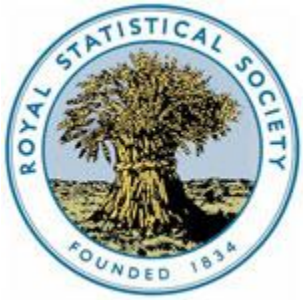

Invariant Co-Ordinate Selection [with Discussion]
Author(s): David E. Tyler, Frank Critchley, Lutz Dümbgen and Hannu Oja
Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 71, No. 3 (Jun., 2009), pp. 549-592
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/40247589
Accessed: 23-11-2015 17:50 UTC

J. R. Statist. Soc. B (2009)
71, *Part* 3, *pp.* 549–592

Invariant co-ordinate selection

David E. Tyler,

Rutgers University, Piscataway, USA

Frank Critchley,

The Open University, Milton Keynes, UK

Lutz Dümbgen

University of Berne, Switzerland

and Hannu Oja

University of Tampere, Finland

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, December 17th, 2008*, Professor I. L. Dryden *in the Chair*]

Summary. A general method for exploring multivariate data by comparing different estimates of multivariate scatter is presented. The method is based on the eigenvalue–eigenvector decomposition of one scatter matrix relative to another. In particular, it is shown that the eigenvectors can be used to generate an affine invariant co-ordinate system for the multivariate data. Consequently, we view this method as a method for *invariant co-ordinate selection*. By plotting the data with respect to this new invariant co-ordinate system, various data structures can be revealed. For example, under certain independent components models, it is shown that the invariant co-ordinates correspond to the independent components. Another example pertains to mixtures of elliptical distributions. In this case, it is shown that a subset of the invariant co-ordinates corresponds to Fisher's linear discriminant subspace, even though the class identifications of the data points are unknown. Some illustrative examples are given.

Keywords: Affine invariance; Cluster analysis; Independent components analysis; Mixture models; Multivariate diagnostics; Multivariate scatter; Principal components; Projection pursuit; Robust statistics

1. Introduction

When sampling from a multivariate normal distribution, the sample mean vector and sample variance–covariance matrix are a sufficient summary of the data set. To protect against non-normality, and in particular against longer-tailed distributions and outliers, we can replace the sample mean and covariance matrix with robust estimates of multivariate location and scatter (or pseudocovariance). A variety of robust estimates of the multivariate location vector and scatter matrix have been proposed. Among them are multivariate M-estimates (Huber, 1981; Maronna, 1976), the minimum volume ellipsoid estimate and the minimum covariance determinant estimate (Rousseeuw, 1986), S-estimates (Davies, 1987; Lopuhaä, 1985), projection-based estimates (Maronna *et al.*, 1992; Tyler, 1994), τ-estimates (Lopuhaä, 1991), constrained M-estimates (Kent and Tyler, 1996) and MM estimates (Tatsuoka and Tyler, 2000; Tyler, 2002), as well

Address for correspondence: David E. Tyler, Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019, USA.
E-mail: dtyler@rci.rutgers.edu

 1369–7412/09/71549

as one-step versions of these estimates (Lopuhaä, 1999). After computing robust estimates of multivariate location and scatter, outliers can often be detected by examining the corresponding robust Mahalanobis distances; see for example Rousseeuw and Leroy (1987).

Summarizing a multivariate data set via a location and a scatter statistic, and then inspecting the corresponding Mahalanobis distance plot for possible outliers, is appropriate if the bulk of the data arises from a multivariate normal distribution or, more generally, from an elliptically symmetric distribution. However, if the data arise from a distribution which is not symmetric, then different location statistics are estimating different notions of central tendency. Moreover, if the data arise from a distribution other than an elliptically symmetric distribution, even one which is symmetric, then different scatter statistics are not necessarily estimating the same population quantity but rather are reflecting different aspects of the underlying distribution. This suggests that comparing different estimates of multivariate scatter may help to reveal interesting departures from an elliptically symmetric distribution. Such data structures may not be apparent in a Mahalanobis distance plot.

In this paper, we present a general multivariate method based on the comparison of different estimates of multivariate scatter. This method is based on the eigenvalue–eigenvector decomposition of one scatter matrix relative to another. An important property of this decomposition is that the corresponding eigenvectors generate an affine invariant co-ordinate system for the multivariate observations, and so we view this method as a method for *invariant co-ordinate selection* (ICS). By plotting the data with respect to this new invariant co-ordinate system, various data structures can be revealed. For example, when the data arise from a mixture of elliptical distributions, the space that is spanned by a subset of the invariant co-ordinates gives an estimate of Fisher's linear discriminant subspace, even though the class identifications of the data points are unknown. Another example pertains to certain independent components models. Here the variables that are obtained by using the invariant co-ordinates correspond to estimates of the independent components.

The paper is organized as follows. Section 2 sets up some notation and concepts to be used in the paper. In particular, the general concept of affine equivariant scatter matrices is reviewed in Section 2.1 and some classes of scatter matrices are briefly reviewed in Section 2.2. The idea of comparing two different scatter matrices by using the eigenvalue–eigenvector decomposition of one scatter matrix relative to another is discussed in Section 3, with the invariance properties of the ICS transformation being given in Section 4. Section 5 gives a theoretical study of the ICS transformation under the aforementioned elliptical mixture models (Section 5.1), and under independent components models (Section 5.2). The results in Section 5.1 represent a broad generalization of results given under the heading of *generalized principal components analysis* by Ruiz-Gazen (1993) and Caussinus and Ruiz-Gazen (1993, 1995). Readers who are primarily interested in how ICS works in practice may wish to skip Section 5 at a first reading. In Section 6, a general discussion on the choice of scatter matrices that we may consider when implementing ICS, along with some examples illustrating the utility of the ICS transformation for diagnostic plots, is given. Further discussion, open research questions and the relationship of ICS to other approaches are given in Section 7. All formal proofs are reserved for Appendix A. An R package entitled ICS (Nordhausen, Oja and Tyler, 2008) is freely available for implementing the ICS methods.

2. Scatter matrices

2.1. Affine equivariance

Let F_Y denote the distribution function of the multivariate random variable $Y \in \Re^p$, and let $\mathcal{P}_p$ represent the set of all symmetric positive definite matrices of order p. Affine equivariant

multivariate location and scatter functionals, say $\mu(F_Y) \in \Re^p$ and $V(F_Y) \in \mathcal{P}_p$ respectively, are functions of the distribution satisfying the property that for $Y^* = AY + b$, with A non-singular and $b \in \Re^p$,

$$\begin{aligned} \mu(F_{Y^*}) &= A\,\mu(F_Y) + b, \\ V(F_{Y^*}) &= A\,V(F_Y)A'. \end{aligned} \tag{1}$$

Classical examples of affine equivariant location and scatter functionals are the mean vector $\mu_Y = E[Y]$ and the variance–covariance matrix $\Sigma_Y = E[(Y-\mu_Y)(Y-\mu_Y)']$ respectively, provided that they exist. For our purposes, affine equivariance of the scatter matrix can be relaxed slightly to require only affine equivariance of its shape components. A shape component of a scatter matrix $V \in \mathcal{P}_p$ refers to any function of V, say $\mathcal{S}(V)$, such that

$$\mathcal{S}(V) = \mathcal{S}(\lambda V) \qquad \text{for any } \lambda > 0. \tag{2}$$

Thus, we say that the 'shape' of $V(F_Y)$ is affine equivariant if

$$V(F_{Y^*}) \propto A\,V(F_Y)A'. \tag{3}$$

For a p-dimensional sample of size n, $\mathbf{Y} = \{y_1, \ldots, y_n\}$, affine equivariant multivariate location and scatter statistics, say $\hat{\mu}$ and $\hat{V}$ respectively, are defined by applying the above definition to the empirical distribution function, i.e. they are statistics satisfying the property that, for any non-singular A and any $b \in \Re^p$,

$$y_i \to y_i^* = Ay_i + b \text{ for } i = 1, \ldots, n \Rightarrow (\hat{\mu}, \hat{V}) \to (\hat{\mu}^*, \hat{V}^*) = (A\hat{\mu} + b, A\hat{V}A'). \tag{4}$$

Likewise, the shape of $\hat{V}$ is said to be affine equivariant if

$$\hat{V}^* \propto A\hat{V}A'. \tag{5}$$

The sample mean vector $\bar{y}$ and sample variance–covariance matrix S_n are examples of affine equivariant location and scatter statistics respectively, as are all the estimates that were cited in Section 1.

Typically, in practice, $\hat{V}$ is normalized so that it is consistent at the multivariate normal model for the variance–covariance matrix. The normalized version is thus given as $\tilde{V} = \hat{V}/\beta$, where $\beta > 0$ is such that $V(F_Z) = \beta I$ when Z has a standard multivariate normal distribution. For our purposes, it is sufficient to consider only the unnormalized scatter matrix $\hat{V}$ since our proposed methods depend only on the scatter matrix up to proportionality, i.e. only on the shape of the scatter matrix.

Under elliptical symmetry, affine equivariant location and scatter functionals have relatively simple forms. Recall that an elliptically symmetric distribution is defined to be one arising from an affine transformation of a spherically symmetric distribution, i.e., if $Z \sim QZ$ for any $p \times p$ orthogonal matrix Q, then the distribution of $Y = AZ + \mu$ is said to have an elliptically symmetric distribution with centre $\mu \in \Re^p$ and shape matrix $\Gamma = AA'$; see for example Bilodeau and Brenner (1999). If the distribution of Y is also absolutely continuous, then it has a density of the form

$$f(y; \mu, \Gamma, g) = \det(\Gamma)^{-1/2}\, g\{(y-\mu)'\Gamma^{-1}(y-\mu)\} \qquad \text{for } y \in \Re^p, \tag{6}$$

for some non-negative function g and with $\Gamma \in \mathcal{P}_p$. As defined, the shape parameter Γ of an elliptically symmetric distribution is only well defined up to a scalar multiple, i.e., if Γ satisfies the definition of a shape matrix for a given elliptically symmetric distribution, then $\lambda\Gamma$ also does for any $\lambda > 0$. In the absolutely continuous case, if no restrictions are placed on

the function g, then the parameter Γ is confounded with g. One could normalize the shape parameter by setting, for example, $\det(\Gamma) = 1$ or $\mathrm{tr}(\Gamma) = p$. Again, this is not necessary for our purposes since only the shape components of Γ, as defined in expression (2), are of interest in this paper, and these shape components for an elliptically symmetric distribution are well defined.

Under elliptical symmetry, any affine equivariant location functional corresponds to the centre of symmetry and any affine equivariant scatter functional is proportional to the shape matrix, i.e. $\mu(F_Y) = \mu$ and $V(F_Y) \propto \Gamma$. In particular, $\mu_Y = \mu$ and $\Sigma_Y \propto \Gamma$ when the first and second moments exist respectively. More generally, if $V(F_Y)$ is any functional satisfying condition (3), then $V(F_Y) \propto \Gamma$.

As noted in Section 1, for general distributions, affine equivariant location functionals are not necessarily equal and affine equivariant scatter functionals are not necessarily proportional to each other. The corresponding sample versions of these functionals are therefore estimating different population features. The difference in these functionals reflects in some way how the distribution differs from an elliptically symmetric distribution.

Remark 1. The class of distributions for which all affine equivariant location functionals are equal and all equivariant scatter functionals are proportional to each other is broader than the class of elliptical distributions. For example, this can be shown to be true for F_Y when $Y = AZ + \mu$ with the distribution of Z being exchangeable and symmetric in each component, i.e. $Z \sim DJZ$ for any permutation matrix J and any diagonal matrix D having diagonal elements ± 1. We conjecture that this is the broadest class for which this property holds. This class contains the elliptical symmetric distributions, since these correspond to Z having a spherically symmetric distribution.

2.2. Classes of scatter statistics

Conceptually, the simplest alternatives to the sample mean $\bar{y}$ and sample covariance matrix S_n are the weighted sample means and sample covariance matrices respectively, with the weights dependent on the classical Mahalanobis distances. These are defined by

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{n} u_1(s_{0,i}) y_i \Big/ \sum_{i=1}^{n} u_1(s_{0,i}), \\ \hat{V} &= \sum_{i=1}^{n} u_2(s_{0,i})(y_i - \bar{y})(y_i - \bar{y})' \Big/ \sum_{i=1}^{n} u_2(s_{0,i}), \end{aligned} \tag{7}$$

where $s_{0,i} = (y_i - \bar{y})' S_n^{-1} (y_i - \bar{y})$, and $u_1(s)$ and $u_2(s)$ are some appropriately chosen weight functions. Other simple alternatives to the sample covariance matrix can be obtained by applying only the scatter equation above to the sample of pairwise differences, i.e. to the symmetrized data set

$$\mathbf{Y}^{\mathrm{s}} = \{ y_i - y_j \mid i, j = 1, \ldots, n,\ i \neq j \}, \tag{8}$$

for which the sample mean is 0. Even though the weighted mean and covariance matrix, as well as the symmetrized version of the weighted covariance matrix, may downweight outliers, they have unbounded influence functions and zero breakdown points.

A more robust class of multivariate location and scatter statistics is given by the multivariate M-estimates, which can be viewed as adaptively weighted sample means and sample covariance matrices respectively. More specifically, they are defined as solutions to the M-estimating equations

$$\hat{\mu} = \sum_{i=1}^{n} u_1(s_i) y_i \Big/ \sum_{i=1}^{n} u_1(s_i),$$
$$\hat{V} = \sum_{i=1}^{n} u_2(s_i)(y_i - \hat{\mu})(y_i - \hat{\mu})' \Big/ \sum_{i=1}^{n} u_3(s_i), \tag{9}$$

where $s_i = (y_i - \hat{\mu})' \hat{V}^{-1} (y_i - \hat{\mu})$, and $u_1(s)$, $u_2(s)$ and $u_3(s)$ are again some appropriately chosen weight functions. We refer the reader to Huber (1981) and Maronna (1976) for the general theory regarding the multivariate M-estimates. The equations given in expression (9) are implicit equations in $(\hat{\mu}, \hat{V})$ since the weights depend on the Mahalanobis distances relative to $(\hat{\mu}, \hat{V})$, i.e. on $d_i(\hat{\mu}, \hat{V}) = \sqrt{s_i}$. Nevertheless, relatively simple algorithms exist for computing the multivariate M-estimates. The maximum likelihood estimates of the parameters μ and Γ of an elliptical distribution for a given spread function g in expression (6) are special cases of M-estimates.

