Principal Axis Analysis

Frank Critchley[1], Ana Pires[2] and Conceição Amado[2]

Department of Statistics, The Open University, Milton Keynes[1] *and*
Department of Mathematics, CEMAT, Instituto Superior Técnico, Lisbon[2]

SUMMARY

Principal axis analysis rotates principal components to optimally detect cluster structure, rotation being based on a second spectral decomposition identifying preferred axes in the sphered data. As such, it complements principal component analysis as an extremely fast, general, projection pursuit method, particularly well-suited to detecting mixtures of elliptical distributions. Examples show that it can perform comparably to linear discriminant analysis *without* using group (cluster) membership information, while its sphered and unsphered forms offer complementary views. Points of contact with a range of multivariate methods are noted and further developments briefly indicated.

Some key words: cluster analysis; directional statistics; elliptical distributions; exploratory analysis; independent component analysis; linear discriminant analysis; mixture models; multivariate analysis of variance; principal component analysis; projection pursuit; spherical distributions.

1 OVERVIEW

Thought of as a projection pursuit method, principal component analysis (PCA) chases variability. In contrast, principal axis analysis (PAA) chases structure. More specifically, any cluster structure that may be present among the cases. In both cases, pursuit is almost instant, computation of $x \to x^{PCA}$ requiring just one spectral decomposition, and the orthogonal transformation $x^{PCA} \to x^{PAA} = U^T x^{PCA}$ to preferred axes in the sphered data just one more.

Eigenvectors being indeterminate up to overall sign, the same is true of each element of both score vectors, x^{PCA} and x^{PAA}. These indeterminacies have two immediate consequences. Intrinsic interest lies, not in directions, but in *axes* (that is, in pairs $\pm q$ ($\|q\| = 1$) of opposed directions) and, second, we may speak of principal axis analysis rotating principal components while, in practice, allowing reflections.

The mean and covariance (assumed nonsingular) of $x \sim F$ are respectively denoted μ and Ω, the latter having spectral decomposition $\Omega = Q\Lambda Q^T$, $\Lambda = \text{diag}(\lambda_r)$ ($\lambda_1 \geq ... \geq \lambda_k > 0$) being the diagonal matrix of ordered eigenvalues and $Q = (q_1|...|q_k)$ an orthogonal matrix of corresponding normalised

eigenvectors. Thus, $x^\circ = Q^T(x-\mu)$ is the usual vector of centred, uncorrelated, PCA scores, while $x^* = \Lambda^{-\frac{1}{2}}x^\circ$ is its *sphered* version having mean $\mu^* = 0$ and covariance $\Omega^* = I$.

Alternative, complementary, versions of $x^{PAA} = U^T x^{PCA}$ follow from taking x^{PCA} to be x° (unsphered form) or x^* (sphered form). Sample versions of both PCA and PAA follow in the usual way.

Deferring definition of U until later, we begin with three illustrative examples in each of which $x^{PCA} = x^*$.

1.1 Three examples

Example 1 illustrates the complementary goals of PCA and PAA as projection pursuit methods. It concerns the bivariate haemophilia data set of Hermans and Habbema (1975) – as complemented with additional cases in Johnson and Wichern (1992) – and displayed in Figure 1, where we have observations on $x_1 = \log$(Factor-VIII activity) and $x_2 = \log$(Factor-VIII like antigen) for 60 non-carriers, denoted '$\triangle$', and 45 carriers, denoted '+'. Principal component analysis transforms this to Figure 2 in which sample variance is maximal horizontally. In contrast, principal axis analysis spheres and rotates this plot some 90° to obtain Figure 3 in which the 'empirical alignment' measure of cluster structure (defined later) is maximal horizontally. Partitioning cases by their sign on this first principal axis and comparing the result with Figure 4, this example also illustrates PAA performing comparably to linear discriminant analysis (LDA) – yet without using group (cluster) membership information!

FIGURES 1 TO 4 ABOUT HERE

The k principal axes are ordered by their empirical alignments, these positive measures having average value 1. Example 2 illustrates that PAA can reveal cluster structure missed in plots of the data or their principal components. And that it aims to do so in the first few principal axes, whose empirical alignments can suggest how many to retain. This example concerns a five dimensional data set comprising 50 cases from each of two, well-separated, normal distributions with the same covariance matrix. This sub-population structure is not evident in the full scatterplotmatrix of the original data (Figure 5), or of their principal components (Figure 6), but is clear from a plot of the first two principal axes (Figure 8). Indeed, only the first principal axis is required, as suggested by the empirical alignment plot (Figure 7) where only one axis is clearly preferred.

FIGURES 5 TO 8 ABOUT HERE

Example 3 illustrates that PAA remains trivial to compute for large sample size and/or dimension, just two spectral decompositions of order k being required. And that it can sometimes recover underlying structure somewhat better than linear discriminant analysis when the constant within-group covariance assumption of LDA fails. It concerns a ten dimensional data set comprising 250 cases from each of four normal distributions, with differing covariance matrices

and whose means lie at the corners of a square. This structure is not evident in either the original or principal component scatterplotmatrix (not shown), but is clear in a plot of the first two principal axes (Figure 9) and, with a pronounced 'tilt', in the LDA plot (Figure 10). There are just two clearly preferred axes in the empirical alignment plot (not shown), correctly suggesting that there is no need to explore higher dimensions for additional cluster structure.

FIGURES 9 AND 10 ABOUT HERE

Further examples are given later. Next, we note that principal axis analysis is a DREaM methodology. That is, it has points of contact with each of Diagnostics, Robustness, Exploration and Modelling.

1.2 A DREaM methodology

Given a multivariate data set $\{x_i \in \mathcal{R}^k : i = 1, ..., n\}$ – equivalently, given n and the empirical distribution $\widehat{F}$ – PAA addresses the exploratory question: 'What is F?'. It does so with a meta-model in mind in which

$$F = \sum_{g=1}^{G} \pi_g F_g \ \ (\pi_g > 0, \ \sum_{g=1}^{G} \pi_g = 1) \tag{1}$$

is a mixture of an unknown number G of elliptical, but otherwise unspecified, distributions with well-separated means $\{\mu_g\}_{g=1}^{G}$, each assumed to differ from the overall mean $\mu = \sum_{g=1}^{G} \pi_g \mu_g$. Both extremes $G = 1$ and $G = n$ are entertained, the former subsuming many classical parametric models and the latter corresponding to classical nonparametric statistics, where $\widehat{F}$ estimates F.

There are a number of approaches to fitting models such as (1), though none is universally ideal. There is an extensive literature on mixture models, while a range of cluster analysis methods is also available (see, for example, Gordon (1981), Everitt et al. (2001) and, in the robust case, Cuesta-Albertos et al. (1997)). However, many such methods impose restrictions on the $\{F_g\}_{g=1}^{G}$ and/or their covariance matrices $\{\Omega_g\}_{g=1}^{G}$ and are relatively expensive computationally, especially for large n or k. There may also be difficulties in choosing the unknown number of groups G.

