
P. kes – P. tae	–2.59	2.83	–5.52	4.80	–130.35	55.29	–7.20	7.52
P. max – P. pat	–3.07	1.45	–3.01	5.59	–118.45	36.25	–6.18	6.08
P. max – P. tae	–1.20	3.99	–3.90	5.98	–199.79	–22.05	–9.46	4.63
P. pat–P. tae	–0.19	4.60	–4.80	4.30	–151.68	12.04	–8.85	4.12

Contrast	Tensile		Tear		Burst		Growth	
P. ell – P. kes	–385.62	363.80	–1.06	5.43	–0.94	0.09	–4.81	0.84
P. ell – P. max	–315.17	390.02	–2.01	4.10	–0.89	0.08	–2.08	3.23
P. ell – P. pat	–137.61	486.91	–4.01	1.40	–0.47	0.38	–3.29	1.42
P. ell – P. tae	–383.62	365.80	–3.13	3.36	–0.92	0.11	–4.09	1.55
P. kes – P. max	–372.35	469.02	–4.78	2.51	–0.56	0.60	–0.60	5.74
P. kes – P. pat	–201.95	573.06	–6.84	–0.13	–0.15	0.91	–1.87	3.97
P. kes – P. tae	–437.39	441.39	–5.87	1.74	–0.58	0.62	–2.59	4.03
P. max – P. pat	–228.94	503.38	–5.53	0.82	–0.14	0.86	–4.27	1.25
P. max – P. tae	–467.02	374.35	–4.58	2.71	–0.58	0.57	–5.02	1.32
P. pat–P. tae	–571.06	203.95	–1.94	4.78	–0.89	0.17	–3.26	2.58

Table 3: *Simultaneous Bonferroni 90% confidence intervals for all pairwise contrasts among five pine species. Intervals excluding zero are highlighted.*

The close ties between CVA and MANOVA (see for example Kshirsagar (1972) and Gittins (1985)) allow simultaneous confidence statement techniques associated with MANOVA to be used for investigating the significance of contrasts among the five species with respect to the individual variables. The overall MANOVA null hypothesis of equal group mean vectors is rejected at a very small significance level (the *p*-value approaching zero) and the resulting Bonferroni simultaneous 90% confidence intervals for all contrasts are given in Table 3. Perusal of Table 3 suggests that the only variables that differ statistically significant when comparing the species pairwise are TotYield, Density and Tear *viz.* when investigating *P. ell – P. max*; *P. ell – P. pat*; *P. max – P. tae* and *P. kes – P. pat.*

3 Specifications for different usages of paper products

Diverse qualities of wood pulp are of importance to different users of paper products. The suitability of wood pulps for specific end uses, such as kraft liner or sack paper, may be expressed in terms of the wood and processing characteristics. Based on practical experience, guiding range values, as specified in Table 4, were compiled.

	Kraft liner		Sack paper	
	Min	Max	Min	Max
Total Yield	43	50	43	50
Alkali	70	85	70	85
Density	422	550	422	550
TEA	85	Large	90	Large
Tensile	1400	2400	1800	2600
Tear	8	15	10	17
Burst	5	6	6	7

Table 4: *Specifications for two different end uses of paper products.*

In addition to displaying sample points and all variables in a single graph, biplot methodology allows for interpolating product specifications, such as given in Table 4, in the form of acceptance regions into the plot. This is illustrated in the CVA biplot in Figure 3.

It follows from Figure 3 that none of the five pine species may be regarded as good for kraft liner, but that *P. maximinoi* and *P. patula* are suitable for sack paper. It appears that, in the case of the remaining three species, both Density and Tensile are on the high side of what is required by manufacturers of sack paper.

4 Analysis of distances

CVA was initially derived by Fisher as a non-parametric method, but it requires homogeneity of group dispersion matrices, since the within covariance matrix is formed by pooling the covariance matrices of the different groups (see, for example, Gittins (1985)). If hypothesis testing is to be performed as part of a CVA the assumption of multivariate normality also has to be made, establishing a close connection between CVA and MANOVA. Inspecting the variances of the variables for the different pine species casts serious doubt upon the use of a pooled covariance matrix in CVA and MANOVA in the data set considered in the previous sections. Furthermore, due to the small sample sizes, some of the species have singular covariance matrices while multivariate normality is doubtful. However, the basic formula for a variance can be expressed in terms of the squared distances between all pairs of points, since

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2. \tag{1}$$

Equation (1) thus provides the connection between analysis of variance and analysis of distance (AOD). Gower and Krzanowski (1999) showed that AOD is akin to analysis of variance in that the intersample distances can be broken down into a within sum of squared distances component and a between sum of squared distances component ($T = W + B$). Moreover, in AOD no distributional assumptions are made nor any assumption of homogeneity of dispersion matrices.