From a robustness perspective, an often-cited drawback to the multivariate M-estimates is their relatively low breakdown in higher dimension. Specifically, their breakdown point is bounded above by $1/(p+1)$. Subsequently, numerous high breakdown point estimates have been proposed, such as the minimum volume ellipsoid, the minimum covariance determinant, the S-estimates, the projection-based estimates, the τ-estimates, the constrained M-estimates and the MM estimates, all of which are cited in Section 1. All the high breakdown point estimates are computationally intensive and, except for small data sets, are usually computed by using approximate or probabilistic algorithms. The computational complexity of high breakdown point multivariate estimates is especially challenging for extremely large data sets in high dimensions, and this remains an open and active area of research.

The definition of the weighted sample means and covariance matrices given by expression (7) can be readily generalized by using any initial affine equivariant location and scatter statistic, say $\hat{\mu}_0$ and $\hat{V}_0$ respectively, i.e.

$$\hat{\mu} = \sum_{i=1}^{n} u_1(s_{0,i}) y_i \Big/ \sum_{i=1}^{n} u_1(s_{0,i}),$$
$$\hat{V} = \sum_{i=1}^{n} u_2(s_{0,i})(y_i - \hat{\mu}_0)(y_i - \hat{\mu}_0)' \Big/ \sum_{i=1}^{n} u_2(s_{0,i}), \tag{10}$$

where now $s_{0,i} = (y_i - \hat{\mu}_0)' \hat{V}_0^{-1} (y_i - \hat{\mu}_0)$. In the univariate setting such weighted sample means and variances are sometimes referred to as one-step W-estimates (Hampel *et al.*, 1986; Mosteller and Tukey, 1977), and so we refer to their multivariate versions as multivariate one-step W-estimates. Given a location and a scatter statistic, a corresponding one-step W-estimate provides a computationally simple choice for an alternative location and scatter statistic.

Any method that one uses for obtaining location and scatter statistics for a data set $\mathbf{Y}$ can also be applied to its symmetrized version $\mathbf{Y}^s$ to produce a scatter statistic. For symmetrized data, any affine equivariant location statistic is always 0.

The functional or population versions of the location and scatter statistics that were discussed in this section are readily obtained by replacing the empirical distribution of $\mathbf{Y}$ with the population distribution function F_Y. For the M-estimates and the one-step W-estimates, this simply implies replacing the averages in expressions (9) and (10) respectively with expected values. For symmetrized data, the functional versions are obtained by replacing the empirical distribution of $\mathbf{Y}^s$ with its almost sure limit F_Y^s, the distribution function of $Y^s = Y_1 - Y_2$, where Y_1 and Y_2 are independent copies of Y.

3. Comparing scatter matrices

Comparing positive definite symmetric matrices arises naturally within a variety of multivariate statistical problems. Perhaps the most obvious case is when we wish to compare the covariance structures of two or more different groups; see for example Flury (1988). Other well-known cases occur in multivariate analysis of variance, wherein interest lies in comparing the within-group and between-group sum of squares and cross-products matrices, and in canonical correlation analysis, wherein interest lies in comparing the covariance matrix of one set of variables with the covariance matrix of its linear predictor based on another set of variables. These methods involve either multiple populations or two different sets of variables. Less attention has been given to the comparison of different estimates of scatter for a single set of variables from a single population. Some work in this direction, though, can be found in Art *et al.* (1982), Caussinus and Ruiz-Gazen (1990, 1993, 1995), Caussinus *et al.* (2003) and Ruiz-Gazen (1993), which will be discussed in later sections.

Typically, the difference between two positive definite symmetric matrices can be summarized by considering the eigenvalues and eigenvectors of one matrix with respect to the other. More specifically, suppose that $V_1 \in \mathcal{P}_p$ and $V_2 \in \mathcal{P}_p$. An eigenvalue, say ρ_j, and a corresponding eigenvector, say h_j, of V_2 relative to V_1 correspond to a non-trivial solution to the matrix equations

$$V_2 h_j = \rho_j V_1 h_j. \tag{11}$$

Equivalently, ρ_j and h_j are an eigenvalue and corresponding eigenvector respectively of $V_1^{-1} V_2$. Since most readers are probably more familiar with the eigenvalue–eigenvector theory of symmetric matrices, we note that ρ_j also represents an eigenvalue of the symmetric matrix $M = V_1^{-1/2} V_2 V_1^{-1/2} \in \mathcal{P}$, where $V_1^{1/2} \in \mathcal{P}_p$ denotes the unique positive definite symmetric square root of V_1. Hence, we can choose p ordered eigenvalues, $\rho_1 \geqslant \rho_2 \geqslant \ldots \geqslant \rho_p > 0$, and an orthonormal set of eigenvectors q_j, $j = 1, \ldots, p$, such that $Mq_j = \rho_j q_j$. The relationship between h_j and the eigenvectors of M is given by $q_j \propto V_1^{1/2} h_j$, and so $h_i' V_1 h_j = 0$ for $i \neq j$. This yields the following simultaneous diagonalization of V_1 and V_2:

$$\begin{aligned} H' V_1 H &= D_1, \\ H' V_2 H &= D_2 \end{aligned} \tag{12}$$

where $H = (h_1 \ldots h_p)$, D_1 and D_2 are diagonal matrices with positive entries and $D_1^{-1} D_2 = \Delta = \mathrm{diag}(\rho_1, \ldots, \rho_p)$. Without loss of generality, we can take $D_1 = I$ by normalizing h_j so that $h_j' V_1 h_j = 1$. Alternatively, we can take $D_2 = I$. Such a normalization is not necessary for our purposes and we simply prefer the general form (12) since it reflects the exchangeability between the roles of V_1 and V_2. Note that the matrix $V_1^{-1} V_2$ has the spectral value decomposition

$$V_1^{-1} V_2 = H \Delta H^{-1}. \tag{13}$$

Various useful interpretations of the eigenvalues and eigenvectors in equation (11) can be given whenever V_1 and V_2 are two different scatter matrices for the same population or sample. We first note that the eigenvalues $\rho_1, \ldots, \rho_p$ are the maximal invariants under affine transformation for comparing V_1 and V_2, i.e., if we define a function $G(V_1, V_2)$ such that $G(V_1, V_2) = G(AV_1A', AV_2A')$ for any non-singular A, then $G(V_1, V_2) = G(D_1, D_2) = G(I, \Delta)$, with D_1, D_2 and Δ being defined as above. Furthermore Δ is invariant under such transformations. Since scatter matrices tend to be well defined only up to a scalar multiple, it is more natural to be

interested in the difference between V_1 and V_2 up to proportionality. In this case, if we consider a function $G(V_1, V_2)$ such that $G(V_1, V_2) = G(\lambda_1 A V_1 A', \lambda_2 A V_2 A')$ for any non-singular A and any $\lambda_1 > 0$ and $\lambda_2 > 0$, then $G(V_1, V_2) = G\{I, \Delta/\det(\Delta)^{1/p}\}$, i.e. maximal invariants in this case are

$$(\rho_1, \ldots, \rho_p) \Big/ \left(\prod_{i=1}^{p} \rho_i\right)^{1/p}$$

or, in other words, we are interested in $(\rho_1, \ldots, \rho_p)$ up to a common scalar multiplier.

A more useful interpretation of the eigenvalues arises from the following optimality property, which follows readily from standard eigenvalue–eigenvector theory. For $h \in \Re^p$, let

$$\kappa(h) = h' V_2 h / h' V_1 h. \tag{14}$$

For $V_1 = V_1(F_Y)$ and $V_2 = V_2(F_Y)$, $\kappa(h)$ represents the square of the ratio of two different measures of scale for the variable $h'Y$. Recall that the classical measure of kurtosis corresponds to the fourth power of the ratio of two scale measures, namely the fourth root of the fourth central moment and the standard deviation. Thus, the value of $\kappa(h)^2$ can be viewed as a generalized measure of 'relative' kurtosis. The term relative is used here since the scatter matrices V_1 and V_2 are not necessarily normalized. If both V_1 and V_2 are normalized so that they are both consistent for the variance–covariance matrix under a multivariate normal model, then a deviation of $\kappa(h)$ from 1 would indicate non-normality. In general, though, the ratio $\kappa(h_1)^2/\kappa(h_2)^2$ does not depend on any particular normalization.

The maximal possible value of $\kappa(h)$ over $h \in \Re^p$ is ρ_1 with the maximum being achieved in the direction of h_1. Likewise, the minimal possible value of $\kappa(h)$ is ρ_p with the minimum being achieved in the direction of h_p. More generally, we have

$$\sup\{\kappa(h) | h \in \Re^p, h' V_1 h_j = 0, j = 1, \ldots, m-1\} = \rho_m, \tag{15}$$

with the supremum being obtained at h_m, and

$$\inf\{\kappa(h) | h \in \Re^p, h' V_1 h_j = 0, j = m+1, \ldots, p\} = \rho_m, \tag{16}$$

with the infimum being obtained at h_m. These successive optimality results suggest that plotting the data or distribution by using the co-ordinates $Z = H'Y$ may reveal interesting structures. We explore this idea in later sections.

Remark 2. An alternative motivation for the transformation $Z = H'Y$ is as follows. Suppose that Y is first 'standardized' by using a scatter functional $V_1(F)$ satisfying condition (3), i.e. $X = V_1(F_Y)^{-1/2} Y$. If Y is elliptically symmetric about μ_Y, then X is spherically symmetric about the centre $\mu_X = V_1(F_Y)^{-1/2} \mu_Y$. If a second scatter functional is then applied to X, say $V_2(F)$ satisfying condition (3), then $V_2(F_X) \propto I$, and hence no projection of X is any more interesting than any other projection of X. However, if Y is not elliptically symmetric, then $V_2(F_X)$ is not necessarily proportional to I. This suggests that a principal components analysis of X based on $V_2(F_X)$ may reveal some interesting projections. By taking the spectral value decomposition $V_2(F_X) = QDQ'$, where Q is an orthogonal matrix, and then constructing the principal component variables $Q'X$, we obtain

$$Q'X = H'Y = Z, \qquad \text{with } D = \Delta, \tag{17}$$

whenever H is normalized so that $H' V_1(F_Y) H = I$.

4. Invariant co-ordinate systems

In this and the following section we study the properties of the transformation $Z = H'Y$ in more detail, and in Section 6 we give some examples illustrating the utility of the transformation when used in diagnostic plots. For simplicity, unless otherwise stated, we hereafter state any theoretical properties by using the functional or population version of scatter matrices. The sample version then follows as a special case based on the empirical distributions. Examples are, of course, given for the sample version. The following condition is assumed throughout and the following notation is used hereafter.

Condition 1. For $Y \in \Re^p$ having distribution F_Y, let $V_1(F)$ and $V_2(F)$ be two scatter functionals satisfying condition (3). Further, suppose that both $V_1(F)$ and $V_2(F)$ are uniquely defined at F_Y.

Definition 1. Let $H(F) = (h_1(F) \ldots h_p(F))$ be a matrix of eigenvectors defined as in equations (11) and (12), with $\rho_1(F) \geqslant \ldots \geqslant \rho_p(F)$ being the corresponding eigenvalues, whenever V_1 and V_2 are taken to be $V_1(F)$ and $V_2(F)$ respectively.

It is well known that principal component variables are invariant under translations and orthogonal transformations of the original variables, but not invariant under other general affine transformations. An important property of the transformation that is proposed here, i.e. $Z = H(F_Y)'Y$, is that the resulting variables are invariant under any affine transformation.

Theorem 1. In addition to condition 1, suppose that the roots $\rho_1(F_Y), \ldots, \rho_p(F_Y)$ are all distinct. Then for the affine transformation $Y^* = AY + b$, with A being non-singular,

$$\rho_j(F_{Y^*}) = \gamma\, \rho_j(F_Y) \qquad \text{for } j = 1, \ldots, p \tag{18}$$

for some $\gamma > 0$. Moreover, the components of $Z = H(F_Y)'Y$ and $Z^* = H(F_{Y^*})'Y^*$ differ at most by co-ordinatewise location and scale, i.e., for some constants $\alpha_1, \ldots, \alpha_p$ and $\beta_1, \ldots, \beta_p$, with $\alpha_j \neq 0$ for $j = 1, \ldots, p$,

$$Z_j^* = \alpha_j Z_j + \beta_j \qquad \text{for } j = 1, \ldots, p. \tag{19}$$

Owing to property (19) we refer to the transformed variables $Z = H(F_Y)'Y$ as an *invariant co-ordinate system*, and the method for obtaining them as ICS. If a univariate standardization is applied to the transformed variables, then the standardized versions of Z_j and Z_j^* differ only by a factor of ± 1.

A generalization of the previous theorem, which allows for possible multiple roots, can be stated as follows.

Theorem 2. Let Y, Y^*, Z and Z^* be defined as in theorem 1. In addition to condition 1, suppose that the roots $\rho_1(F_Y), \ldots, \rho_p(F_Y)$ consist of m distinct values, say $\rho_{(1)} > \ldots > \rho_{(m)}$, with $\rho_{(k)}$ having multiplicity p_k for $k = 1, \ldots, m$, and hence $p_1 + \ldots + p_m = p$. Then, expression (18) still holds. Furthermore, suppose that we partition $Z' = (Z'_{(1)}, \ldots, Z'_{(m)})$, where $Z_{(k)} \in \Re^{p_k}$. Then, for some non-singular matrix C_k of order p_k and some p_k-dimensional vector β_k,

$$Z^*_{(k)} = C_k Z_{(k)} + \beta_k \qquad \text{for } k = 1, \ldots, m, \tag{20}$$

i.e. the space that is spanned by the components of $Z^*_{(k)}$ is the same as the space that is spanned by the components of $Z_{(k)}$.