Alternatively, a method of robust principal component analysis may be used. This growing field can be accessed via the recent papers of Locantore et al. (1999), Marden (1999), Croux and Haesbroeck (2000), Croux et al. (2002), Visuri et al. (2003) and Hubert et al. (2005). However, these methods assume that there is a $\pi_g \geq \frac{1}{2}$, which may not be the case.

Principal axis analysis is, in itself, an extremely fast, general method of exploratory multivariate analysis, particularly well-suited to detecting mixtures of the form (1). One of its key features is the use of diagnostic information to indicate when cluster structure is present ($G > 1$). In later work, we plan to use PAA as a first, exploratory, step in a flexible, robust-diagnostic strategy to identify empirically, and then estimate efficiently, a set of candidate models within this general class. It may also be useful more generally as a fast precursor to mixture, cluster or robust analyses.

The interesting algorithm proposed in Peña and Prieto (2001) is similar in spirit to principal axis analysis but differs in important ways, being kurtosis-based, affine equivariant and requiring specification of adjustable parameters.

1.3 Organisation

The rest of the paper is organised as follows. Sections 2 and 3 describe principal axis analysis in the population and in the sample respectively. Section 4 discusses two further examples with large-scale real data sets, illustrating the method's robustness to departures from (1) and the complementarity of its sphered and unsphered forms. Section 5 notes further points of contact with other multivariate methods, including independent component analysis, and offers a quantification of PAA's complementarity with PCA. Further developments are briefly indicated in the final section.

A number of results on spherical and related distributions and on directional statistics are employed. For details of these see, respectively, Fang et al. and Ng (1990) and Mardia and Jupp (2000).

We adopt the usual Euclidean geometry for vectors z and matrices M so that, in particular, $\|z\|^2 = z^T z$ and $\|M\|^2 = \text{trace}(M^T M)$. For any nonzero vector z, $\text{dir}(z) = z/\|z\|$ denotes its direction. An axis is identified with the pair $\pm q$ ($\|q\| = 1$) of opposed direction vectors which it contains and, again, with the 'axial matrix' qq^T representing orthogonal projection onto it.

With probability 1, $x^* \neq 0$ and we have the length-times-direction factorisation $x^* = D.d,\ (D > 0,\ d = \text{dir}(x^*))$, where $D^2 = \|x^*\|^2$ denotes the squared Mahalanobis distance of x from F, whose expectation is k.

We say that a random vector and its distribution are sphered if it has zero mean and identity covariance. Finally, F^* denotes the sphered distribution of x^* induced by $x \sim F$.

2 PAA IN THE POPULATION

The key idea behind principal axis analysis is that preferred axes in F^* indicate the presence of cluster structure ($G > 1$), in which case they can help find it.

Consider first the null case $G = 1$ when no cluster structure is present. In this elliptical case, F has preferred axes, but F^* does not. Indeed, as is intuitive geometrically, the major axis $\pm q_1$ is typically most preferred in F, Marden (1999, p. 351) proving that here $\text{cov}(\text{dir}(x - \mu)) = Q\widetilde{\Lambda}Q^T$ for some diagonal matrix $\widetilde{\Lambda}$ confirmed, empirically, to have nonincreasing diagonal elements for many elliptical distributions. As is again intuitive geometrically, the sphering $F \to F^*$ removes these preferred axes. Formally, recall that a distribution is spherical if it is unchanged by orthogonal transformation, while an elliptical distribution is an affine transformation of a spherical distribution. Thus, $x \sim F$ is elliptical precisely when $x^* \sim F^*$ is spherical. This in turn arises precisely when D and d are independent with d *uniformly* distributed. In particular, the expected axial matrix $A = E_{F^*}(dd^T)$ is, here, a multiple of the identity.

For $G > 1$, F is not elliptical and so either (a) D and d are dependent or (b) d is not uniform, providing complementary sources of diagnostic information with which to detect cluster structure. This paper restricts attention to (b), the geometric and algebraic insight which follows indicating that the nonuniformity of d has structure whose preferred axes can help find the cluster mean subspace $\mathcal{M} = \text{Span}\{(\mu_g - \mu) : g = 1, ..., G\}$ *without* knowing group (cluster) memberships. Recalling that $\mu^* = 0$, the nonsingular affine transformation $x \to x^*$ sends $\mathcal{M}$ to $\mathcal{M}^* = \text{Span}\{\text{dir}\left(\mu_g^*\right) : g = 1, ..., G\}$, a subspace of the same dimension $m \leq \min\{G - 1, k\}$. Note that m is manageably small in many practically encountered problems, being 1 in the two group problems of Examples 1 and 2 above, while m reduced to 2 in the four group problem of Example 3 due to coplanarity of the group means.

Recall that (1) supposes well-separated means $\{\mu_g\}_{g=1}^G$, each different from the overall mean μ. The essential geometric insight of principal axis analysis is clear from the $G = 2$ case. Here, μ lies along the line joining μ_1 and μ_2. This collinearity is preserved under $x \to x^*$ which sends $\mu \to \mu^* = 0$ so that $\text{dir}(\mu_1^*) = -\text{dir}(\mu_2^*)$. Accordingly, F^* comprises complementary proportions π_1 and π_2 of cases biased in opposed directions from the origin. In general, for each $g = 1, ..., G$, F^* has a proportion π_g of cases biased in the direction of $\text{dir}(\mu_g^*)$, these directions together spanning $\mathcal{M}^*$.

The better separated the $\{F_g\}$, the more marked this pattern becomes, within group noise distracting less from between group signal. In the limit as each $\Omega_g \to 0$, $d = \text{dir}(\mu_g^*)$ with probability π_g, so that the expected axial matrix A reduces to $A_0 = \sum_{g=1}^G \pi_g \,\text{dir}(\mu_g^*)\,\text{dir}(\mu_g^*)^T$, whose range is the subspace $\mathcal{M}^*$ which we seek.

In general under (1), Range(A) coincides with $\mathcal{M}^*$ to first order, in the following sense. Let $x \to x^*$ induce $F_g \to F_g^*$. Then $x^* = x_g^* \sim F_g^*$ with probability π_g so that $A = \sum_{g=1}^G \pi_g E_{F_g^*}(d_g d_g^T)$ where $d_g = \text{dir}\left(x_g^*\right)$. As $x_g^* = \left\|\mu_g^*\right\| \left\{\text{dir}\left(\mu_g^*\right) + \Gamma_g z_g\right\}$ where $\Gamma_g = \Lambda^{-\frac{1}{2}} Q^T \Omega_g^{\frac{1}{2}} / \left\|\mu_g^*\right\|$ and z_g is sphered, while by well-separatedness each $\|\Gamma_g\|^2 = \text{trace}\left(\Omega^{-1}\Omega_g\right) / \left\|\mu_g^*\right\|^2$ is small, it follows that there is no linear term in the expansion $A = A_0 + O\left(\gamma^2\right)$, where $\gamma = \max\{\|\Gamma_g\| : g = 1, ..., G\}$.