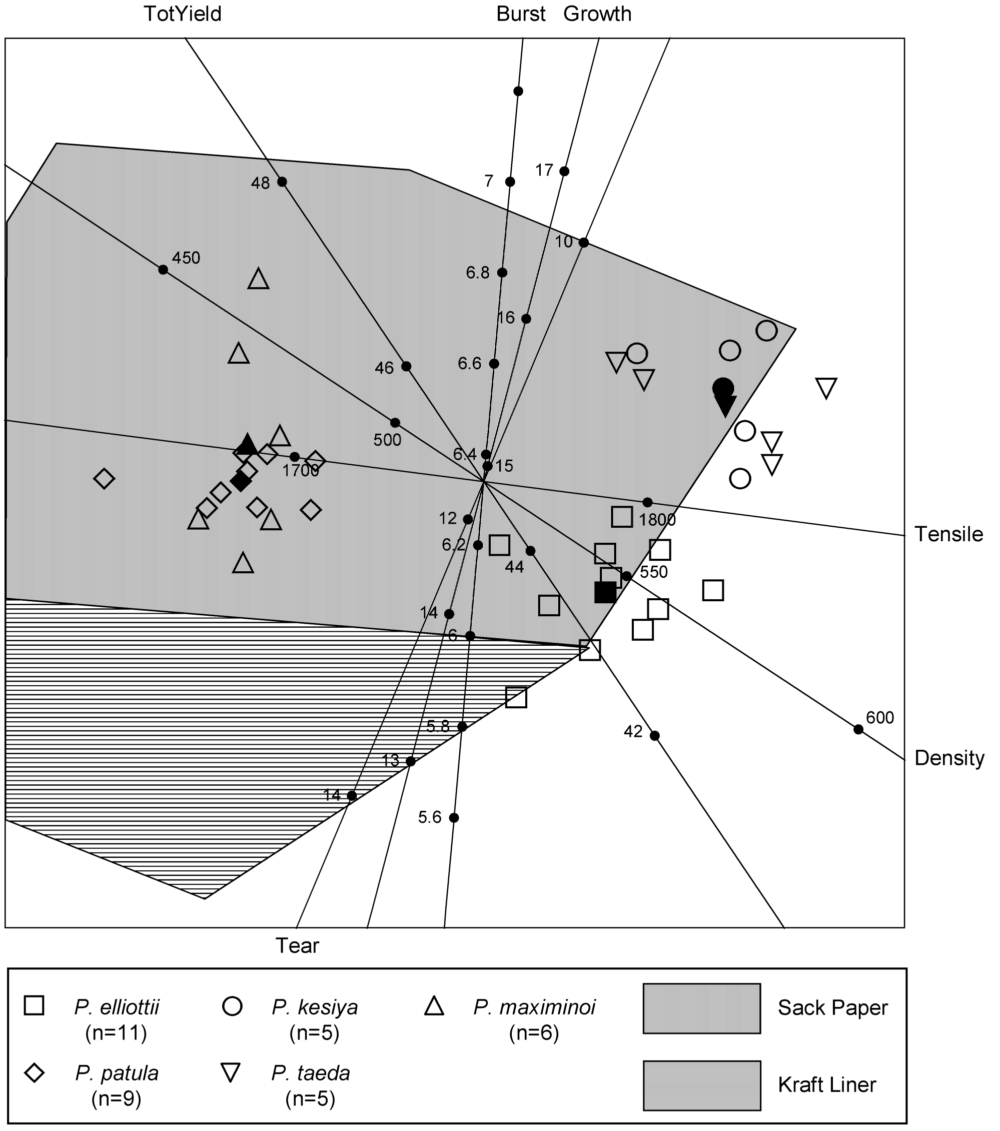

Figure 3: *CVA biplot with acceptance regions for two different end uses of paper products.*

The pine species data may be arranged in a matrix $\mathbf{X}$ of size $n \times p$ with $n = 36$ and $p =$