As with any eigenvalue–eigenvector problem, eigenvectors are not well defined. For a distinct root, the eigenvector is well defined up to a scalar multiple. For a multiple root, say with multiplicity p_0, the corresponding p_0 eigenvectors can be chosen to be any linearly independent

vectors spanning the corresponding p_0-dimensional eigenspace. Consequently $Z_{(k)}$ in theorem 2 is not well defined. One could construct some arbitrary rule for defining $Z_{(k)}$ uniquely. However, this is not necessary here since, no matter which rule we may use to define $Z_{(k)}$ uniquely, the results of theorem 2 hold.

5. Invariant co-ordinate selection under non-elliptical models

When Y has an elliptically symmetric distribution, all the roots $\rho_1(F_Y), \ldots, \rho_p(F_Y)$ are equal, and so the ICS transformation $Z = H(F_Y)'Y$ is arbitrary. The aim of ICS though is to detect departures of Y from an elliptically symmetric distribution. In this section, the behaviour of the ICS transformation is demonstrated theoretically for two classes of non-elliptically symmetric models, namely for mixtures of elliptical distributions and for independent components models.

5.1. *Mixture of elliptical distributions*

In practice, data often appear to arise from mixture distributions, with the mixing being the result of some unmeasured grouping variable. Uncovering the different groups is typically viewed as a problem in cluster analysis. One clustering method, which was proposed by Art *et al.* (1982), is based on first reducing the dimension of the clustering problem by attempting to identify Fisher's linear discriminant subspace. To do this, they gave an iterative algorithm for approximating the within-group sum of squares and cross-products matrix, say W_n, and then considered the eigenvectors of $W_n^{-1}(T_n - W_n)$, where T_n is the total sum-of-squares and cross-products matrix. The approach that was proposed by Art *et al.* (1982) was motivated primarily by heuristic arguments and was supported by a Monte Carlo study.

Subsequently, Ruiz-Gazen (1993) and Caussinus and Ruiz-Gazen (1993, 1995) showed for a location mixture of multivariate normal distributions with equal variance–covariance matrices that Fisher's linear discriminant subspace can be consistently estimated even when the group identification is not known, provided that the dimension, say q, of the subspace is known. Their results are based on the eigenvectors that are associated with the q largest eigenvalues of $S_{1,n}^{-1}S_n$, where S_n is the sample variance–covariance matrix and $S_{1,n}$ is either the one-step W-estimate (7) or its symmetrized version. They also required that the $S_{1,n}$ differs from S_n by only a small perturbation, since their proof involves expanding the functional version of $S_{1,n}$ about the functional version of S_n. In this subsection, it is shown that these results can be extended essentially to any pair of scatter matrices, and also that the results hold under mixtures of elliptical distributions with proportional scatter parameters.

For simplicity, we first consider properties of the ICS transformation for a mixture of two multivariate normal distributions with proportional covariance matrices. Considering proportional covariance matrices allows for the inclusion of a point mass contamination as one of the mixture components, since a point mass contamination is obtained by letting the proportionality constant go to 0.

Theorem 3. In addition to condition 1, suppose that

$$Y \sim_{\mathrm{d}} (1-\alpha)N_p(\mu_1, \Gamma) + \alpha\, N_p(\mu_2, \lambda\Gamma),$$

where $0 < \alpha < 1$, $\mu_1 \neq \mu_2$, $\lambda > 0$ and $\Gamma \in \mathcal{P}_p$. Then either

(a) $\rho_1(F_Y) > \rho_2(F_Y) = \ldots = \rho_p(F_Y)$,
(b) $\rho_1(F_Y) = \ldots = \rho_{p-1}(F_Y) > \rho_p(F_Y)$, or
(c) $\rho_1(F_Y) = \ldots = \rho_p(F_Y)$.

For $p > 2$, if case (a) holds, then $h_1(F_Y) \propto \Gamma^{-1}(\mu_1 - \mu_2)$ and, if case (b) holds, then $h_p(F_Y) \propto \Gamma^{-1}(\mu_1 - \mu_2)$. For $p = 2$, if $\rho_1(F_Y) > \rho_2(F_Y)$, then either $h_1(F_Y)$ or $h_2(F_Y)$ is proportional to $\Gamma^{-1}(\mu_1 - \mu_2)$.

Thus, depending on whether case (a) or case (b) holds, h_1 or h_p respectively corresponds to Fisher's linear discriminant function (see for example Mardia *et al.* (1980)), even though the group identity is unknown. An intuitive explanation about why we might expect this to hold is that any estimate of scatter contains information on the between-group variability, i.e. the difference between μ_1 and μ_2, and the within-group variability or shape, i.e. Γ. Thus, one might expect that we could separate these two sources of variability by using two different estimates of scatter. This intuition though is not used in our proof of theorem 3; nor is our proof based on generalizing the perturbation arguments that were used by Ruiz-Gazen (1993) and Caussinus and Ruiz-Gazen (1995) in deriving their aforementioned results. Rather, the proof of theorem 3 that is given in Appendix A relies solely on invariance arguments.

Whether case (a) or case (b) holds in theorem 3 depends on the choice of $V_1(F)$ and $V_2(F)$ and on the nature of the mixture. Obviously, if case (a) holds and then the roles of $V_1(F)$ and $V_2(F)$ are reversed, then case (b) would hold. Case (c) holds only in very specific situations. In particular, case (c) holds if $\mu_1 = \mu_2$, in which case Y has an elliptically symmetric distribution. When $\mu_1 \neq \mu_2$, i.e. when the mixture is not elliptical itself, it is still possible for case (c) to hold. This though is dependent not only on the specific choice of $V_1(F)$ and $V_2(F)$ but also on the particular value of the parameters α, μ_1, μ_2, Γ and λ.

For example, suppose that $V_1(F) = \Sigma(F)$, the population covariance matrix, and $V_2(F) = \mathcal{K}(F)$ where

$$\mathcal{K}(F) = E[(Y - \mu_Y)' \Sigma(F)^{-1}(Y - \mu_Y) \times (Y - \mu_Y)(Y - \mu_Y)'] . \tag{21}$$

Beside being analytically tractable, the scatter functional $\mathcal{K}(F)$ is one which arises in a classical algorithm for independent components analysis and is discussed in more detail in later sections. For the special case $\lambda = 1$ and when $\mu_1 \neq \mu_2$, if we let $\eta = \alpha(1 - \alpha)$, then it can be shown that case (a) holds for $\eta > 1/6$, case (b) holds for $\eta < 1/6$ and case (c) holds for $\eta = 1/6$. Also, for any of these three cases, we have $\rho_1(F_Y) - \rho_p(F_Y) = \eta|1 - 6\eta|\theta^2/(1 + \eta\theta)^2$, where $\theta = (\mu_1 - \mu_2)'\Gamma^{-1}(\mu_1 - \mu_2)$.

Other examples have been studied in Caussinus and Ruiz-Gazen (1993, 1995). In their work, $V_2(F) = \Sigma(F)$ and $V_1(F)$ corresponds to the functional version of the symmetrized version of the one-step W-estimate (7). Paraphrasing, they showed for the case $\lambda = 1$ and for the class of weight functions $u_2(s) = u(\beta s)$ that case (a) holds for sufficiently small β provided that $\eta < 1/6$. They did not note, though, that case (a) or (b) can hold for other values of β and η. The reason that the condition $\eta < 1/6$ arises in their work, as well as in the discussion in the previous paragraph, is because their proof involves expanding $u(\beta s)$ about $u(s)$, with the matrix $\mathcal{K}(F)$ then appearing in the linear term of the corresponding expansion of the one-step W-estimate about $\Sigma(F)$.

Theorem 3 readily generalizes to a mixture of two elliptical distributions with equal shape matrices, but with possibly different location vectors and different spread functions, i.e., if Y has density

$$f_Y(y) = (1 - \alpha)\, f(y; \mu_1, \Gamma, g_1) + \alpha\, f(y; \mu_2, \Gamma, g_2),$$

where $0 < \alpha < 1$, $\mu_1 \neq \mu_2$ and $f(y; \mu, \Gamma, g)$ is defined by expression (6), then the results of theorem 3 hold. Note that this mixture distribution includes the case where both mixture components are from the same elliptical family but with proportional shape matrices. This special case corresponds to setting $g_2(s) = g_1(s/\lambda)$, and hence $f(y; \mu_2, \Gamma, g_2) = f(y; \mu_2, \lambda\Gamma, g_1)$.

An extension of these results to a mixture of k elliptically symmetric distributions with possibly different centres and different spread functions, but with equal shape matrices, is given in the following theorem. Stated more heuristically, this theorem implies that Fisher's *linear discriminant subspace* (see for example Mardia *et al.* (1980)) corresponds to the span of some subset of the invariant co-ordinates, even though the group identifications are not known.

Theorem 4. In addition to condition 1, suppose that Y has density

$$f_Y(y) = \det(\Gamma)^{-1/2} \sum_{j=1}^{k} \alpha_j g_j\{(y-\mu_j)'\Gamma^{-1}(y-\mu_j)\},$$

where $\alpha_j > 0$ for $j = 1, \ldots, k$, $\alpha_1 + \ldots + \alpha_k = 1$, $\Gamma \in \mathcal{P}_p$ and $g_1, \ldots, g_k$ are non-negative functions. Also, suppose that the centres $\mu_1, \ldots, \mu_k$ span some q-dimensional hyperplane, with $0 < q < p$. Then, using the notation of theorem 2 for multiple roots, there is at least one root $\rho_{(j)}$, $j = 1, \ldots, m$, with multiplicity greater than or equal to $p-q$. Furthermore, if no root has multiplicity greater than $p-q$, then there is a root with multiplicity $p-q$, say $\rho_{(t)}$, such that

$$\text{span}\{\Gamma^{-1}(\mu_j - \mu_k) | j = 1, \ldots, k-1\} = \text{span}\{H_q(F_Y)\}, \tag{22}$$

where $H_q(F_Y) = (h_1(F_Y), \ldots, h_{p_1+\ldots+p_{t-1}}(F_Y), h_{p_1+\ldots+p_{t+1}}(F_Y), \ldots, h_p(F_Y))$.

The condition in theorem 4 that only one root has multiplicity $p-q$ and no other root has a greater multiplicity reduces to case (a)–(b) in theorem 3 when $k = 2$. Analogously to the discussion given after theorem 3, this condition generally holds except for special cases. For a given choice of $V_1(F_Y)$ and $V_2(F_Y)$, these special cases depend on the particular values of the parameters.

5.2. *Independent components analysis models*

Independent components analysis (ICA) is a highly popular method within many applied areas which routinely encounter multivariate data. For a good overview, see Hyvärinen *et al.* (1981). The most common ICA model presumes that Y arises as a convolution of p independent components or variables, i.e. $Y = BX$, where B is non-singular, and the components of X, say $X_1, \ldots, X_p$, are independent. The main objective of ICA is to recover the mixing matrix B so that we can 'unmix' Y to obtain independent components $X^* = B^{-1}Y$. Under this ICA model, there is some indeterminacy in the mixing matrix B, since the model can also be expressed as $Y = B_0 X_0$, where $B_0 = BQ\Lambda$ and $X_0 = \Lambda^{-1}Q'X$, Q being a permutation matrix and Λ a diagonal matrix with non-zero entries. The components of X_0 are then also independent. Under the condition that at most one of the independent components $X_1, \ldots, X_p$ has a normal distribution, it is well known that this is the only indeterminacy for B, and consequently the independent components $X = B^{-1}Y$ are well defined up to permutations and componentwise scaling factors.

The relationship between ICS and ICA for symmetric distributions is given in the next theorem.

Theorem 5. In addition to condition 1, suppose that $Y = BX + \mu$, where B is non-singular, and the components of X, say $X_1, \ldots, X_p$, are mutually independent. Further, suppose that X is symmetric about 0, i.e. $X \sim_d -X$, and the roots $\rho_1(F_Y), \ldots, \rho_p(F_Y)$ are all distinct. Then, the transformed variable $Z = H(F_Y)'Y$ consists of independent components or, more specifically, Z and X differ by at most a permutation and/or componentwise location and scale.

From the proof of theorem 5, it can be noted that the condition that X be symmetrically distributed about 0 can be relaxed to require that only $p-1$ of the components of X be symmetrically distributed about 0. It is also worth noting that the condition that all the roots be

distinct is more restrictive than the condition that at most one of the components of X is normal. This follows since it is straightforward to show in general that, if the distributions of two components of X differ from each other by only a location shift and/or scale change, then there is at least one root having multiplicity greater than 1.

If X is not symmetric about 0, then we can symmetrize Y before applying theorem 5, i.e. suppose that $Y = BX + \mu$ with X having independent components, and let Y_1 and Y_2 be independent copies of Y. Then $Y^s = Y_1 - Y_2 = BX^s$, where $X^s = X_1 - X_2$ is symmetric about zero and has independent components. Thus, theorem 5 can be applied to Y^s. Moreover, since the convolution matrix B is the same for both Y and Y^s, it follows that the transformed variable $Z = H(F_Y^s)'Y$ and X differ by at most a permutation and/or componentwise location and scale, where F_Y^s refers to the symmetrized distribution of F_Y, i.e. the distribution of Y^s.

An alternative to symmetrizing Y is to choose both $V_1(F)$ and $V_2(F)$ so that they satisfy the following *independence property*.

Definition 1. An affine equivariant scatter functional $V(F)$ is said to have the 'independence property' if $V(F_X)$ is a diagonal matrix whenever the components of X are mutually independent, provided that $V(F_X)$ exists.

Assuming this property, Oja *et al.* (2006) proposed to use principal components on standardized variables as defined in remark 2 to obtain a solution to the ICA problem. Their solution can be restated as follows.

Theorem 6. In addition to condition 1, suppose that $Y = BX + \mu$, where B is non-singular, and the components of X, say $X_1, \ldots, X_p$, are mutually independent. Further, suppose that both scatter functionals $V_1(F)$ and $V_2(F)$ satisfy the independence property that is given in definition 1, and the roots $\rho_1(F_Y), \ldots, \rho_p(F_Y)$ are all distinct. Then, the transformed variable $Z = H(F_Y)'Y$ consists of independent components or, more specifically, Z and X differ by at most a permutation and/or componentwise location and scale.