Overall, the insight is that the greater the expected alignment of an axis with $\pm d$, the more cluster structure it exhibits. We formalise this as follows. Let A have spectral decomposition $A = U\Delta U^T$, where $U = (u_1|...|u_k)$ is orthogonal and $\Delta = \text{diag}(\delta_r)$ with $\delta_1 \geq ... \geq \delta_k \geq 0$ and, we note, $\sum_{r=1}^k \delta_r = 1$. For any fixed axis $\pm u$, its alignment with the random axis $\pm d$ can be measured by

$$\delta_{\pm u} = (u^T d)^2 = \cos^2(\theta) = \cos^2(\pi - \theta),$$

where $\theta \leq (\pi - \theta)$ are the complementary angles made by these axes. Since $\delta_{\pm u}$ has expectation $u^T A u$, $\pm u_1$ has maximal cluster structure, this maximum being δ_1. Orthogonal to this axis, $\pm u_2$ has maximal cluster structure δ_2. And so on. When there is no cluster structure ($G = 1$), uniformity of d implies $A = k^{-1} I$ so that every axis has the same expected alignment, k^{-1}. In general,

$\pm u_1, ..., \pm u_k$ are the ordered principal axes of F^*, those with above average expected alignment – that is, with 'population alignment' $k.\delta_r > 1$ – being preferred.

Principal axis analysis in the population comprises, then, the pair of transformations:

$$x \to x^{PCA} = x^\circ \text{ (unsphered) or } x^* \text{ (sphered), followed by}$$
$$x^{PCA} \to x^{PAA} = U^T x^{PCA}.$$

The sphered and unsphered forms of principal axis analysis are complementary. In the former case, x^{PAA} itself is sphered, so that its elements are uncorrelated but carry no shape information in the sense of differential axis variance. In the latter case, the elements of x^{PAA} are, in general, correlated but do carry such shape information, which can be insightful. This complementarity is useful in practice, as examples show (see below).

Like principal component analysis, principal axis analysis is invariant to changes of location, rotation and reflection and is equivariant under overall rescaling. The fact that $\|d\| = 1$ confers a certain robustness on its additional transformation $x^{PCA} \to x^{PAA}$.

3 PAA IN THE SAMPLE

Let $\{x_i : i = 1, ..., n\}$ be a multivariate data set of size $n > k$ with empirical distribution $\widehat{F}$. Replacing F by $\widehat{F}$ and x by x_i, principal axis analysis in the sample is defined by direct analogy with its population version.

With probability 1, the empirical covariance matrix $\widehat{\Omega}$ is nonsingular. Denoting its spectral decomposition by $\widehat{\Omega} = \widehat{Q}\widehat{\Lambda}\widehat{Q}^T$, principal component analysis transforms each $x_i \to x_i^{PCA}$ taken to be $x_i^\circ = \widehat{Q}^T(x_i - \overline{x})$ or $x_i^* = \widehat{\Lambda}^{-\frac{1}{2}} x_i^\circ$.

With probability 1, each $x_i^* \neq 0$ and we have the length-times-direction factorisation $x_i^* = D_i.d_i, \quad (D_i = \|x_i^*\| > 0, \ d_i = \text{dir}(x_i^*))$. The mean axial matrix $\widehat{A} = n^{-1} \sum_{i=1}^n d_i d_i^T$ having spectral decomposition $\widehat{A} = \widehat{U}\widehat{\Delta}\widehat{U}^T$, where $\widehat{U} = (\widehat{u}_1|...|\widehat{u}_k)$ is orthogonal and $\widehat{\Delta} = \text{diag}(\widehat{\delta}_r)$ with $\widehat{\delta}_1 \geq ... \geq \widehat{\delta}_k \geq 0$ and $\sum_{r=1}^k \widehat{\delta}_r = 1$, $\pm\widehat{u}_1, ..., \pm\widehat{u}_k$ are the ordered principal axes of the empirical distribution $\widehat{F}^*$ of the $\{x_i^*\}_{i=1}^n$. Those with above average mean alignment with the empirical axes $\{\pm d_i\}_{i=1}^n$ – that is, with 'empirical alignment' $k.\widehat{\delta}_r > 1$ – are preferred.

Finally, principal axis analysis in the sample transforms each $x_i^{PCA} \to x_i^{PAA} = \widehat{U}^T x_i^{PCA}$, inheriting the invariance, equivariance and robustness properties of its population counterpart. In particular, the robustness of the transformation $x_i^{PCA} \to x_i^{PAA}$ means that it is not unduly distracted by cases from relatively rare groups with means $\overline{x}_g^*$ lying outside the span of the others.

In the exploratory spirit of Section 1.2, we recommend examining a range of dimensionalities suggested by the empirical alignment plot. Experience shows that this procedure works well, the transformed cluster mean subspace $\mathcal{M}^*$

being well estimated by the first principal axis in Examples 1 and 2 above, and by the span of the first two in Example 3.

4 FURTHER EXAMPLES

The two examples discussed here illustrate the robustness of principal axis analysis to departures from the elliptical mixture model (1) and the complementarity of its sphered and unsphered forms. In Examples 1 to 3 above the sphered form offered the more revealing view, while in the examples here the unsphered form shows little correlation between its first two elements and offers the better view.

Examples 4 and 5 concern a large data set on pen-based recognition of handwritten digits (available at http://www.ics.uci.edu/mlearn/MLSummary.html) containing some 1100 observations on the digits 0 to 9, $k = 16$ variables being automatically retrieved as the horizontal and vertical coordinates of 8 predefined points of each digit. It has been used as a challenge data set for several discrimination procedures. Here, we proceed without using group membership information.

Example 4 concerns the digits 4 and 6, denoted '$\triangle$' and '+' respectively. The scatterplotmatrix of the data shows clear departures from (1), the subplot for the first four variables being given as Figure 11. Nevertheless, a single principal axis is clearly preferred in the empirical alignment plot (not shown), the two-group structure of the data being seen along the first principal axis in the sphered scatterplot Figure 12 and, more clearly, in its unsphered version, Figure 13.

FIGURES 11 TO 13 ABOUT HERE

Example 5 concerns the digits 5 and 8, denoted '$\triangle$' and '+' respectively, the scatterplotmatrix of the data (not shown) again showing clear departures from (1). A known feature of this data set is that there are two distinct subgroups of observations for the digit 5 depending on how its top-left part is written, as a pronounced right-angle or as a curve. This results, in effect, in there being three groups in the data, '5', 'S' and '8' say. In the sphered analysis (Figure 14), the first of these groups is separated out on the first principal axis, but the other two overlap. The unsphered analysis (Figure 15) separates all three groups on the first principal axis, the '8's being central. However, their means are not collinear. Indeed, the empirical alignment plot (Figure 16) suggests at least 3-D structure. This is confirmed in the rotated 3-D plot of Figure 17 where the '8's, here shown as the larger '$\triangle$' symbol, are seen to have their own internal structure.

FIGURES 14 TO 17 ABOUT HERE

5 FURTHER POINTS OF CONTACT

5.1 A basic decomposition

Working again in the population, some further points of contact of principal axis analysis flow from the between-plus-within decomposition of the covariance matrix under (1) as:

$$\Omega = B + W \text{ where } B = \sum_{g=1}^{G} \pi_g (\mu_g - \mu)(\mu_g - \mu)^T \text{ and } W = \sum_{g=1}^{G} \pi_g \Omega_g. \quad (2)$$

A key fact is that the range of B is the cluster mean subspace $\mathcal{M}$.