8. Let d_{ij} denote the Euclidean distance $\{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\}^{1/2}$ between samples i and j for $i,j=1,\dots,n$. The squared distances may be gathered to form the symmetric $n\times n$ matrix $\mathbf{D}$ with ij-th element $\frac{1}{2}d_{ij}^2$. The matrix $\mathbf{X}$ satisfies the equation

$$-(\mathbf{I}_n-\mathbf{1}\mathbf{s}')\mathbf{D}(\mathbf{I}_n-\mathbf{s}\mathbf{1}')=\mathbf{X}\mathbf{X}', \tag{2}$$

where the vector $\mathbf{s}$ satisfies $\mathbf{s}'\mathbf{1}=1$, with $\mathbf{1}$ a vector of ones and $\mathbf{I}_n$ the $n\times n$ identity matrix. Any solution of (2) for $\mathbf{X}$ yields coordinates, say $\mathbf{Y}$ of the sample points. There are various ways of obtaining $\mathbf{Y}$; in particular, classical metric scaling (see, for example, Borg and Groenen (1997), Cox and Cox (2001) as well as the references contained therein). Classical metric scaling is also known as principal coordinate analysis (PCO) (Gower (1966)). The PCO solution of (2) for $\mathbf{s}=\frac{1}{n}\mathbf{1}$ is based upon the spectral decomposition

$$-(\mathbf{I}_n-\tfrac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{D}(\mathbf{I}_n-\tfrac{1}{n}\mathbf{1}\mathbf{1}')=\mathbf{V}\mathbf{\Lambda}\mathbf{V}' \tag{3}$$

with $\mathbf{V}$ an $n\times n$ the matrix of orthonormalised eigenvectors and $\mathbf{\Lambda}$ the diagonal matrix containing the corresponding eigenvalues. Thus the PCO coordinate matrix is given by $\mathbf{Y}=\mathbf{V}\mathbf{\Lambda}^{1/2}$. If the n sample points are divided into g groups with corresponding sizes $n_1,n_2,\dots,n_g$, the group structure may be described by the $n\times g$ indicator matrix $\mathbf{G}$ with ij-th element equal to 1 if sample point i is in group j and zero otherwise. Let $\mathbf{N}$ be the diagonal matrix with diagonal elements the respective group sizes, $n_1,n_2,\dots,n_g$. Then the coordinates of the group means are given by

$$\bar{\mathbf{Y}}=\mathbf{N}^{-\mathbf{1}}\mathbf{G}'\mathbf{Y}. \tag{4}$$

As mentioned, Gower and Krzanowski (1999) proved that the total sum of squared distances T may be decomposed as $T=W+B$, where

$$\begin{aligned} T &= \tfrac{1}{n}\mathbf{1}'\mathbf{D}\mathbf{1}, & (5)\\ W \text{ (within-group sum of squared distances)} &= \sum_{j=1}^{g}\tfrac{1}{n_j}\mathbf{1}_j'\mathbf{D}_{jj}\mathbf{1}_j,\ \text{and} & (6)\\ B \text{ (between-group sum of squared distances)} &= \tfrac{1}{n}\mathbf{n}'\bar{\mathbf{\Delta}}\mathbf{n}. & (7)\end{aligned}$$

In equations (6) and (7) $\mathbf{D}_{jj}$ refers to the matrix $\mathbf{D}$ being partitioned into g^2 submatrices $\mathbf{D}_{rs}$ of dimensions $n_r\times n_s$ such that $\mathbf{D}_{rs}$ consists of 0.5 times the squared distances between each individual in group r and each individual in group s; $\mathbf{n}$ is a vector containing the group sizes, while $\bar{\mathbf{\Delta}}$ is a symmetrical $g\times g$ matrix with rs-th element being 0.5 times the squared distance between the means of groups r and s.

Since the units of measurement of the variables used in the pine species data set are not commensurable, each variable was centred and scaled to unit variance before performing the AOD. The AOD of the centred and scaled values results in $T=280.00$, $B=84.67$ and $W=195.33$. Similar to MANOVA the question of importance now is whether these results indicate a significant difference between the respective group means. This question can be addressed without having to make any distributional or homogeneity of covariance matrices assumptions by using permutation tests. The basic idea in the above context is as follows: if the B-term in the above AOD expression indicates no group difference then its contribution to T should remain approximately constant over random permutations of the sampling units into g groups of fixed sample sizes. However, if the null hypothesis does not hold, permutation of sampling units should tend to reduce between-group variability. A large number of random permutations can thus be carried out and the B-value (7) calculated for each of these permutations. The obtained value of $B = 84.67$ is now compared with the distribution of permutation B-values. The achieved significance level of this testing procedure is the proportion of times the value of $B = 84.67$ is exceeded by the permutation replicates of this statistic. We employed the algorithm provided by Good (2000) with 10 000 permutations leading to a significance level of less than 1 in 10 000 of rejecting the null hypothesis of equal group means. This seems to support the MANOVA p-value we found to be approaching zero.