The covariance matrix $\Sigma(F)$ is of course well known to satisfy definition 1. It is also straightforward to show that the scatter functional $\mathcal{K}(F)$ that is defined in equation (21) does as well. Theorem 6 represents a generalization of an early ICA algorithm that was proposed by Cardoso (1989) based on the spectral value decomposition of a *kurtosis matrix*. Cardoso's algorithm, which he called the fourth-order blind identification algorithm, can be shown to be equivalent to choosing $V_1(F) = \Sigma(F)$ and $V_2(F) = \mathcal{K}(F)$ in theorem 6.

It is worth noting that the independence property that is given by definition 1 is weaker than the property

$$X_i \text{ and } X_j \text{ are independent} \Rightarrow V(F_X)_{i,j} = 0. \tag{23}$$

The covariance matrix satisfies property (23), whereas $\mathcal{K}(F)$ does not.

An often overlooked observation is that property (23) does not hold for robust scatter functionals in general, i.e. independence does not necessarily imply a zero pseudocorrelation. It is an open problem what scatter functionals other than the covariance matrix, if any, satisfy property (23). Furthermore, robust scatter functionals tend not to satisfy in general even the weaker definition 1. At symmetric distributions, though, the independence property can be shown to hold for general scatter matrices in the following sense.

Theorem 7. Let $V(F)$ be a scatter functional satisfying condition (3). Suppose that the distribution of X is symmetric about some centre $\mu \in \Re^p$, with the components of X being mutually independent. If $V(F_X)$ exists, then it is a diagonal matrix.

Consequently, given a scatter functional $V(F)$, we can construct a new scatter functional satisfying definition 1 by defining $V^s(F) = V(F^s)$, where F^s represents the symmetrized distribution of F. Using symmetrization to obtain scatter functionals which satisfy the independence property has been studied recently by Taskinen *et al.* (2007).

Finally, we note that the results of this section can be generalized in two directions. First, we consider the case of multiple roots, and next we consider the case where only blocks of the components of X are independent.

Theorem 8. In addition to condition 1, suppose that $Y = BX + \mu$, where B is non-singular, and the components of X, say $X_1, \ldots, X_p$, are mutually independent. Further, suppose that either

(a) X is symmetric about 0, i.e. $X \sim_d -X$, or
(b) both $V_1(F)$ and $V_2(F)$ satisfy definition 1.

Then, using the notation of theorem 2 for multiple roots, for the transformed variable $Z = H(F_Y)'Y$ the random vectors $Z_{(1)}, \ldots, Z_{(m)}$ are mutually independent.

Theorem 9. In addition to condition 1, suppose that $Y = BX + \mu$, where B is non-singular, and $X' = (X'_{(1)}, \ldots, X'_{(m)})$ has mutually independent components $X_{(1)} \in \Re^{p_1}, \ldots, X_{(m)} \in \Re^{p_m}$, with $p_1 + \ldots + p_m = p$. Further, suppose that X is symmetric about 0, and the roots $\rho_1(F_Y), \ldots, \rho_p(F_Y)$ are all distinct. Then, there is a partition $\{J_1, \ldots, J_m\}$ of $\{1, \ldots, p\}$ with the cardinality of J_k being p_k for $k = 1, \ldots, m$ such that for the transformed variable $Z = H(F_Y)'Y$ the random vectors

$$Z_{(1)} = \{Z_j, j \in J_1\}, \ldots, Z_{(m)} = \{Z_j, j \in J_m\}$$

are mutually independent. More specifically, $Z_{(j)}$ and $X_{(j)}$ are affine transformations of each other.

From the proof of theorem 9 in Appendix A, it can be noted that the theorem still holds if one of the $X_{(j)}$s is not symmetric. If the distribution of X is not symmetric, theorems 8 and 9 can be applied to Y^s, the symmetrized version of Y. To generalize theorem 6 to the case where blocks of the components of X are independent, a modification of the independence property is needed. Such generalizations of definition 1, theorem 6 and theorem 7 are fairly straightforward and so are not treated formally here.

Remark 3. The general case of multiple roots for the setting that is given in theorem 9 is more problematic. The problem stems from the possibility that a multiple root may not be associated with a particular $X_{(j)}$ but rather with two or more different $X_{(j)}$s. For example, consider the case $X' = (X'_{(1)}, X_{(2)})$, with $X_{(1)} \in \Re^2$ and $X_{(2)} \in \Re$. For this case, $V_1(F_X)^{-1} V_2(F_X)$ is block diagonal with diagonal blocks of order 2 and 1. The three eigenvalues $\rho_1(F_Y)$, $\rho_2(F_Y)$ and $\rho_3(F_Y)$ correspond to the two eigenvalues of the diagonal block of order 2 and to the last diagonal element, but not necessarily respectively. So, if $\rho_1(F_Y) = \rho_2(F_Y) > \rho_3(F_Y)$, this does not imply that the last diagonal element corresponds to $\rho_3(F_Y)$, and hence $Z_{(1)} \in \Re^2$ and $Z_{(2)} \in \Re$, as defined in theorem 2, are not necessarily independent.

6. Discussion and examples

Although the theoretical results of this paper essentially apply to any pair of scatter matrices, in practice the choice of scatter matrices can affect the resulting ICS method. From our experience, for some data sets, the choice of the scatter matrices does not seem to have a big influence on the

diagnostic plots of the ICS variables, particularly when the data are consistent with one of the mixture models or one of the independent component models that were considered in Section 5. For some other data sets, however, the resulting diagnostic plots can be quite sensitive to the choice of the scatter matrices. In general, different pairs of scatter matrices may reveal different types of structure in the data, since departures from an elliptical distribution can come in many forms. Consequently, it is doubtful whether any specific pair of scatter matrices is best for all situations. Rather than choosing two scatter matrices beforehand, especially when one is in a purely exploratory situation having no idea of what to expect, it would be reasonable to consider a number of different pairs of scatter matrices and to consider the resulting ICS transformations as complementary.

A general sense of how the choice of the pair of scatter matrices may impact the resulting ICS method can be obtained by a basic understanding of the properties of the scatter matrices being used. For the purpose of this discussion, we divide the scatter matrices into three broad classes. Class I scatter statistics will refer to those which are not robust in the sense that their breakdown point is essentially zero. This class includes the sample covariance matrix, as well as the one-step W-estimates defined by expression (7) and their symmetrized version. Other scatter statistics which lie within this class are the multivariate sign and rank scatter matrices; see for example Visuri *et al.* (2000). Class II scatter statistics will refer to those which are moderately robust in the sense that they have bounded influence functions as well as positive breakdown points, but with breakdown points being no greater than $1/(p+1)$. This class primarily includes the multivariate M-estimates, but it also includes among others the sample covariance matrices that are obtained after applying either convex hull peeling or ellipsoid hull peeling to the data; see Donoho and Gasko (1992). Class III scatter statistics will refer to the high breakdown point scatter matrices which are discussed in Section 2.2. The symmetrized version of a class II or III scatter matrix, as well as the one-step W-estimates of scatter (10) which uses an initial class II or III scatter matrix for downweighting, are viewed respectively as class II or III scatter matrices themselves.

If one or both scatter matrices are from class I, then the resulting ICS transformation may be heavily influenced by a few outliers at the expense of finding other structures in the data. In addition, even if there are no spurious outliers and a mixture model or an independent components model of the form that was discussed in Section 5 holds, but with long-tailed distributions, then the resulting sample ICS transformation may be an inefficient estimate of the corresponding population ICS transformation. Simulation studies that were reported in Nordhausen, Oja and Ollila (2008) have shown that for ICA an improved performance is obtained by choosing robust scatter matrices for the ICS transformation. Nevertheless, since they are simple to compute, the use of class I scatter matrices can be useful if the data set is known not to contain any spurious outliers or if the objective of the diagnostics is to find such outliers, as recommended in Caussinus and Ruiz-Gazen (1990).

If we use class II or III scatter matrices, then we can still find spurious outliers by plotting the corresponding robust Mahalanobis distances. The resulting ICS transformation, though, would not be heavily affected by the spurious outliers. Outliers affect class II scatter matrices more so than class III scatter matrices, although even a high proportion of spurious outliers may not necessarily affect the class II scatter matrices. For outliers to affect a class II scatter matrix heavily, they usually need to lie in a cluster; see for example Dümbgen and Tyler (2005). The results of Section 5.1 though suggest that such clustered outliers can be identified after making an ICS transformation, even if they cannot be identified by using a robust Mahalanobis distance based on a class II statistic.

Using two class III scatter matrices for an ICS transformation may not necessarily give good results, unless we are interested only in the structure of the 'inner' 50% of the data. For

example, suppose that the data arise from a 60–40 mixture of two multivariate normal distributions with widely separated means but equal covariance matrices. A class III scatter matrix is then primarily determined by the properties of the 60% component. Consequently, when using two class III scatter matrices for ICS the corresponding ICS roots will tend to be equal or nearly equal. In the case where all the roots are equal, theorem 3 does not apply. In the case where the roots are nearly equal, owing to sampling variation, the sample ICS transformation may not satisfactorily uncover Fisher's linear discriminant function.

A reasonable general choice for the pair of scatter matrices to use for an ICS transformation would be to use one class II and one class III scatter matrix. If we wish to avoid the computational complexity that is involved with a class III scatter matrix, then using two class II scatter matrices may be adequate. In particular, we could choose a class II scatter matrix whose breakdown point is close to $1/(p+1)$, such as the M-estimate corresponding to the maximum likelihood estimate for an elliptical Cauchy distribution (Dümbgen and Tyler, 2005), together with a corresponding one-step W-estimate for which $\psi(s) = s\,u_2(s) \to 0$ as $s \to \infty$. Such a one-step W-estimate of scatter has a redescending influence function. From our experience, the use of a class III scatter matrix for ICS does not seem to reveal any data structures that cannot be obtained otherwise.

The remarks and recommendations that are made here are highly conjectural. The question of what pairs of scatter matrices are best at detecting specific types of departure from an elliptical distribution remains a broad open problem. In particular, it would be of interest to discover for what types of data structures it would be advantageous to use at least one class III scatter matrix in the ICS method. Most likely, some advantages may arise when working with very high dimensional data sets, in which case the computational intensity that is needed to compute a class III scatter matrix is greatly amplified; see for example Rousseeuw and van Driessen (1999).

We demonstrate some of the concepts in the following examples. These examples illustrate for several data sets the use of the ICS transformation for constructing diagnostic plots. They also serve as illustrations of the theory that has been presented in the previous sections.

6.1. Example 1

Rousseeuw and van Driessen (1999) analysed a data set consisting of $n = 677$ metal plates on which $p = 9$ characteristics are measured. For this data set they computed the sample mean and covariance matrix as well as the minimum covariance determinant estimate of centre and scatter. Their paper helps to illustrate the advantage of using high breakdown point multivariate estimates, or class III statistics, for uncovering multiple outliers in a data set.

For our illustration, we choose two class II location and scatter statistics. The first estimate $(\hat{\mu}_1, \hat{V}_1)$ is taken to be the maximum likelihood estimate that is derived from an elliptical Cauchy distribution. This corresponds to an M-estimate (9) with $u_1(s) = u_2(s) = (p+1)/(s+1)$ and $u_3(s) = 1$. The M-estimating equations for this M-estimate are known to admit a unique solution in general, which can be found via a simple reweighting algorithm regardless of the initial value.

For our second estimate $(\hat{\mu}_2, \hat{V}_2)$, we take the sample mean vector and sample covariance matrix, using only the inner 50% of the data as measured by the Mahalanobis distances that are derived by using the Cauchy M-estimate, i.e. $d_i(\hat{\mu}_1, \hat{V}_1)$. This corresponds to a multivariate one-step W-estimate of scatter (10) with $u_1(s) = u_2(s) = I(s \leqslant 1)$ and $u_3(s) = 1$, and with initial estimates $\hat{\mu}_0 = \hat{\mu}_1$ and $\hat{V}_0 = \hat{\sigma}_1^2 \hat{V}_1$, where $\hat{\sigma} = \text{median}\{d_i(\hat{\mu}_1, \hat{V}_1), i = 1, \ldots, n\}$.

Figs 1(a) and 1(b) show the Mahalanobis distances plots for $d_i(\hat{\mu}_1, \hat{V}_1)$ and $d_i(\hat{\mu}_2, \hat{V}_2)$ respectively, with Fig. 1(c) being a scatter plot of these two sets of distances. These plots are somewhat similar to the plots that are based on the classical Mahalanobis distances and those based on

the minimum covariance determinant given in Rousseeuw and van Driessen (1999). As noted in Rousseeuw and van Driessen (1999), Figs 1(b) and 1(c) indicate that there are at least three distinct groups: the first 100 points, those with index 491–565 and the rest. The index itself is a factor that is not taken into account in obtaining the Mahalanobis distances. It represents an order of production and is clearly an important factor. The effect of the index is also apparent in some of the plots of the individual variables.

The comparative Mahalanobis distance plot that is given in Fig. 1(c) indicates that the data do not arise from an elliptically symmetric distribution. Otherwise, the scatter plot of the two distances would be approximately linear since the two location statistics would be estimating the same centre and the two scatter matrices would be estimating the same population shape matrix up to a proportionality constant, and consequently the resulting Mahalanobis distances would be approximately proportional to each other. The non-elliptical nature of the data can most likely be attributed to a mixture resulting from the index factor.