The decomposition (2) establishes direct points of contact with multiple analysis of variance. However, PAA is more widely applicable, being useable when group membership and/or the number of groups is unknown and placing fewer restrictions on the distributions $\{F_g\}_{g=1}^{G}$, such as normality or equal covariances $\{\Omega_g\}_{g=1}^{G}$. PAA is similarly more general than linear discriminant analysis, discussed next.

One way to think of each of principal component analysis, principal axis analysis and linear discriminant analysis is that they reduce dimension while preserving, as far as possible, a specified feature of a corresponding matrix M. In terms of the decomposition (2), we have

$$M = \Omega \text{ in PCA, } M = B \text{ in PAA and } M = W^{-1}B \text{ in LDA,}$$

preservation of the range of B being implicit in PAA.

5.2 Complementarity of PCA and PAA

Principal component and principal axis analysis are complementary methods with logically distinct objectives. Example 1 illustrates this well: the maximal cluster structure and maximal variance directions being almost orthogonal, they provide complementary summaries of this bivariate data set. Again, Examples 2 and 3 illustrate that a method aimed at explaining variability can miss other important features, even when strongly present. Nevertheless, it is possible for PCA and PAA to perform similarly – as with the Fisher iris data, where principal component analysis in known to perform well as an implicit cluster analysis method. The decomposition (2) can be used to throw light on these findings by quantifying the complementarity of PCA and PAA, as follows.

The greater the percentage $TV(\mathcal{M})$ of total variance within the mean subspace $\mathcal{M}$, the greater the potential for principal component analysis to reveal its cluster structure. Noting that $m = \dim(\mathcal{M}) = \text{rank}(B)$, let $B = Q^B \Lambda^B (Q^B)^T$ with $\Lambda^B = \text{diag}(\lambda_1^B, ..., \lambda_m^B, 0, ..., 0)$ and $Q^B = (q_1^B | ... | q_k^B)$ orthogonal. Then, using (2), we have

$$TV(\mathcal{M}) = 100 \times \frac{\sum_{r=1}^{m} \text{var}((q_r^B)^T x)}{\text{trace}(\Omega)} = 100 \times \frac{\sum_{r=1}^{m} (\lambda_r^B + w_{rr}^B)}{\sum_{r=1}^{k} \lambda_r},$$

where w_{rr}^B denotes the r^{th} diagonal element of $W^B = (Q^B)^T W Q^B$.

For data with known group memberships and estimated dimension $\widehat{m}$, we may compute $TV(\widehat{\mathcal{M}})$ where, in an obvious notation, $\widehat{\mathcal{M}} = \text{Span}\{\widehat{q}_1^B, ..., \widehat{q}_{\widehat{m}}^B\}$. For example, this numerical summary is 97% for the Fisher iris data, 52% in Example 1, but just 12% in Example 3 and 11% in Example 2.

5.3 Two scatter matrices are better than one

Independent recent work by Lutz Duembgen, Hannu Oja and Dave Tyler has shown that the ability of principal axis analysis to discriminate, without knowing group membership, is shared by other methods including certain versions of independent component analysis (Hyvarinen et al., 2001). A joint paper is planned on this and related topics, emphasising the joint use of two scatter matrices. A direct link between this theme and principal axis analysis arises as follows. See also Bugrien and Kent (2005).

We note first that the affine transformations $x \to \widetilde{x}$ making $\widetilde{x}$ sphered are precisely those of the form $x \to Ox^*$ for some orthogonal matrix O and that the sphered version of PAA does not depend on the particular choice $O = I$ made above. For, replacing x^* by $\widetilde{x} = Ox^*$, $A \to \widetilde{A} = OAO^T$ so that $U \to \widetilde{U} = OU$, leaving x^{PAA} unchanged.

Taking now $O = Q$, as we may, and denoting by $\Omega^{-\frac{1}{2}}$ the unique symmetric square root of Ω^{-1}, the sphered version of PAA is obtained by transforming $x \to \widetilde{x} = \Omega^{-\frac{1}{2}}(x - \mu)$ and then $\widetilde{x} \to x^{PAA} = \widetilde{U}^T\widetilde{x}$ where $\widetilde{U}$ is an orthogonal matrix of eigenvectors of

$$\widetilde{A} = \Omega^{-\frac{1}{2}}\widetilde{\Omega}\Omega^{-\frac{1}{2}} \text{ in which } \widetilde{\Omega} = E\left(\frac{(x-\mu)(x-\mu)^T}{(x-\mu)^T\Omega^{-1}(x-\mu)}\right).$$

In this last expression, both Ω and $\widetilde{\Omega}$ are scatter matrices in the generalised sense of affine equivariant, nonnegative definite, symmetric matrices, the latter downweighting observations at large Mahalanobis distances from the mean.

6 CONCLUSION

Overall, our message is clear. When using principal component analysis as an exploratory tool, we recommend performing a complementary principal axis analysis. This costs essentially nothing, while two views are better than one. In particular, PAA can reveal important structure missed by PCA.

Several possible further developments were noted in Section 1.2. Others include (a) accommodating singular covariance matrices, as arises when there are more variables than cases, by using a generalised inverse in the sphering step, (b) replacing mean and covariance by general location and scatter measures respectively, (c) adapting principal axis analysis to specific problems arising with gene expression data, and (d) extending its current Euclidean geometry to Riemannian geometries appropriate to clustering curves and surfaces. There may also be helpful connections to partial least squares.

Acknowledgements

The authors are grateful to many colleagues for helpful comments and for several sources of research support: the European Science Foundation SACD network, the Open University, U.K. (Frank Critchley) and the Programa Operacional 'Ciência, Tecnologia, Inovação' (POCTI) of the Fundação para a Ciência e a Tecnologia (FCT), cofinanced by the European Community fund FEDER (A. M. Pires and C. Amado).

References

Bugrien, J.B. & Kent, J.T. (2005). Independent component analysis: An approach to clustering. In *Quantitative Biology, Shape Analysis, and Wavelets*, Eds. S. Barber, P.D. Baxter, K.V. Mardia & R.E. Walls, pp. 111-114. Leeds: Leeds University Press.

Croux, C. & Haesbroeck, G. (2000). Principal component analysis based on robust estimates of the covariance and correlation matrix: influence functions and efficiencies. *Biometrika*, **72**, 603-18.

Croux, C., Ollila, E. & Oja, H. (2002). Sign and rank covariance matrices: statistical properties and application to principal components analysis. In *Statistical Data Analysis Based on the L1-Norm and Related Methods*, Ed. Y. Dodge, pp. 257-69. Basel: Birkhaüser.

Cuesta-Albertos, J. A., Gordaliza, A. & Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *Annals of Statistics*, **25**, 553-76.

Everitt, B. S., Landau, S. & Leese, M. (2001). *Cluster Analysis*. Oxford: Oxford University Press.

Fang, K.-T., Kotz, S., & Ng, K.-W. (1990). *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall.

Gordon, A. D. (1981). *Classification*. London: Chapman and Hall.

Hermans, J. & Habbema, J. D. F. (1975). Comparison of five methods to estimate posterior probabilities, *EDV in Medizin und Biologie*, **6**, 14-9.