Biplots to accompany the above AOD may also be constructed. Our AOD of the pine species data set focused on the corresponding group means (4). A biplot that provides a graphical display of the separation and overlap between the five groups, analogous to the CVA biplots that accompany the MANOVA, may be obtained as follows:

The coordinates, $\bar{\mathbf{Y}}$, for the group means may be found from the spectral decomposition of

$$-(\mathbf{I}_g - \tfrac{1}{g}\mathbf{1}\mathbf{1}')\mathbf{D}^*(\mathbf{I}_g - \tfrac{1}{g}\mathbf{1}\mathbf{1}') = \bar{\mathbf{Y}}\bar{\mathbf{Y}}', \tag{8}$$

where $\mathbf{D}^* = \mathbf{N^{-1}G'DGN^{-1}}$. This amounts to a standard PCO, leading to a display in which the origin of the axes is at the centroid of the group means points not weighted by the group sizes. Alternatively, a weighted PCO may be performed using

$$-(\mathbf{I}_g - \tfrac{1}{n}\mathbf{1}\mathbf{n}')\mathbf{D}^*(\mathbf{I}_g - \tfrac{1}{n}\mathbf{n}\mathbf{1}') = \bar{\mathbf{Y}}\,\bar{\mathbf{Y}}'. \tag{9}$$

This leads to a graphical display in which the origin of the axes is at the centroid of the n individual points. After finding a graphical representation of the group means in, say, 2 dimensions it is of interest to display the individual sample points about their respective means. Following Gower (1968) this may be accomplished by plotting

$$-\tfrac{1}{2}(\bar{\mathbf{Y}}^{*\prime}\bar{\mathbf{Y}}^*)^{-1}\bar{\mathbf{Y}}^{*\prime}(\mathbf{I}_g - \tfrac{1}{n}\mathbf{1}\mathbf{1}')(\mathbf{d}_i - \mathbf{d}_0) \tag{10}$$

for $i = 1, 2, \ldots, n$. In (10) $\bar{\mathbf{Y}}^*$ denotes the two dimensional approximation of $\bar{\mathbf{Y}}$ obtained from the spectral decomposition of (9). The vector $\mathbf{d}_i$ contains the squared distances of the i-th sample point to each of the g group means on the graph, while the elements of $\mathbf{d}_0$ are the squared distances of the respective group means to the origin of the axes. If a weighted PCO is performed the individual points may similarly be plotted by replacing (10) with

$$-\tfrac{1}{2}(\bar{\mathbf{Y}}^{*\prime}\bar{\mathbf{Y}}^*)^{-1}\bar{\mathbf{Y}}^{*\prime}(\mathbf{I}_g - \tfrac{1}{n}\mathbf{1}\mathbf{n}')(\mathbf{d}_i - \mathbf{d}_0). \tag{11}$$

In order to complete the biplot, information about the original variables must be added in the form of calibrated axes. We implement the procedure Gower and Hand (1996) proposed for metric multidimensional scaling, leading to the biplot displayed in Figure 4. This biplot results from a standard PCO and has an overall quality of display of 75.8%. The corresponding biplot for a weighted PCO is almost identical and is therefore not repeated here. Note also that although the data were scaled to unit variances before performing the AOD the biplot axes in Figure 4 are calibrated to show the original units of measurement. Performing the AOD on the unscaled data leads to a biplot in which all the data points lie along the Tensile axis. In this case the achieved significance level obtained with the permutation testing procedure turns out to be 0.1966. We do not include the biplot for the unscaled case here. Suffice it to say that, although a CVA biplot is invariant with respect to the scaling of variables to unit variances, the above example shows this not to be the case for AOD biplots. Therefore users should carefully consider the influence of the scaling of variables on the appearance of AOD biplots and associated hypothesis testing.