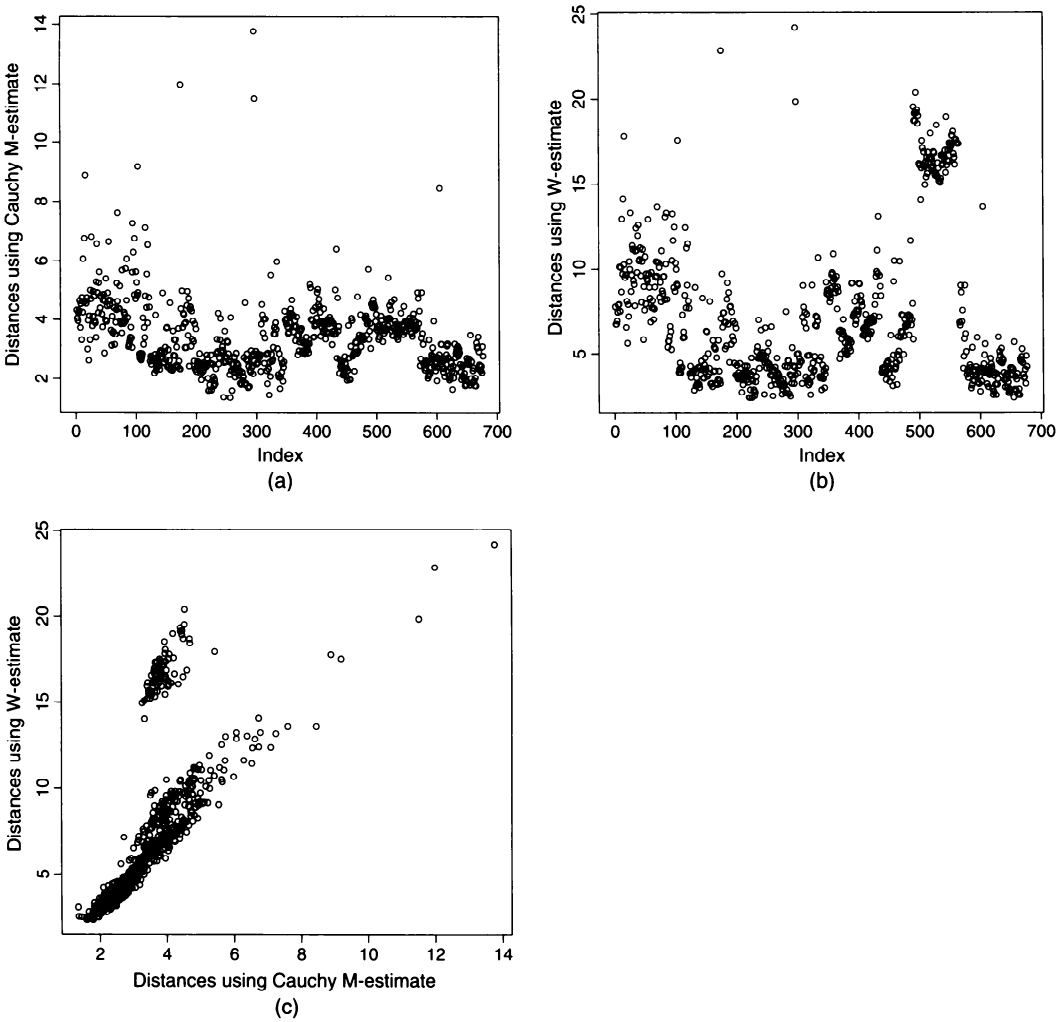

Fig. 1. Example 1: Mahalanobis distances based on (a) $\hat{V}_1$, (b) $\hat{V}_2$ and (c) $\hat{V}_1$ *versus* $\hat{V}_2$

The affine invariant plots that are given in Fig. 1 do not reveal whether the three groups that are observed in the plots correspond to three clusters, since the Mahalanobis distances give no indication of the relative distance of the points from each other. A more complete affine invariant view of the data can be obtained from a pairs plot of the ICS transformation of the data based on the scatter matrices $\hat{V}_1$ and $\hat{V}_2$ described above. For this analysis, the resulting ICS roots are $(\hat{\rho}_1, \ldots, \hat{\rho}_9) = (19.94, 5.27, 3.68, 3.41, 2.89, 2.61, 2.12, 1.69, 1.62)$. Fig. 2(a) shows the scatter plot for the first two ICS components, with Figs 2(b) and 2(c) showing the first two ICS components separately. The three groups can also be seen in these plots. Moreover, we can ascertain how the groups differ. In particular, it can be noted that the group that is associated with index 491–565 essentially lies in a particular direction from the rest of the data, namely that determined by the first ICS component, whereas the first 100 points essentially lie in a different direction, determined by the second ICS component. Finally, if we plot the other ICS components, various isolated outliers also become visible.

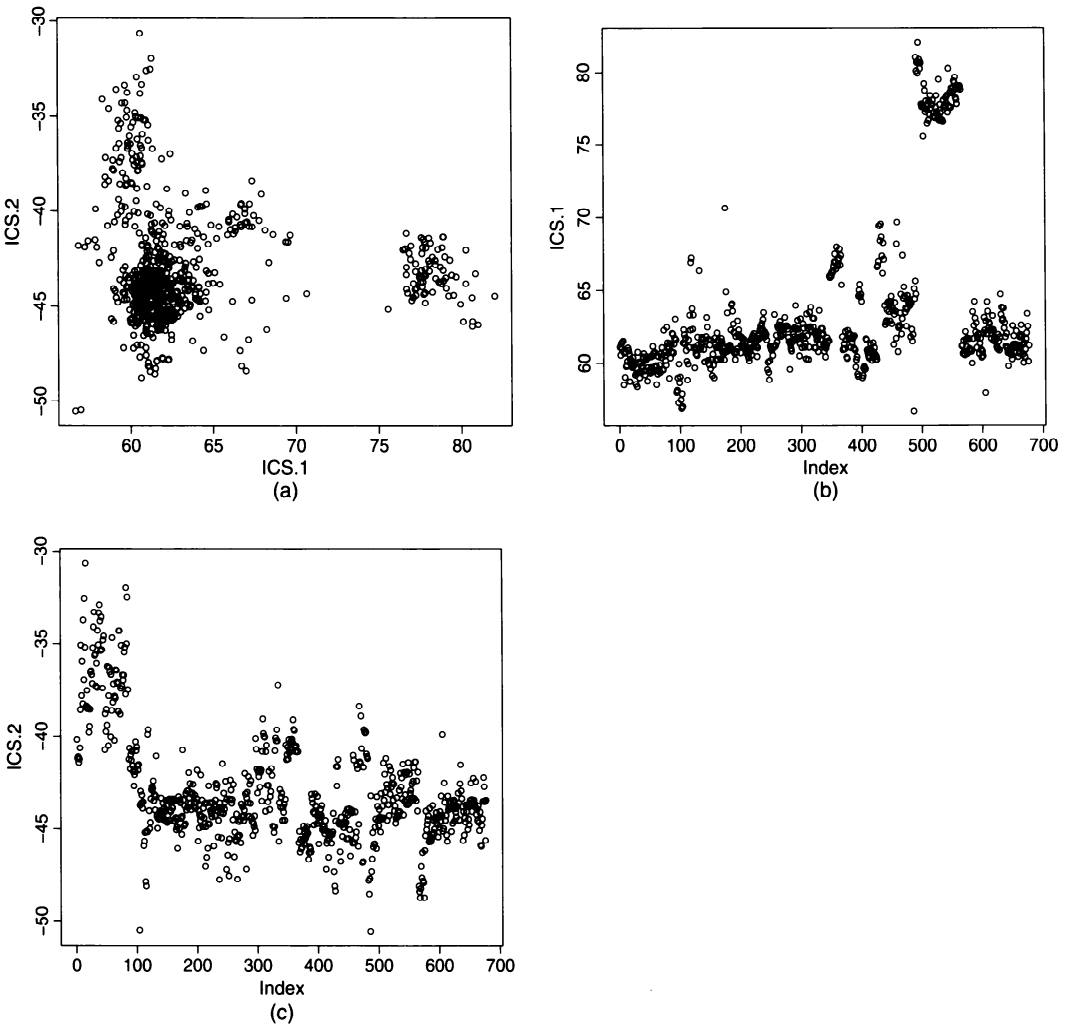

Fig. 2. Example 1: first and second ICS co-ordinates based on $\hat{V}_1$ and $\hat{V}_2$

6.2. Example 2

The pairs plot that is given in Fig. 3(a) arises from simulation of a random sample of size $n = 500$ from a $p = 4$ dimensional distribution. Arguably nothing seems particularly remarkable about this data set. The sample variance–covariance matrix of the data set in Fig. 3(a) is the identity matrix and so a principal components analysis does not indicate any particular direction of interest. If we apply an ICS transformation to these data, however, we can uncover an interesting hidden structure in the data as seen in the pairs plot given in Fig. 3(b). The corresponding ICS roots are $(\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4) = (1.49, 1.13, 0.81, 0.73)$. For this example, the original data y_i correspond to an affine transformation of a distribution that is generated by simulating a uniform distribution on the unit circle, to which independent normal noise with mean 0 and standard deviation 0.01 is added, concatenated with a standard normal distribution and a t-distribution on 5 degrees of freedom. Note that, no matter how the simulated data are affinely transformed, the resulting ICS co-ordinates are always given by Fig. 3(b).

The two scatter matrices that are used here for the ICS transformation are the sample covariance matrix S_n and the sample version of the scatter matrix $\mathcal{K}(F)$ given in equation (21), namely

$$\mathcal{K}_n = \frac{1}{n}\sum_{i=1}^{n}\{(y_i - \bar{y})' S_n^{-1}(y_i - \bar{y}) \times (y_i - \bar{y})(y_i - \bar{y})'\}. \tag{24}$$

$\mathcal{K}_n$ can be viewed as a one-step W-estimate of scatter (7) obtained by weighting each point by its classical Mahalanobis distance squared, i.e. choosing $u_2(s) = s$ and $u_3(s) = 1$, with $s_i = (y_i - \bar{y})' S_n^{-1}(y_i - \bar{y})$. As an estimate of scatter, $\mathcal{K}_n$ is obtained by actually upweighting outliers. Even though neither S_n nor $\mathcal{K}_n$ are robust estimates of scatter, they can uncover the structure in this particular data set since it contains no spurious outliers. Such a structure would be difficult to detect, with or without spurious outliers, if we were to consider only the Mahalanobis distances or a robust version of them. Similar results to those displayed in Fig. 3(b) arise when using almost any other pair of scatter matrices for the ICS transformation.

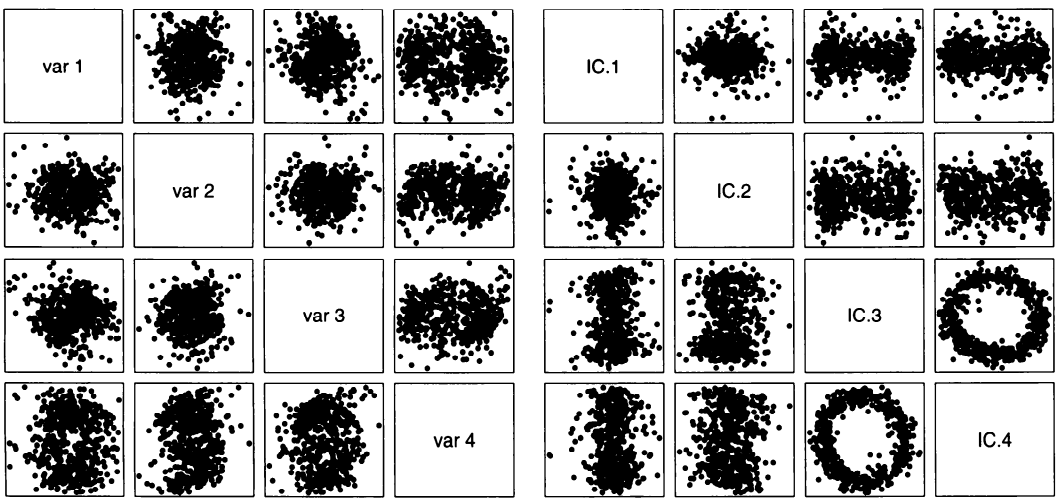

Fig. 3. Example 2: (a) a simulated four-dimensional data set and (b) the ICS co-ordinates by using S_n and $\mathcal{K}_n$

Note that none of the theorems that were given in Section 5.2 are directly applicable to this example, but rather this example is of the type that is discussed in remark 3. For the functional version of this example, we have $X' = (X'_{(1)}, X_{(2)}, X_{(3)})$, with the distribution of $X_{(1)} \in \Re^2$ being that of the uniform distribution on the unit circle plus bivariate spherical normal noise with variances 0.1, $X_{(2)} \in \Re$ having a standard normal distribution, $X_{(3)} \in \Re$ having a t-distribution on 5 degrees of freedom and with $X_{(1)}$, $X_{(2)}$ and $X_{(3)}$ being mutually independent. By using invariance arguments, it can be shown that, regardless of the choice of the two scatter matrices, at least two of the roots $\rho_1(F_Y), \ldots, \rho_4(F_Y)$ are equal, and hence there are at most three distinct roots. For the case of three distinct roots, the two ICS variables that are associated with the multiple root correspond to an affine transformation of $X_{(1)}$, and the ICS variables that are associated with the two distinct roots correspond to univariate linear transformations of $X_{(2)}$ and $X_{(3)}$. The case of three distinct roots tends to hold except for very special choices of $V_1(F)$ and $V_2(F)$. In particular, it can be shown to hold for the choice $V_1(F) = \Sigma(F)$ and $V_2(F) = \mathcal{K}(F)$, with the smallest root being the multiple root, the largest root being associated with $X_{(2)}$ and the second-largest root being associated with $X_{(3)}$. Hence, the results that are displayed in Fig. 3 are as expected.

6.3. *Other examples*

We briefly explain here the results of some other examples. The first is the classical Fisher iris data, which can be found in the statistical package R (R Development Core Team, 2005). This data set consists of $p = 4$ measurements, namely sepal length, sepal width, petal length and petal width, on $n = 150$ iris flowers. The 150 flowers belong to three different varieties of irises. Suppose that we ignore the group classification of the data and perform an ICS transformation of the $n = 150$ data points by using the sample covariance matrix and a Cauchy M-estimate. It turns out that the first ICS component is almost identical with the first linear discriminant function that we would obtain if we did a discriminant analysis using the varieties as the group variable, with a sample correlation between the two being 0.99, even though the former does not take the group classification into account. The results of the ICS method for this example are similar for almost any pair of scatter matrices that we may choose. This can be attributed to the data being consistent with the mixture models that are discussed in Section 5.1 together with the absence of any obvious outliers.