Hubert, M., Rousseeuw, P. J. & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **47**, 64-79.

Hyvarinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*. New York: Wiley.

Johnson, R.A. & Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. Third Edition. London: Prentice-Hall.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. & Kohen, K. L. (1999). Robust principal components for functional data. *Test*, **8**, 1-73.

Marden, J. I. (1999). Some robust estimates of principal components, *Statistics and Probability Letters*, **43**, 349-59.

Mardia, K.V. & Jupp, P.E. (2000). *Directional Statistics.* New York: Wiley.

Peña, D. Y. & Prieto, F. J. (2001). Cluster identification using projections. *Journal of the American Statistical Association*, **96**, 1433-45.

Visuri, S., Ollila, E., Koivunen, V., Möttönen, J. & Oja H. (2003). Affine equivariant multivariate rank methods. *Journal of Statistical Planning and Inference*, **114**, 161-85.

Figures

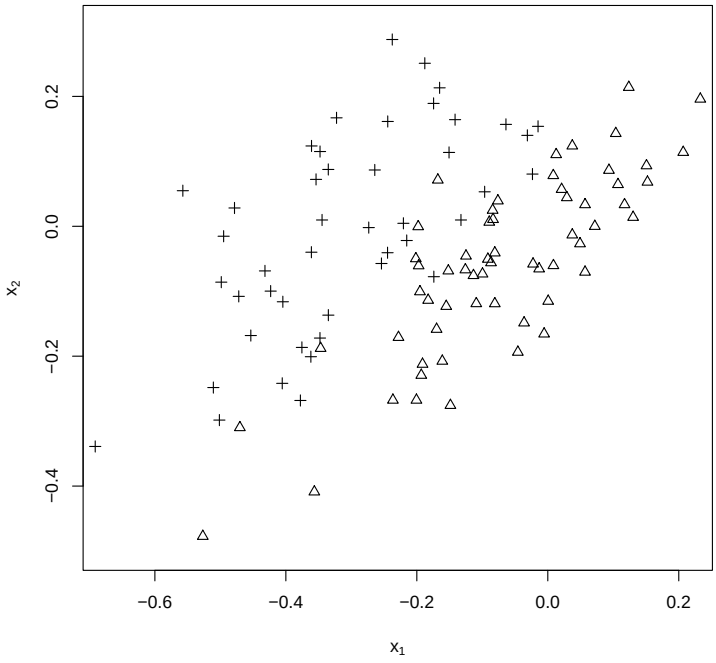

Fig. 1. Example 1: scatterplot of the data.

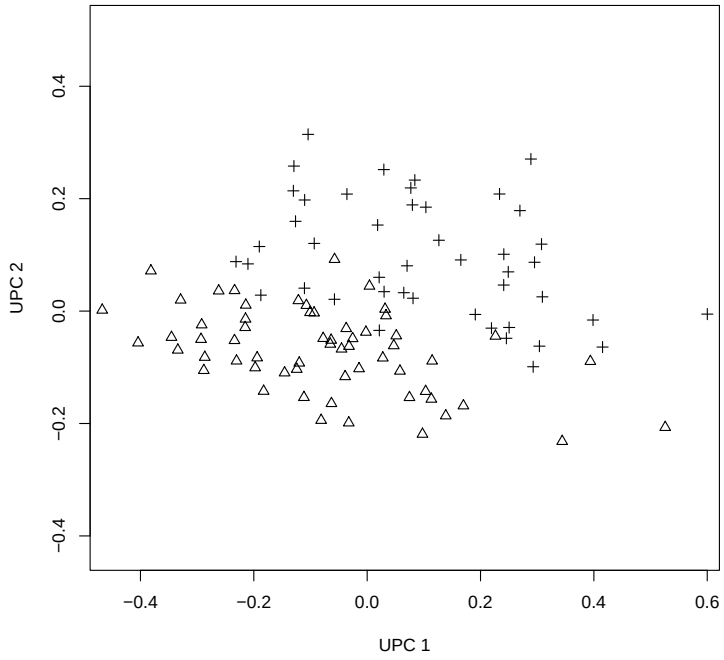

Fig. 2. Example 1: scatterplot of the unsphered PCA scores.

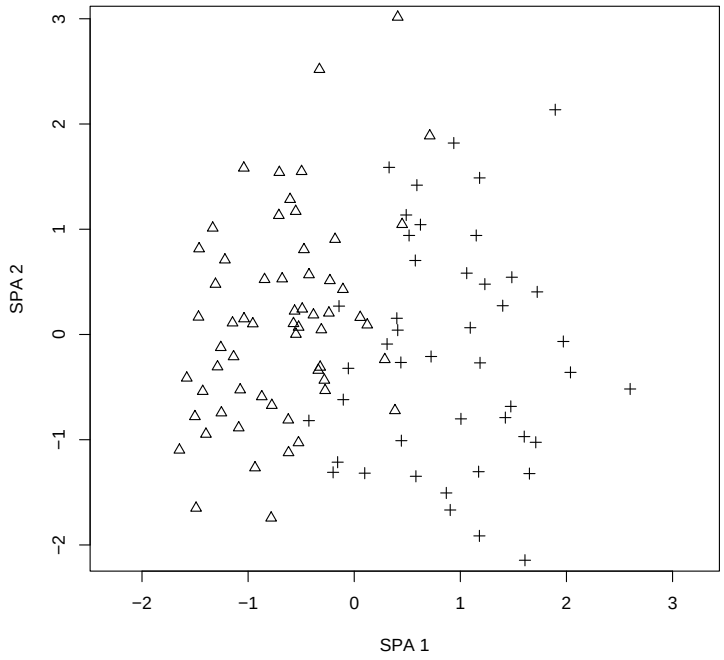

Fig. 3. Example 1: scatterplot of the PAA scores of the sphered PCA scores.

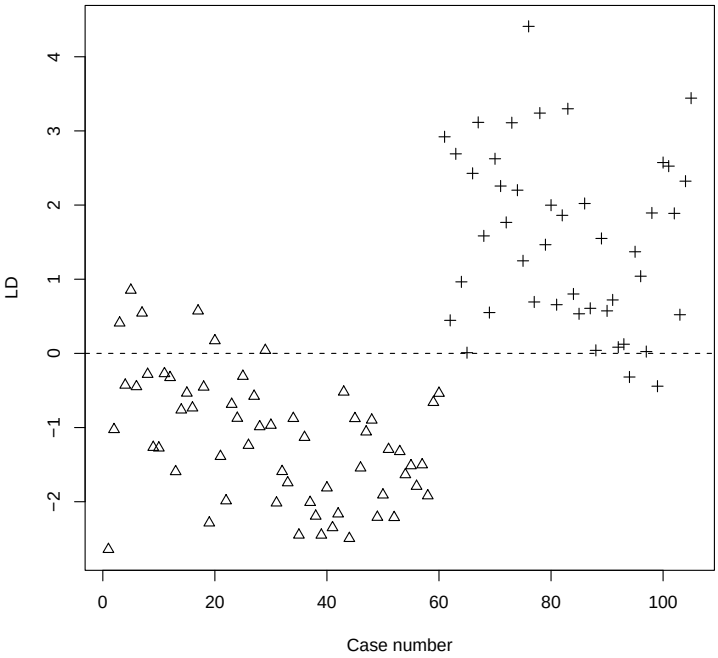

Fig. 4. Example 1: plot of LDA score against case number.