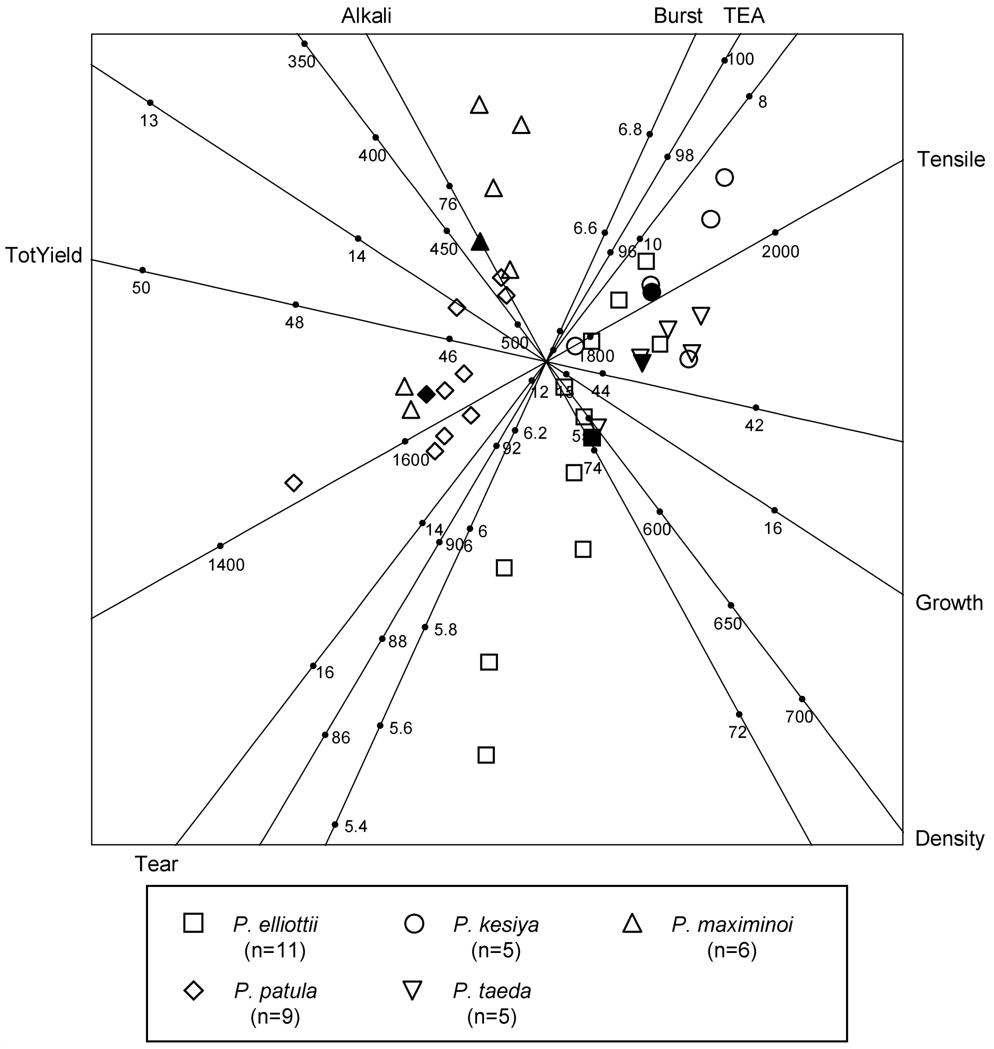

Figure 4: *Standard principal coordinate analysis biplot displaying the AOD of the pine species data set.*

Although there are conspicuous similarities between Figure 4 and Figures 1 through 3 the AOD biplot in Figure 4 shows the five groups to be differently separated than suggested by the CVA biplots: *P. maximinoi* has been shifted away from *P. patula*, while *P. taeda* and *P. kesiya* are further apart from each other than in the CVA biplots. Distortion due to dimension reduction may be responsible for these differences. In view of the relatively large quality values obtained it seems likely that these differences are attributable to the influence of the incorrect pooling of heterogeneous (and in some cases singular) class covariance matrices. The biplot in Figure 4 not only provides suggestions of the variables responsible for the differences among the various pine species, but also portrays the role of the relatively large variances characteristic of some of the groups. In particular, the large tensile variance of *P. elliottii* is noteworthy.

5 Conclusion

We have demonstrated the capabilities of a CVA biplot when regarded as a multivariate extension of an ordinary scatterplot for providing information regarding overlap and separation between five different pine species, as well as the respective roles played by eight wood and processing characteristics of these species. A further extension allows users of pulp to inspect visually whether their specifications for a particular usage are met by a particular pine species.