The next example uses the modified wood gravity data set that was given in Rousseeuw and Leroy (1987). This data set is frequently used as an example illustrating outlier detection methods. It consists of $n = 20$ observations in $p = 6$ dimensions, of which four of the observations are artificial outliers that had been put into the data set by the original authors. Rousseeuw and Leroy (1987) demonstrated how classical outlier detection methods fail to uncover these outliers, whereas they are readily uncovered by using Mahalanobis distances based on high breakdown point location and scatter statistics. For this data set, we compute a Cauchy M-estimate and a t_2 M-estimate, which have breakdown points of $1/7 = 0.143$ and $1/8 = 0.125$ respectively. Unlike example 1, neither corresponding Mahalanobis distance plot, which are given by Figs 4(a) and 4(b), reveals any outliers, and the two plots are fairly similar. Since the proportion of contamination is $4/20 = 0.20$, we would not expect that the Mahalanobis distances based on either of the two M-estimates would reveal the outliers if the outliers formed a cluster; see Tyler (2002). However, if the outliers do form a cluster then the results of Section 5.1 suggest an ICS transformation based on these two scatter estimates may separate the main cluster of data from the cluster of four outliers. Such is the case here, with all four outliers clearly appearing in the first ICS co-ordinate; see Fig. 4(c). The results here are again not heavily dependent on the scatter matrices being used in the ICS transformation.

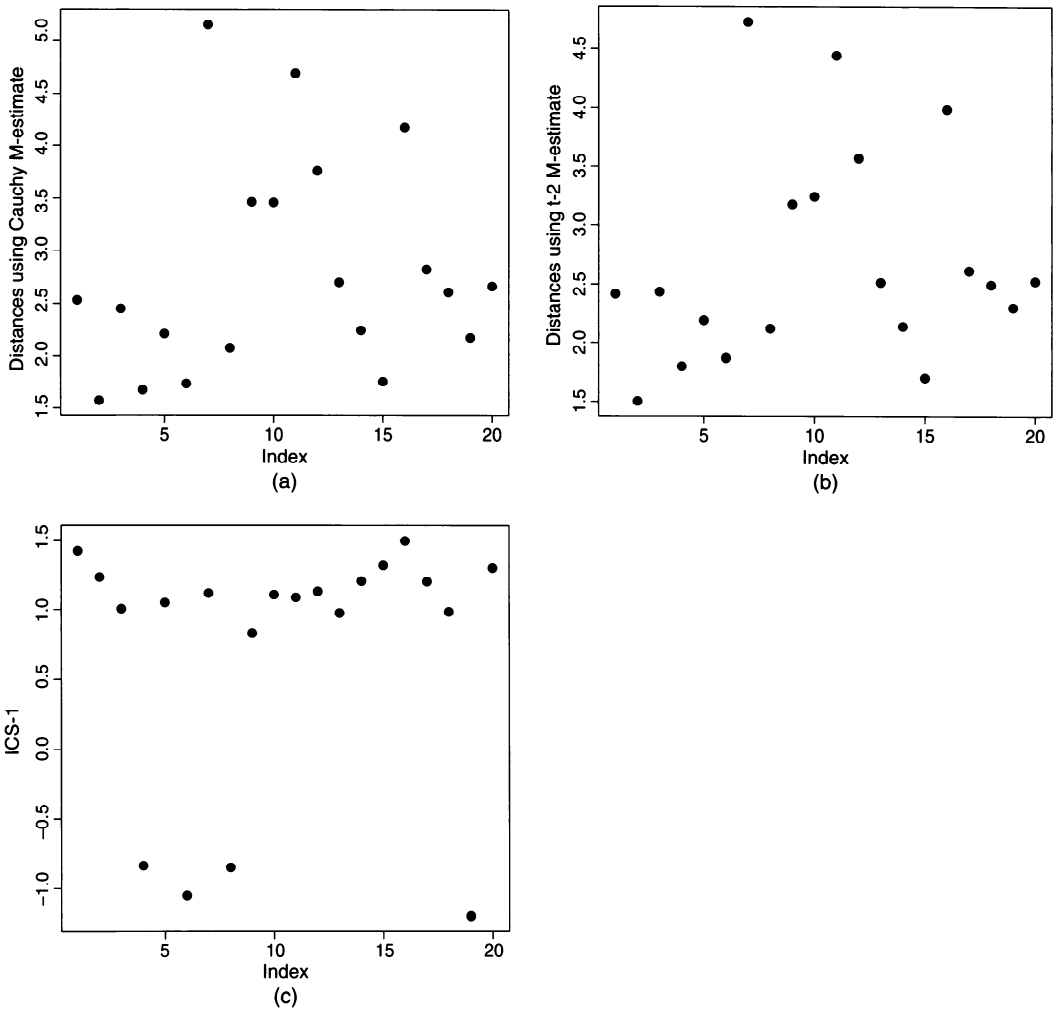

Fig. 4. Example of ICS on the modified wood gravity data set

As a final example, consider the RANDU data set which can be also be found in R (R Development Core Team, 2005); Fig. 5(a). This consists of $n = 300$ observations in $p = 3$ dimensions which are supposedly obtained by a random-number generator. In reality, though, the data lie on parallel planes which are not apparent in the original co-ordinates. However, if we transform this data set to the ICS co-ordinates by using the sample covariance matrix S_n and the one-step W-estimate based on pairwise differences given by

$$\hat{V} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(y_i - y_j)(y_i - y_j)'}{\{(y_i - y_j)' S_n^{-1} (y_i - y_j)\}^2}, \tag{25}$$

then the parallel plane structure in the data becomes apparent in a pairs plot; see Fig. 5(b).

In the last example, the presumed distribution from which the data arise is not an elliptical distribution but rather a uniform distribution within the unit cube, which falls within the class of distributions that was discussed in remark 1. Thus, we might expect that a departure from this presumed distribution would be reflected in an ICS analysis. The parallel lines in the

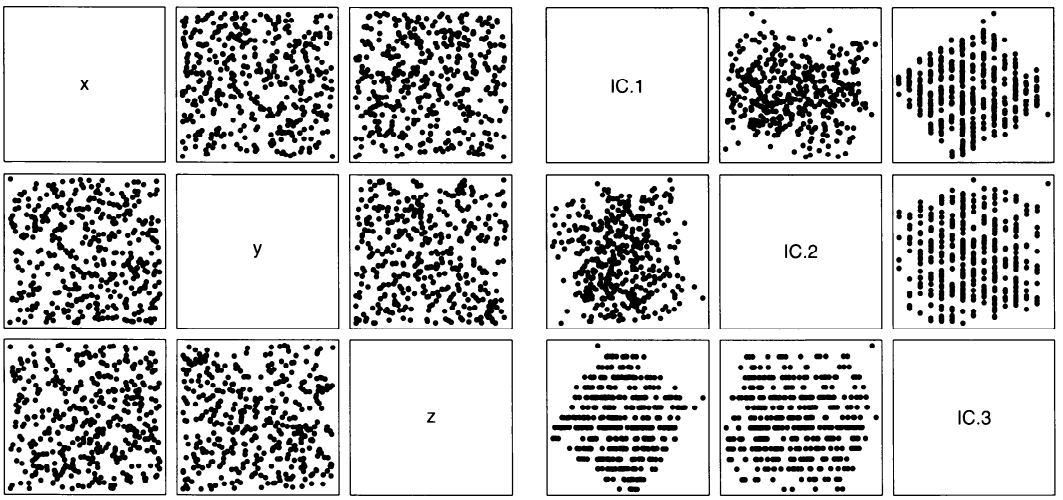

Fig. 5. Example of ICS on the RANDU data set

data set may be viewed as arising from a location mixture of symmetric singular distributions. Consequently the behaviour of an ICS transformation for mixture distributions that was given in Section 5.1, although not directly applicable since the mixture components are not strictly elliptically distributed, gives some rationale about why such a pattern can be detected. For this example, though, the resulting ICS pairs plots are fairly sensitive to the choice of the scatter matrices being used. In particular, if we replace the square term in the denominator of equation (25) with a power of q, then for $q < 1.5$ the lines in the RANDU data set are not very apparent. In contrast, the results do not appear to be heavily dependent on q for $q > 1.5$.

7. Concluding remarks

7.1. Relationship to projection pursuit

Aside from its relationship to mixture models and to ICA, the concept underlying the ICS method has similarities to projection pursuit methods, and in particular to what Huber (1985) referred to as class III projection pursuit approaches, i.e. approaches which investigate the affine invariant aspects of the data. In projection pursuit, we typically seek interesting, usually meaning non-normal, projections of the data; see for example Cook *et al.* (1993), Friedman and Tukey (1974), Huber (1985) and Jones and Sibson (1987). The evaluation of what makes a projection interesting in the projection pursuit context depends solely on the distribution of the particular projection. In general, the pursuit in projection pursuit methods tends to be computationally intensive. In contrast, the value of $\kappa(h)$ that is given by equation (14) used in ICS is not strictly a function of the distribution of the linear combination $h'Y$ but rather is dependent on the multivariate distribution F through $V_1(F)$ and $V_2(F)$. The sequential optimization of $\kappa(h)$ as given by equations (15) and (16) has an analytic solution in terms of eigenvectors and so is not computationally intensive. In this sense, ICS can be viewed as a projection pursuit without the pursuit effort.

The relationship between projection pursuit itself and ICA is well documented; see for example Hyvärinen *et al.* (2001). Almost all algorithms proposed for the ICA problem tend to be of a projection pursuit nature. One notable exception, though, is the previously cited

algorithm that was proposed by Cardosa (1989). It is also worth noting that a relationship between projection pursuit based on kurtosis and multivariate normal mixtures has been observed by Peña and Prieto (2001). They showed that, for a mixture of two multivariate normal distributions with equal covariance matrices, either the projection which minimizes or the projection which maximizes the classical kurtosis coefficient corresponds to Fisher's linear discrimination function between the two elements of the mixture. An analogous result was obtained by Yenyukov (1988) when using the ratio of a robust variance to the sample variance as a projection index. He also proposed to use $\kappa(h)$ based on the sample covariance matrix and a robust estimate of the covariance matrix as an approximation to such a projection index.

7.2. Other related methods

As noted in Section 1, ICS can be viewed as a more general formulation of what was referred to by Ruiz-Gazen (1993) and Caussinus and Ruiz-Gazen (1995) as generalized principal components analysis. They used this terminology since the matrix equation (11) is commonly referred to as a generalized eigenvalue–eigenvector problem. The term generalized principal components analysis, however, is often used in the literature to describe various unrelated generalizations of principal components. Hence, to distinguish this method from other methods that are referred to as generalized principal components, as well as to emphasize a central property of the method, we use the term ICS. Furthermore, we do not view ICS as a generalization of principal components analysis; nor do we consider ICS to be a competitor to either principal components analysis or to a robust version of principal components analysis, but rather as a complementary method. Principal components analysis is concerned with understanding the spread or scatter of a data cloud, which is a property which cannot be identified within an affine invariant setting. As suggested by Huber (1985), a fuller understanding of a data set is obtained by exploring its affine invariant aspects in addition to its location–scale information.

Recently, Critchley *et al.* (2007) proposed to perform a principal components analysis on standardized data, as described in remark 2, which they referred to as *principal axis analysis*. Their proposal thus corresponds to the special case of the ICS transformation when we take $\hat{V}_1 = S_n$, the sample covariance matrix, and $\hat{V}_2$ to be a one-step reweighted covariance matrix with the weights corresponding to the inverse of the classical squared Mahalanobis distances, i.e. $\hat{V}_2$ is a one-step W-estimate as defined in expression (7) with weight functions $u_2(s) = 1/s$ and $u_3(s) = 1$, using the sample mean and covariance matrix as the initial estimates. They referred to their approach as principal axis analysis since $\hat{V}_2$ depends on the standardized sample vectors $X_i = S_n^{-1/2}(Y_i - \bar{Y})$ only through their pairs of opposed directions $\pm X_i/\|X_i\|$. Within Critchley *et al.* (2007), heuristic arguments are given to motivate the use of principal axis analysis for detecting well-separated clusters when the data arise from a mixture of elliptical distributions, even one with possibly different shape matrices.

Another approach for generating affine invariant co-ordinates, which is well known within the area of multivariate non-parametric statistics, is the transformation–retransformation (TR) approach that was proposed by Chakraborty and Chaudhuri (1996, 1998). The basic idea behind the TR approach in the one-sample problem is to transform the multivariate data by multiplying each observation by the inverse of a matrix containing p of the observations. The TR approach though is not invariant under permutation of the n observations, unless either the p observations that are used for standardizing are chosen randomly or some permutation invariant criterion is used to select them. In any event, it is difficult to express the TR approach in terms of functionals, and this makes the theoretical properties of the TR transformation

problematic to study. The TR transformation, however, is not meant to be an exploratory transformation of the data, but rather a step that is used for generating affine invariant multivariate non-parametric tests. Likewise, an ICS transformation can also be used for generating such tests (see for example Nordhausen *et al.* (2006)), or in general for defining multivariate versions of univariate concepts. An affine equivariant componentwise multivariate median for $Y \in \Re^p$, for example, can be defined by μ_Y, where $\mu_Y' = (\text{median}(Z_1), \ldots, \text{median}(Z_p))\, H(F_Y)^{-1}$ with $Z = H(F_Y)'Y$ corresponding to an ICS transformation. In such settings, the main focus of ICS is not on dimension reduction but rather on the complete affine invariant co-ordinate system.

7.3. Summary and continuing research

In this paper, we have introduced the concept of ICS as a general affine invariant method for exploring multivariate data. After removing the effect of the centre and scatter from a multivariate data set, ICS essentially addresses the question of whether there is anything else of interest in the data set. The paper also shows how an ICS transformation theoretically behaves under elliptical mixture models and under ICA models.

From a statistical modelling perspective, one might argue that ICA models may seem impractical except for very specific problems. Nevertheless, ICA algorithms have become increasingly popular in many areas that routinely apply multivariate methods and often yield interesting results even when the ICA model may seem unrealistic. One of the original goals of this paper was to give a model-free explanation about why this might be so. The results in this paper relating ICS with both mixture models and with ICA provide one such explanation.

As noted in Section 6, the theoretical results of this paper apply to essentially any choice of two scatter matrices. The statistical variability and the robustness properties of the ICS transformations though do depend on the particular scatter matrices being used. The results concerning ICS under mixture models that was given in Section 5.1 suggest that the method may have some natural robustness properties, at least in terms of detecting clusters of outliers, even if the estimates themselves are not particularly robust.