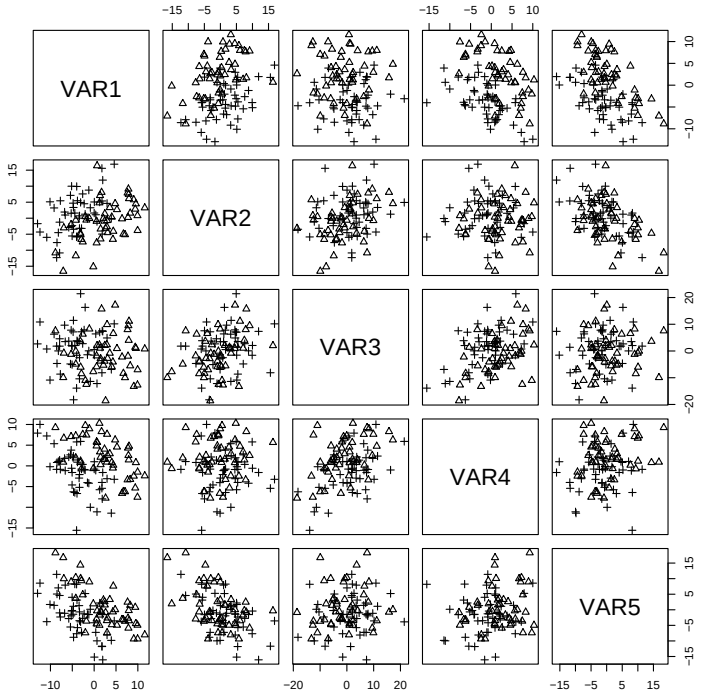

Fig. 5. Example 2: scatterplotmatrix of the data.

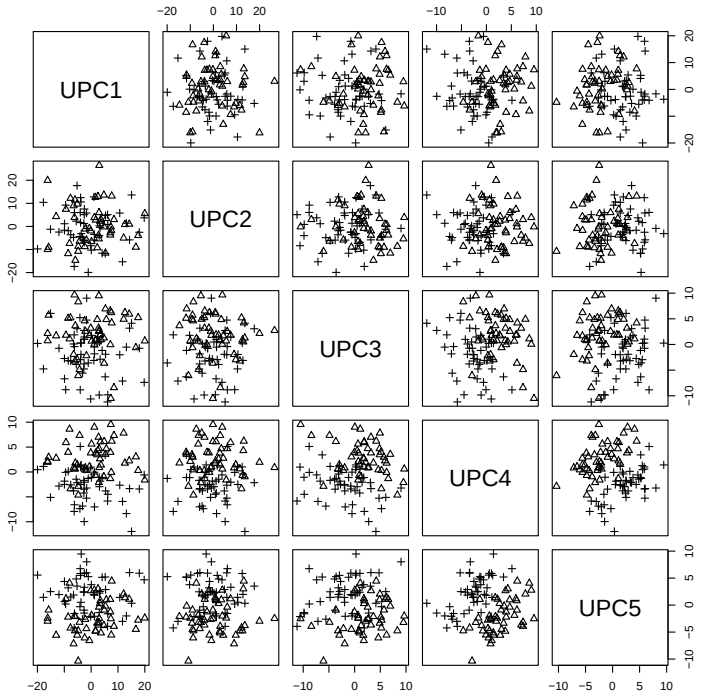

Fig. 6. Example 2: scatterplotmatrix of the unsphered PCA scores.

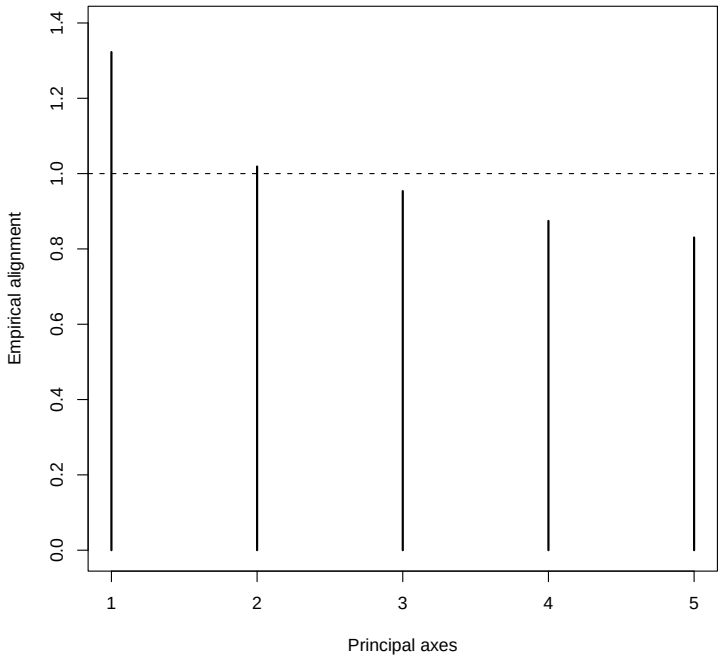

Fig. 7. Example 2: empirical alignment plot.

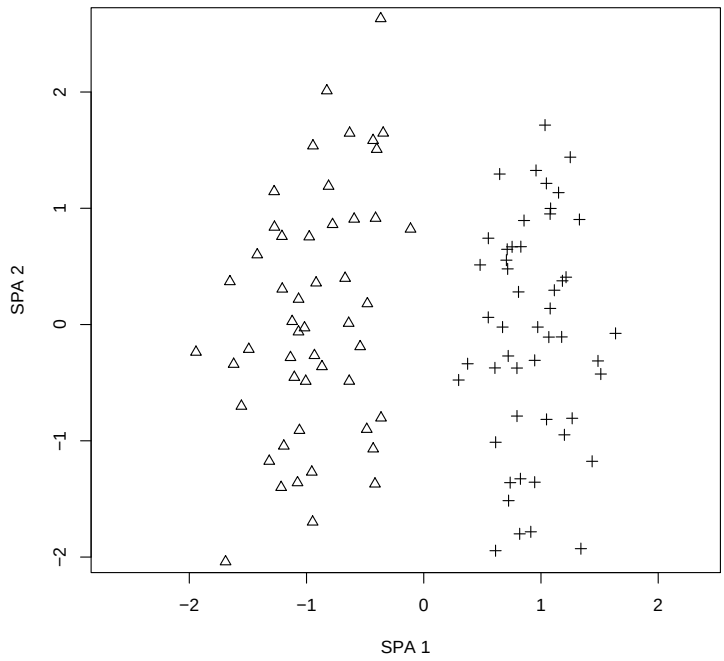

Fig. 8. Example 2: scatterplot of the first two PAA scores of the sphered PCA scores.