We also provide evidence questioning the appropriateness of MANOVA and hence also of ordinary CVA biplots in the case of the pine species data set discussed by Clarke, *et al.* (2003). However, even though lack of homogeneity among the covariance matrices of the different pine species casts some doubt about constructing ordinary CVA biplots for a detailed description of the pine species data set, its usefulness for the visual display of multivariate variation, of group structure and of the roles of more than just two variables together with multidimensional specifications as demanded by users is not to be disputed. Indeed, Figures 2 and 3 illustrate the potential of the CVA biplot as an extension of an ordinary scatterplot for both producers as well as consumers of a product, the qualities of which depend on several correlated variables.

The small sample sizes and heterogeneous group covariance matrices found to question CVA and MANOVA procedures in the case of the pine species data set are often encountered in practice. As an alternative to the analysis of variance we show how the analogous analysis of distance can be used in such cases. The AOD procedure is invariant with respect to distributional assumptions. Furthermore, it is unaffected by small sample sizes and heterogeneous group covariance matrices, while permutation test procedures may be employed when hypothesis testing is required. Since the biplots discussed in this paper are designed for the accurate display of distances the multidimensional biplot illustrated in Figure 4 is an invaluable aid for a visual display of an AOD. Moreover, all features and extensions of CVA biplots are available for AOD biplots.

Acknowledgement

Financial support of the National Research Foundation of South Africa is gratefully acknowledged.

References

[1] Aldrich C, Gardner S & Le Roux NJ, 2004, *Monitoring of metallurgical process plants by using biplots*, American Institute for Chemical Engineering Journal, **50**, pp. 2167–2186.

[2] Barnes RD, Plumptre RA, Quilter AK, Morris AR, Burley J & Palmer ER, 1999, *The use of stem dissection to sample trees of different ages for determining pulping properties of tropical pines*, IAWA Journal, **20**, pp. 37–43.

[3] Borg I & Groenen P, 1997, *Modern multidimensional scaling*, Springer-Verlag, New York (NY).

[4] Broderick G, Paris J & Valade JL, 1995, *A composite representation of pulp quality*, Chemometrics and Intelligent Laboratory Systems, **29**, pp. 19–28.

[5] Clarke CRE, Morris AR, Palmer ER, Barnes RD, Baylis WBH, Burley J, Gourlay ID, O'Brien E, Plumptre RA & Quilter AK, 2003, *Effect of environment on wood density and pulp quality of five pine species grown in southern Africa*, Tropical Forestry Papers No 43, Oxford Forestry Institute, Department of Plant Sciences, University of Oxford, Oxford, 162 pp.

[6] Cox TF & Cox MAA, 2001, *Multidimensional scaling*, Chapman & Hall/CRC, Boca Raton (FL).

[7] Gabriel KR, 1971, *The biplot graphical display of matrices with application to principal component analysis*, Biometrika, **58**, pp. 453–467.

[8] Gittins R, 1985, *Canonical analysis: A review with applications in ecology*, Springer-Verlag, Berlin.

[9] Good P, 2000, *Permutation tests: a practical guide to resampling methods for testing hypotheses*, 2nd edition, Springer-Verlag, Berlin.

[10] Gower JC, 1966, *Some distance properties of latent root and vector methods used in multivariate analysis*, Biometrika, **53**, pp. 325–338.

[11] Gower JC, 1968, *Adding a point to vector diagrams in multivariate analysis*, Biometrika, **55**, pp. 582–585.

[12] Gower JC & Hand DJ, 1996, *Biplots*, Chapman & Hall, New York (NY).

[13] Gower JC & Krzanowski WJ, 1999, *Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance*, Applied Statistics, **48**, pp. 505–519.

[14] Hastie T, Tibshirani R & Friedman J, 2001, *The elements of statistical learning*, Springer, New York (NY).

[15] Johnson RA & Wichern DW, 2002, *Applied multivariate analysis*, 5th edition, Prentice-Hall, Upper Saddle River (NJ).

[16] Kaufman L & Rousseeuw PJ, 1990, *Finding groups in data*, John Wiley & Sons, New York (NY).

[17] Kshirsagar AM, 1972, *Multivariate analysis*, Marcel Dekker, New York (NY).

[18] McLachlan GJ, 1992, *Discriminant analysis and statistical pattern recognition*, John Wiley, New York (NY).

[19] Morris AR, Palmer ER, Barnes RD, Burley J, Plumptre RA & Quilter A, 1997, *The influence of felling age and site altitude on pulping properties of Pinus patula and Pinus elliottii*, Tappi, **80**, pp. 133–138.