As with all eigenvector methods, the stability of the ICS transformations will depend on the spread of the theoretical roots $\rho_1(F), \ldots, \rho_p(F)$, which in turn depends on the choice of the scatter matrices. The effects of the choice of the scatter matrices on the resulting ICS method, as well as the statistical properties that are associated with ICS methods, are currently being studied by the authors and their students. Rather than choose two scatter matrices beforehand, one promising strategy seems to be to allow the data to choose two scatter matrices from among a large class of scatter matrices based on the observed separation of the respective roots. Such a data-driven approach unfortunately makes a theoretical study of the statistical properties of the resulting method far more challenging.

Acknowledgements

The authors are grateful to Klaus Nordhausen for his help in providing R code for the examples and illustrations, to Anne Ruiz-Gazen for bringing to our attention the literature on generalized principal components analysis and to Stefan Van Aelst for providing access to the data set that was used in example 1.

The research of the first author was supported by National Science Foundation grant DMS-0604596. The research of the third author was supported by the Swiss National Science Foundation.

Appendix A: Proofs

A.1. Proof of theorems 1 and 2

From property (3), it follows that there are $\gamma_1 > 0$ and $\gamma_2 > 0$ such that $V_1(F_{Y^*}) = \gamma_1 A V_1(F_Y) A'$ and $V_2(F_{Y^*}) = \gamma_2 A V_2(F_Y) A'$. By definition, $V_2(F_{Y^*})\, h_j(F_{Y^*}) = \rho_j(F_{Y^*})\, V_1(F_{Y^*})\, h_j(F_{Y^*})$ and so

$$V_2(F_Y) A' h_j(F_{Y^*}) = \gamma\, \rho_j(F_{Y^*})\, V_1(F_Y) A'\, h_j(F_{Y^*}), \tag{26}$$

where $\gamma = \gamma_1/\gamma_2$. This implies that condition (18) holds. If $\rho_j(F_Y)$ is a distinct root, then equation (26) also implies that $h_j(F_Y) = a_j A'\, h_j(F_{Y^*})$ for some scalar $a_j \neq 0$, and so

$$Z_j = h_j(F_Y)' Y = a_j\, h_j(F_{Y^*})' A Y = a_j\, h_j(F_{Y^*})'(Y^* - b) = a_j Z_j^* - b_j,$$

which completes the proof for theorem 1. Consider now the case of a multiple root, say $\rho_{(k)} \equiv \rho_{j1}(F_Y) = \ldots = \rho_{j2}(F_Y)$ where $j2 = j1 + p_k - 1$, and let $H_{(k)}(F) = (h_{j1}(F), \ldots, h_{j2}(F))$. As a consequence of a multiple root, the exact choice $H_{(k)}(F)$ is somewhat arbitrary unless some rule is specified about how to choose its columns. However, the span of $H_{(k)}(F)$ is uniquely defined and so, no matter what rule we use to define $H_{(k)}(F)$, equation (26) implies that $H_{(k)}(F_Y) = A' H_{(k)}(F_{Y^*}) B_k'$ for some non-singular matrix B_k. This implies that

$$Z_{(k)} = H_{(k)}(F_Y)' Y = B_k H_{(k)}(F_{Y^*})' A Y = B_k H_{(k)}(F_{Y^*})'(Y^* - b) = B_k Z_{(k)}^* - b_j,$$

which completes the proof for theorem 2.

A.2. Proof of theorems 3 and 4

Since theorem 3 is a special case of theorem 4, it is only necessary to prove the latter. Using the notation of theorem 4, let $M_0 = \Gamma^{-1/2}(M - \mu_k \mathbf{1}_k')$, where $M = (\mu_1 \ldots \mu_k)$ and $\mathbf{1}_k \in \Re^k$ is a vector of 1s. Since M_0 has rank q, the triangular decomposition for matrices gives

$$M_0 = P \begin{pmatrix} T_u & 0 \\ 0 & 0 \end{pmatrix} = P(t_1 \ldots t_q\ 0) = PT$$

with P being an orthogonal matrix of order p and T_u being an upper triangular matrix of order q. The distribution of $X = P'\Gamma^{-1/2}(Y - \mu_k)$ is then a mixture of k spherical distributions with centres $t_1, \ldots, t_k$, where $t_{q+1} = \ldots = t_k = 0$, and spread functions $g_i, i = 1, \ldots, k$, i.e. the density of X is given by

$$f_X(x) = \sum_{j=1}^{k} \alpha_j\, g_j\{(x - t_j)'(x - t_j)\}.$$

The distributions of X and QX are thus the same for any orthogonal Q of the form

$$Q = \begin{pmatrix} I_q & 0 \\ 0 & Q_{22} \end{pmatrix},$$

where I_q is the identity matrix of order q and Q_{22} is an orthogonal matrix of order $p - q$. Thus, given a scatter functional $V(F)$ satisfying condition (3), $V(F_X) = V(F_{QX}) \propto Q V(F_X) Q'$, for any such Q, and so

$$V(F_X) = \begin{pmatrix} V_{11}(F_X) & V_{12}(F_X) \\ V_{12}(F_X)' & V_{22}(F_X) \end{pmatrix} = \begin{pmatrix} V_{11}(F_X) & V_{12}(F_X) Q_{22}' \\ Q_{22}\, V_{12}(F_X)' & Q_{22}\, V_{22}(F_X) Q_{22}' \end{pmatrix} \tag{27}$$

for any orthogonal matrix Q_{22}. Note that equality holds in expression (27) rather than just proportionality since the upper block diagonal matrices are equal (and non-zero). By making appropriate choices for Q_{22} in expression (27) we obtain $V_{12}(F_X) = 0$ and $V_{22}(F_X) = \gamma I_{p-q}$, for some $\gamma > 0$. Thus, for the two scatter functionals $V_1(F)$ and $V_2(F)$,

$$V_1(F_X)^{-1}\, V_2(F_X) = \begin{pmatrix} V_{1,11}(F_X)^{-1}\, V_{2,11}(F_X) & 0 \\ 0 & (\gamma_2/\gamma_1) I \end{pmatrix}.$$

This matrix has at least one root with multiplicity greater than or equal to $p - q$. By theorem 2, we know that the roots of $V_1(F_Y)^{-1}\, V_2(F_Y)$ are proportional to the roots of $V_1(F_X)^{-1}\, V_2(F_X)$, and so at least one of the roots $\rho_{(j)}$ has a multiplicity that is greater than or equal to $p - q$.

Suppose now that no root has multiplicity greater than $p - q$, which by theorem 2 applies to $V_1(F_X)^{-1} V_2(F_X)$ as well as to $V_1(F_Y)^{-1}\, V_2(F_Y)$. For $V_1(F_X)^{-1}\, V_2(F_X)$, one root with multiplicity $p - q$ must

be γ_2/γ_1. Also, the q-dimensional subspace that is spanned by the eigenvectors of $V_1(F_X)^{-1}V_2(F_X)$, other than those associated with γ_2/γ_1, is the same as the subspace that is spanned by $(I_q \quad 0)'$, or equivalently it is the same as the subspace that is spanned by T. From the shape equivariant property (3), we have $V_1(F_X)^{-1}V_2(F_X) \propto P'\Gamma^{1/2}V_1(F_Y)^{-1}V_2(F_Y)\Gamma^{-1/2}P$, and so it follows that if a is an eigenvector of $V_1(F_X)^{-1}V_2(F_X)$ then $h=\Gamma^{-1/2}Pa$ is an eigenvector of $V_1(F_Y)^{-1}V_2(F_Y)$. If the eigenvector a is associated with the root γ_2/γ_1, then h is associated with some root, say $\rho_{(t)}$, with multiplicity $p-q$. The subspace that is spanned by all the eigenvectors of $V_1(F_Y)^{-1}V_2(F_Y)$, other than those associated with $\rho_{(t)}$, is thus the same as the subspace that is spanned by $\Gamma^{-1/2}PT=\Gamma^{-1/2}M_0=\Gamma^{-1}(M-\mu_k 1_k')$, and hence equation (22) holds.

A.3. *Proof of theorem 5*

The symmetry of X, along with X having independent components, implies that $X \sim_d \mathcal{S}X$ for any diagonal matrix $\mathcal{S}$ having only 1s and -1s as entries, i.e. for matrices of the form $\mathcal{S}=\text{diag}(\pm 1,\ldots,\pm 1)$. So, for any scatter functional $V(F)$ which satisfies the shape equivariant property (3), it follows that $V(F_X)=V(F_{\mathcal{S}X})=\mathcal{S}\,V(F_X)\mathcal{S}$, for any such $\mathcal{S}$. The last equality follows from property (3) since the diagonal components of $V(F_X)$ and $S\,V(F_X)S$ are the same. By choosing $\mathcal{S}=(-1,1,\ldots,1)$, we note that all the off-diagonal terms in the first row and in the first column of $V(F_X)$ must be 0. Continuing, we conclude that $V(F_X)$ is a diagonal matrix.

For the two scatter functionals $V_1(F)$ and $V_2(F)$, $V_1(F_X)^{-1}\,V_2(F_X)$ is a diagonal matrix. By theorem 1, it follows that $V_1(F_X)^{-1}\,V_2(F_X) \propto P_\sigma\,\Delta(F_Y)P_\sigma'$, where $\Delta(F_Y)=\text{diag}\{\rho_1(F_Y),\ldots,\rho_p(F_Y)\}$ and P_σ is a permutation matrix. Using property (3) again gives

$$V_1(F_Y)^{-1}\,V_2(F_Y) \propto (B')^{-1}\,V_1(F_X)^{-1}\,V_2(F_X)B' \propto (B')^{-1}P_\sigma\,\Delta(F_Y)P_\sigma'B',$$

which by the spectral value decomposition (13) implies that $H(F_Y)=(B')^{-1}P_\sigma\mathcal{D}$ for some non-singular diagonal matrix $\mathcal{D}$. The theorem then follows since

$$Z=H(F_Y)'Y=\mathcal{D}P_\sigma'B^{-1}(BX+\mu)=\mathcal{D}P_\sigma'X+\mathcal{D}P_\sigma'B^{-1}\mu.$$

A.4. *Proof of theorem 6*

It follows immediately that, if $V(F)$ is a scatter functional satisfying definition 1, then $V(F_X)$ is a diagonal matrix. The remainder of the proof is then identical to the proof of theorem 5.

A.5. *Proof of theorem 7*

By equivariance, we can assume without loss of generality that $\mu=0$. The proof is then given by the first part of the proof to theorem 5.

A.6. *Proof of theorem 8*

The proof of theorem 8 is analogous to the proof of theorems 5 and 6. The only difference is that the matrix $\mathcal{D}$ at the end of the proof is not necessarily a diagonal matrix, but rather a block diagonal matrix with diagonal blocks of order $p_1,\ldots,p_m$.

A.7. *Proof of theorem 9*

The proof of theorem 9 is a generalization of the proof for theorem 5. For this proof, a reference to the blocks of a matrix of order p refers to the partitioning of the matrix in blocks of dimension $p_i \times p_j$ for $i,j=1,\ldots,m$. The symmetry condition on X, along with the assumption that X has mutually independent subvectors, implies that $X \sim_d \mathcal{S}X$ for any block diagonal matrix $\mathcal{S}$ having diagonal blocks of the form $\pm I_{p_k}$, $k=1,\ldots,m$. So, for any scatter functional $V(F)$ which satisfies the shape equivariant property (3), it follows that $V(F_X)=V(F_{\mathcal{S}X})=\mathcal{S}\,V(F_X)\mathcal{S}$, for any such $\mathcal{S}$. The last equality follows from property (3) since the block diagonal components of $V(F_X)$ and $\mathcal{S}\,V(F_X)\mathcal{S}$ are the same. By choosing the first diagonal block of $\mathcal{S}$ to be $-I_{p_1}$ and the other diagonal blocks to be I_{p_k} for $k=2,\ldots,p$, and then continuing in this fashion, we conclude that $V(F_X)$ is a block diagonal matrix.

For the two scatter functionals $V_1(F)$ and $V_2(F)$, $V_1(F_X)^{-1}\,V_2(F_X)$ is a block diagonal matrix. Applying the spectral value decomposition (13) to the block diagonal elements gives $V_{1,jj}(F_X)^{-1}\,V_{2,jj}(F_X)=$

$H_j \Delta_j H_j^{-1}$, with Δ_j being a diagonal matrix of order p_j for $j = 1, \ldots, m$. Let Δ be the diagonal matrix of order p with diagonal blocks Δ_j, and let H be the block diagonal matrix of order p with diagonal blocks H_j. Thus, $V_1(F_X)^{-1} V_2(F_X) = H \Delta H^{-1}$. It follows from theorem 2 that $\Delta \propto P_\sigma \Delta(F_Y) P_\sigma'$, where $\Delta(F_Y) \propto \mathrm{diag}\{\rho_1(F_Y), \ldots, \rho_p(F_Y)\}$ and P_σ is a block permutation matrix. Applying property (3) again gives

$$V_1(F_Y)^{-1} V_2(F_Y) \propto (B')^{-1} V_1(F_X)^{-1} V_2(F_X) B' \propto (B')^{-1} H P_\sigma \Delta(F_Y) P_\sigma' H^{-1} B'.$$

Comparing this with the spectral value decomposition (13) for $V_1(F_Y)^{-1} V_2(F_Y)$ gives $H(F_Y) = (B')^{-1} H P_\sigma \mathcal{D}$ for some non-singular diagonal matrix $\mathcal{D}$. Thus,

$$Z = H(F_Y)' Y = \mathcal{D} P_\sigma' H' B^{-1} (BX + \mu) = \mathcal{D} P_\sigma' H' X + \beta,$$

where $\beta = \mathcal{D} P_\sigma' H' B^{-1} \mu$. Since $\mathcal{D} P_\sigma' = P_\sigma \mathcal{D}$, it then follows that

$$P_\sigma'(Z - \beta) = (Z_{(1)}', \ldots, Z_{(m)}')' - P_\sigma' \beta = \mathcal{D} H' X,$$

with $Z_{(j)} = D_j H_j' X_{(j)}$, for $j = 1, \ldots, m$. Hence, theorem 9 holds.

References

Art, D., Gnanadesikan, R. and Kettenring, J. R. (1982) Data-based metrics for cluster analysis. *Util. Math.* A, **21**, 75–99.

Bilodeau, M. and Brenner, D. (1999) *Theory of Multivariate Statistics*. New York: Springer.

Cardoso, J.-F. (1989) Source separation using higher order moments. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pp. 2109–2112. Glasgow: Institute of Electrical and Electronics Engineers.