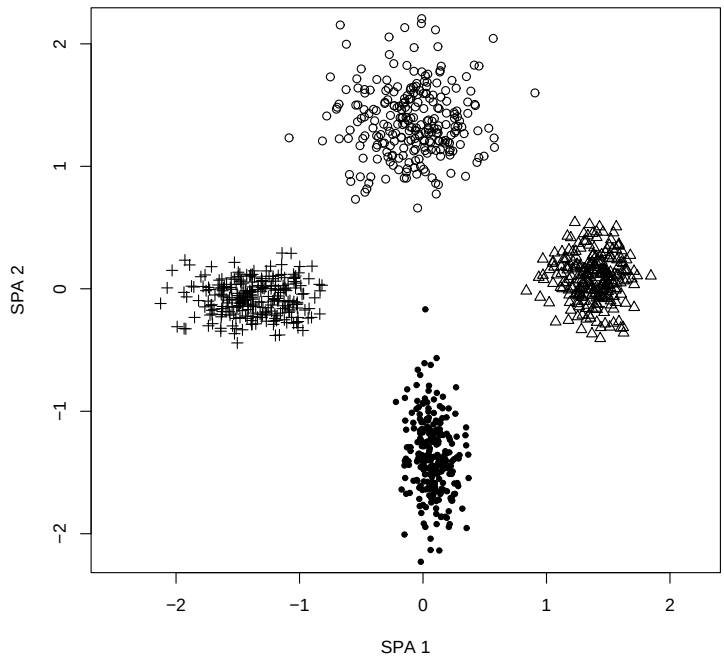

Fig. 9. Example 3: scatterplot of the first two PAA scores of the sphered PCA scores.

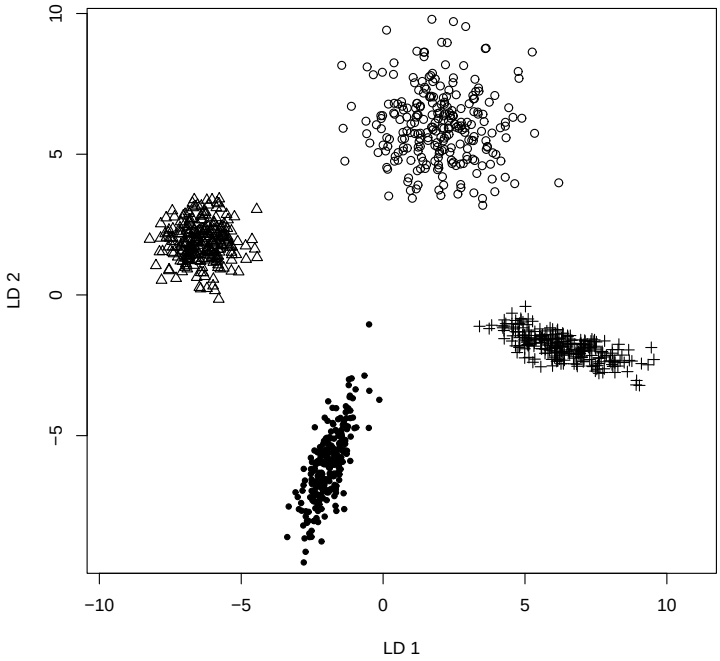

Fig. 10. Example 3: scatterplot of the first two LDA scores.

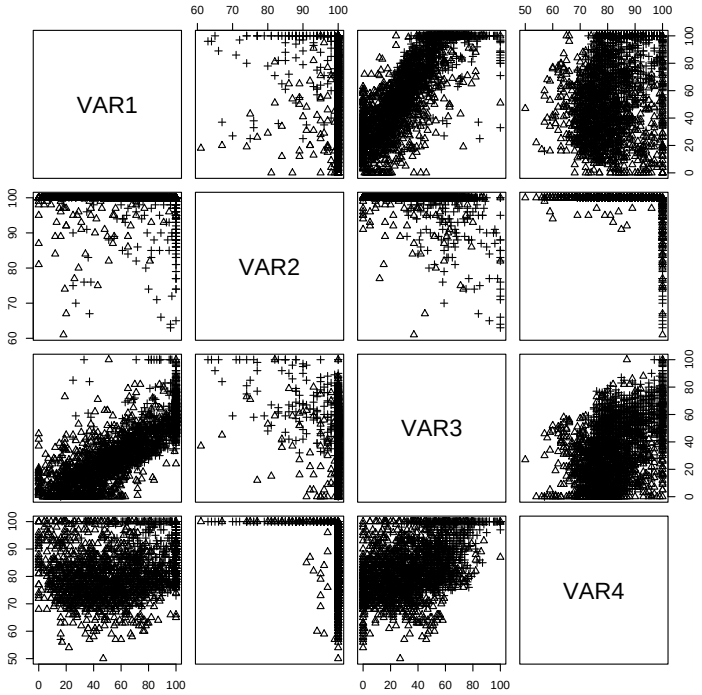

Fig. 11. Example 4: scatterplotmatrix of the first four variables.

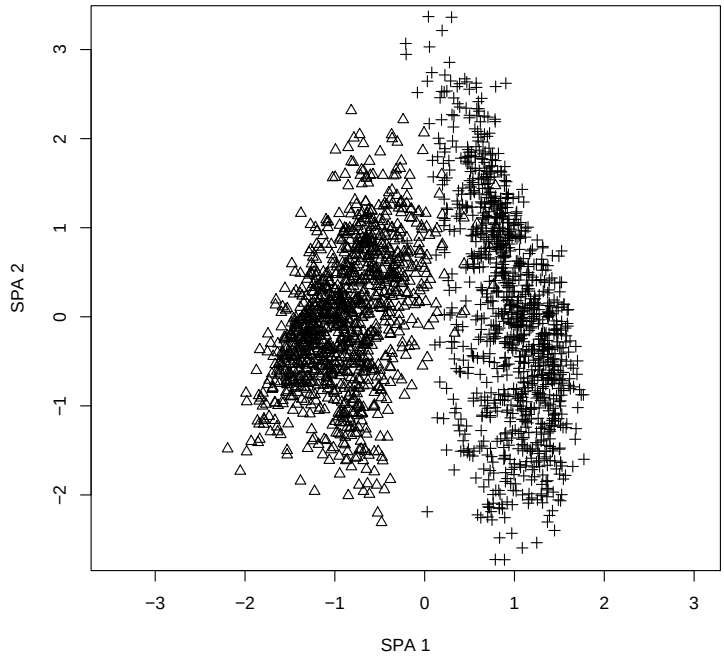

Fig. 12. Example 4: scatterplot of the first two PAA scores of the sphered PCA scores.

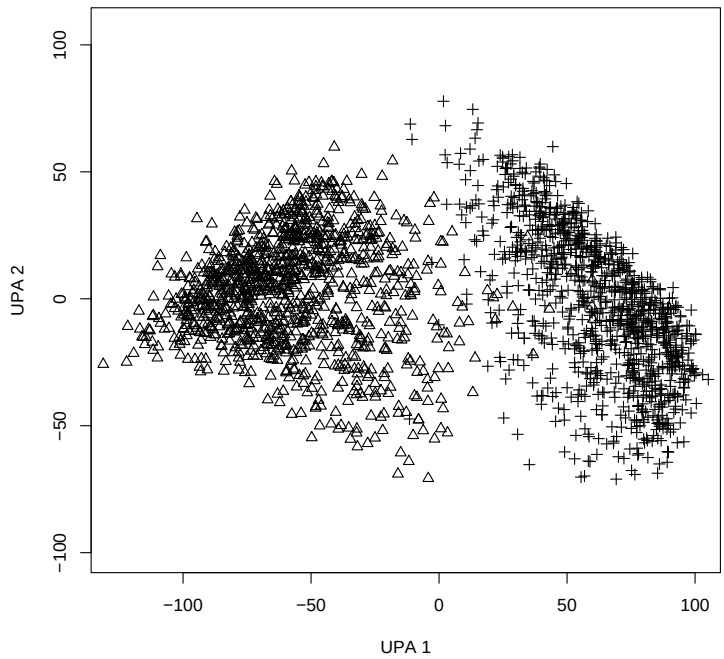

Fig. 13. Example 4: scatterplot of the first two PAA scores of the unsphered PCA scores.

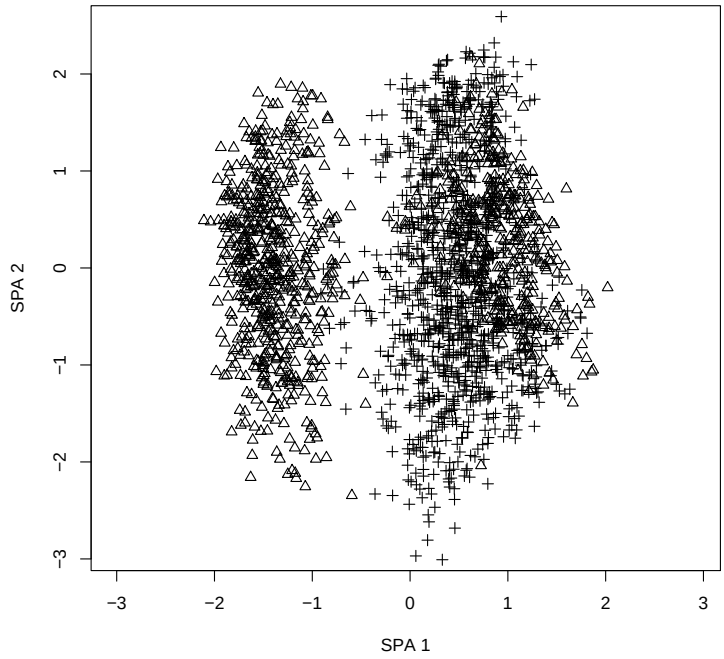

Fig. 14. Example 5: scatterplot of the first two PAA scores of the sphered PCA scores.

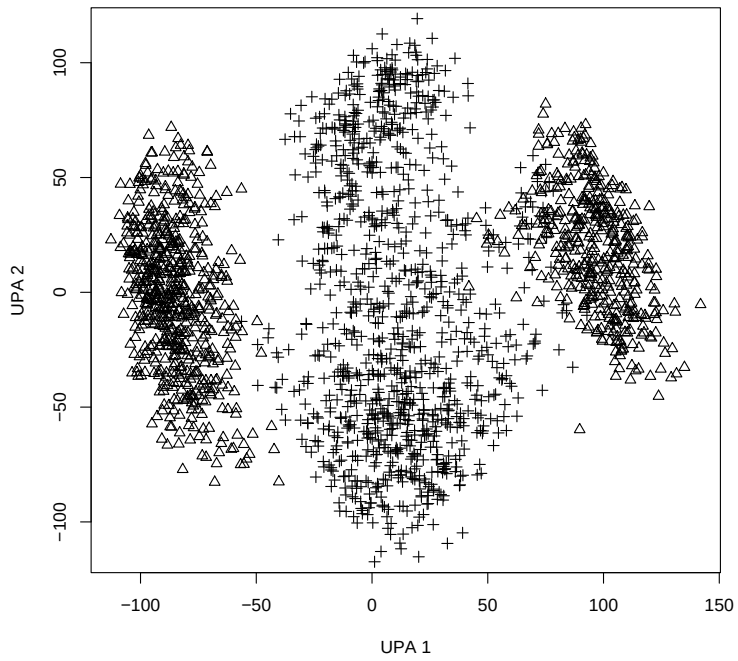

Fig. 15. Example 5: scatterplot of the first two PAA scores of the unsphered PCA scores.

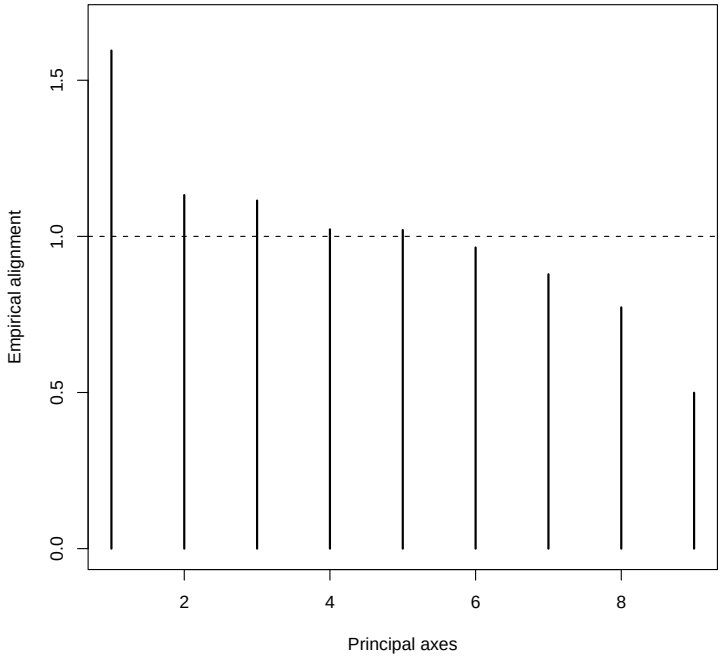

Fig. 16. Example 5: empirical alignment plot.

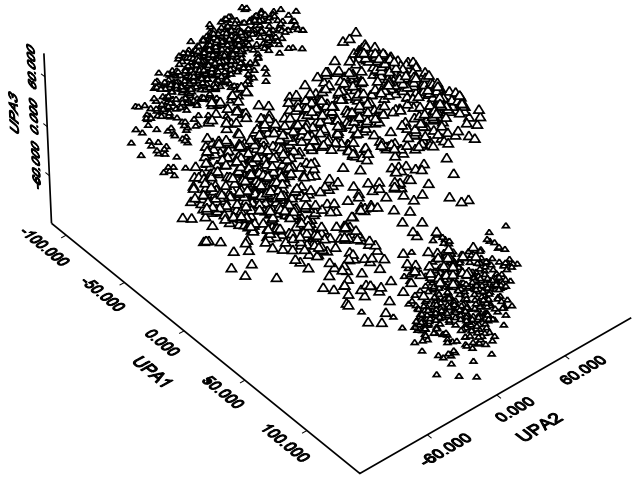

Fig. 17. Example 5: rotated 3-D plot of the first three PAA scores of the unsphered PCA scores.