Caussinus, H., Fekri, M., Hakam, S. and Ruiz-Gazen, A. (2003) A monitoring display of multivariate outliers. *Computnl Statist. Data Anal.*, **44**, 237–252.

Caussinus, H. and Ruiz-Gazen, A. (1990) Interesting projections of multidimensional data by means of generalized principal component analysis. In *Proc. COMPSTAT'90*, pp. 121–126, Heidelberg: Physica.

Caussinus, H. and Ruiz-Gazen, A. (1993) Projection pursuit and generalized principal component analysis. In *New Directions in Statistical Data Analysis and Robustness* (eds S. Morgenthaler, E. Ronchetti and W. A. Stahel), pp. 35–46, Basel: Birkhäuser.

Caussinus, H. and Ruiz-Gazen, A. (1995) Metrics for finding typical structures by means of principal component analysis. In *Data Science and Its Applications* (eds Y. Escoufier and C. Hayashi), pp. 177–192. Tokyo: Academic Press.

Chakraborty, B. and Chaudhuri, P. (1996) On a transformation and retransformation technique for constructing affine equivariant multivariate median. *Proc. Am. Math. Soc.*, **124**, 2539–2547.

Chakraborty, B. and Chaudhuri, P. (1998) On an adaptive transformation–retransformation estimate of multivariate location. *J. R. Statist. Soc.* B, **60**, 145–157.

Cook, D., Buja, A. and Cabrera, J. (1993) Projection pursuit indexes based on orthonormal function expansions. *J. Computnl Graph. Statist.*, **2**, 225–250.

Critchley, F., Pires, A. and Amado, C. (2007) Principal axis analysis. The Open University, Milton Keynes. Unpublished.

Davies, P. L. (1987) Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, **15**, 1269–1292.

Donoho, D. L. and Gasko, M. (1992) Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, **20**, 1803–1827.

Dümbgen, L. and Tyler, D. E. (2005) On the breakdown properties of some multivariate M-functionals. *Scand. J. Statist.*, **32**, 247–264.

Flury, B. (1988) *Common Principal Components and Related Multivariate Models*. New York: Wiley.

Friedman, J. H. and Tukey, J. W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **23**, 881–890.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics: the Approach Based on Influence Functions*. New York: Wiley.

Huber, P. J. (1981) *Robust Statistics*. New York: Wiley.

Huber, P. J. (1985) Projection pursuit. *Ann. Statist.*, **13**, 435–475.

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

Jones, M. C. and Sibson, R. (1987) What is projection pursuit (with discussion)? *J. R. Statist. Soc.* A, **150**, 1–36.

Kent, J. T. and Tyler, D. E. (1996) Constrained M-estimation for multivariate location and scatter. *Ann. Statist.*, **24**, 1346–1370.

Lopuhaä, H. P. (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.*, **17**, 1662–1683.

Lopuhaä, H. P. (1991) Multivariate τ-estimators of location and scatter. *Can. J. Statist.*, **19**, 307–332.

Lopuhaä, H. P. (1999) Asymptotics of reweighted estimators of multivariate location and scatter. *Ann. Statist.*, **27**, 1638–1665.
Mardia, K. V., Kent, J. T. and Bibby, J. M. (1980) *Multivariate Analysis*. London: Academic Press.
Maronna, R. A. (1976) Robust M-estimators of multivariate location and scatter. *Ann. Statist.*, **4**, 51–67.
Maronna, R. A., Stahel, W. A. and Yohai, V. J. (1992) Bias-robust estimators of multivariate scatter based on projections. *J. Multiv. Anal.*, **42**, 141–161.
Mosteller, C. F. and Tukey, J. W. (1977) *Data Analysis and Regression*. Reading: Addison-Wesley.
Nordhausen, K., Oja, H. and Ollila, E. (2008) Robust independent component analysis based on two scatter matrices. *Aust. J. Statist.*, **37**, 91–100.
Nordhausen, K., Oja, H. and Tyler, D. E. (2006) On the efficiency of invariant multivariate sign and rank tests. In *Festschrift for Tarmo Pukkila* (eds J. Isotalo, E. P. Liski, S. Puntanen and G. P. H. Styan), pp. 217–232. Tampere: University of Tampere.
Nordhausen, K., Oja, H. and Tyler, D. E. (2008) Tools for exploring multivariate data: the package ICS. *J. Statist. Softwr.*, **28**, no. 6.
Oja, H., Sirkiä, S. and Eriksson, J. (2006) Scatter matrices and independent component analysis. *Aust. J. Statist.*, **35**, 175–189.
Peña, D. and Prieto, F. J. (2001) Cluster identification using projections. *J. Am. Statist. Ass.*, **96**, 1433–1445.
R Development Core Team (2005) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. (Available from `http://www.R-project.org`.)
Rousseeuw, P. J. (1986) Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications* (eds W. Grossman, G. Pflug, I. Vincze and W. Wertz), pp. 283–297. Dordrecht: Reidel.
Rousseeuw, P. J. and van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.
Rousseeuw, P. J. and Leroy, A. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
Ruiz-Gazen, A. (1993) Estimation robuste d'une matrice de dispersion et projections révélatrices. *PhD Thesis*. Université Paul Sabatier, Toulouse.
Taskinen, S., Sirkiä, S. and Oja, H. (2007) Independent component analysis based on symmetrised scatter matrices. *Computnl Statist. Data Anal.*, **51**, 5103–5111.
Tatsuoka, K. S. and Tyler, D. E. (2000) On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann. Statist.*, **28**, 1219–1243.
Tyler, D. E. (1994) Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.*, **22**, 1024–1044.
Tyler, D. E. (2002) High breakdown point multivariate M-estimation. *Estadística*, **54**, 213–247.
Visuri, S., Koivunen, V. and Oja, H. (2000) Sign and rank covariance matrices. *J. Statist. Planng Inf.*, **91**, 557–575.
Yenyukov, I. S. (1988) Detecting structures by means of projection pursuit. In *Proc. COMPSTAT 88*, pp. 47–58. Heidelberg: Physica.

Discussion on the paper by Tyler, Critchley, Dümbgen and Oja

J. T. Kent (*University of Leeds*)

This is a delightful paper. Like many of the best ideas in statistics, it is based on a simple, yet elegant, idea which leads to powerful new methods of discovering patterns in data. All the user needs to do is to specify two scatter functionals (effectively two metrics for the specific set of data) and then to carry out a relative eigenanalysis. As the examples of the paper make clear, this new methodology has proved its value in a wide range of settings.

Dual metrics have a long history in multivariate analysis. The simplest example is perhaps principal component analysis itself, which involves the eigendecomposition of one matrix (typically a sample covariance matrix) with respect to another matrix (typically the identity matrix, which is often implicit). Another example arises in multivariate analysis of variance, which is also called discriminant analysis, in terms of the 'between-' and 'within-'groups sums of squares and products matrices B and W (or, alternatively, in terms of $T = B + W$ and W), though in this case prior knowledge of the grouping structure is needed. These two examples involve quadratic functions of the data, whereas the focus in this paper is on non-quadratic functions of the data.

The simplest non-quadratic function of a random variable U is the kurtosis, which takes the form $\mathrm{kurt}(U) = E(U^4) - 3$, when U is centred and scaled to have mean 0 and variance 1. It is useful to distinguish three cases:

(a) $\mathrm{kurt}(U) = 0$, which holds under normality, and the alternatives
(b) $\mathrm{kurt}(U) > 0$ and
(c) $\mathrm{kurt}(U) < 0$,

which can be termed the 'super-Gaussian' and 'sub-Gaussian' cases respectively. The super-Gaussian case arises for long-tailed distributions, whereas the sub-Gaussian case arises for what might be called 'balanced' mixtures of two normal distributions with different means. Here balanced means that the mixing proportions are not too far from $\frac{1}{2}$ and the variances are not too dissimilar. When the variances are equal, it is possible to characterize the class of balanced mixtures explicitly; see, for example, Section 5.1 of the paper or Peña and Prieto (2001).

One way to look for structure in a p-dimensional random vector Y (with mean 0 and covariance matrix Σ) is to look for a linear combination a to maximize the *absolute* kurtosis, $|\text{kurt}(a'y)|$, and this criterion forms the basis of one of the standard algorithms in independent component analysis (ICA). However, the above paragraph suggests that, when clustering is suspected, a better approach might be to minimize the *signed* kurtosis $\text{kurt}(a'Y)$, which leads to what can be termed a 'sub-ICA' algorithm; see, for example Bugrien and Kent (2005) for more details.

The set of all fourth-order moments forms a four-way array (and hence does not define a metric). Since the kurtosis involves a quartic function of a acting on this four-way array, optimization for either of these algorithms must generally be carried out numerically.

The paper finesses its way out of this numerical problem by replacing the full four-way array of fourth-order moments by a matrix of selected fourth-order moments, $\mathcal{K} = E\{(Y'\Sigma^{-1}Y)YY'\}$ in equation (21), and replacing the quartic optimization by a quadratic optimization. Thus a natural question is which criterion offers more insight into the structure of multivariate data,

$$\text{kurt}(a'Y) \quad \text{or} \quad a'\mathcal{K}a/a'\Sigma a?$$

Further, do any insights here offer any guidance to the analysis of more general scatter functionals?

Several other questions also spring to mind.

(a) *Ordering of eigenvalues*: the paper makes little distinction between $V_1^{-1}V_2$ and $V_2^{-1}V_1$. Would it be helpful to label one of the matrices as 'more robust' than the other and to distinguish between the interpretation of the largest eigenvalue and the smallest (cf. sub-ICA above)?

(b) *Estimating the centre of the data*: this topic has received little discussion in the paper, but it seems potentially important. Does choice of location functional matter, especially for skew data? Or does symmetrization successfully deal with the issue? If symmetrization is not used, would it be desirable to enforce a common estimate of location when defining the matrices V_1 and V_2?

(c) *Third moments*: this paper works with and extends fourth moments (kurtosis). Is it worth investigating and extending third moments (skewness)?

(d) *High dimension*: the examples in this paper involve data sets of fairly modest dimension. Are there opportunities for insights with high dimensional data ($n < p$ or $n \ll p$), after regularizing?

Let me end with a more philosophical question. The authors motivate the methods in the paper by using ideas from robustness theory, which was developed to protect against outliers. However, this paper is more concerned with pattern detection, which is a more subtle problem. Is it merely serendipity that methods developed for one problem provide tools for another, or is there something deeper going on?

This has been a fascinating paper opening up a whole new direction in the search for patterns in multivariate data. It gives me great pleasure to propose the vote of thanks.

Trevor Ringrose (*Cranfield University, Swindon*)

Multivariate analysis often seems to be a randomly assorted grab-bag of vaguely related methods rather than a coherent field, so it is very encouraging to see a paper which shows the connections between several methods and even more importantly opens up a wide array of potentially useful generalizations and special cases.

The authors rightly point out that when different robust estimators of ostensibly the same parameter produce different answers this is not necessarily a bad thing, as there is information in these differences. Similarly in introductory statistics lectures we often mention that the mean, median and mode are all roughly the same for samples from symmetrical distributions, so it tells us something useful if they are all different. The paper offers a convincing method for making such comparisons in a multivariate setting, which the reader can understand by analogy with the very similar use of within-groups and between-groups covariance matrices in multivariate analysis of variance and canonical variate analysis.

However, the job of the seconder of the vote of thanks is to be more critical, so we might ask the obvious question of how well do the methods proposed work in practice, and in particular what do they add to what we already have? Some of the examples are not very convincing. It was admitted during the verbal presentation that the outlying cluster can in fact be seen quite clearly in a matrix plot of the wood gravity

data. Ignoring the distinction between response and explanatory variables (as in the paper) a biplot of the principal component analysis (PCA) solution (79% of variance on the first two axes) clearly picks out the cluster and shows that they have above-average values on x_2 and x_5 and below-average values for the other variables, as can then be seen clearly in the raw data. Similarly, distinguishing between the species in Fisher's iris data is trivially easy because again a simple matrix plot shows very clear differences in petal sizes, and the authors note in Nordhausen *et al.* (2008) that simple PCA does almost as well as invariant co-ordinate selection. Admittedly this is a mainly theoretical paper so these data sets were chosen for illustration rather than real interest, but even given this they still seem excessively easy. Similarly, the picture mixing example of invariant co-ordinate selection as independent component analysis in Nordhausen *et al.* (2008), pages 24–26, can be performed almost equally well (in R) by using PCA, and in this example PCA seems to cope better with cases where the number of output mixtures exceeds the number of input signals.

It is a criticism of all of us that we tend to use and reuse the same toy examples in published work, which might easily make the cynical outsider suspicious that our methods work only in certain restricted cases. In particular, I would like to propose a moratorium on further published use of Fisher's iris data!

I have two final comments. Firstly, the paper concentrates very much on the co-ordinate scores on the new axes, but in many cases the eigenvector coefficients will also be of interest. Can meaningful biplots be produced? Secondly, the final paragraph mentions data-driven choice of scatter matrices based on the observed separation of eigenvalues (with larger separations regarded as better, one assumes). Sample eigenvalues are usually more spread out than population eigenvalues anyway, and this will then tend to pick the biggest of these overestimates of the separation. Is this good or bad, though? It might turn out to be good when searching for outliers but bad when trying to model the bulk of a distribution.

While reading the paper it seemed very clear that the authors must have started the work separately and from differing perspectives. One of the authors confirmed that this was indeed so: that they had worked independently until three of them realized that they had all talked about the same thing at a conference. It is a pleasingly self-referential aspect of the paper that these three independent components of the work can be unmixed by the reader.

The criticisms above are very minor, however, as this is a very interesting paper which points the way to many more papers developing the methods and their practical application. In recent years developments in multivariate statistics seem to have fallen behind those in areas such as regression modelling and Bayesian methods, and this paper should help to spark new interest in the field. This paper is a very welcome addition, and I have no hesitation in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Davy Paindaveine (*Université Libre de Bruxelles*)

Beyond the role that it plays in detecting departures from ellipticity, invariant co-ordinate selection (ICS) is potentially useful to choose a proper model for the data at hand among the many multivariate models that are available in the literature: (mixtures of) elliptical models, the independent component (IC) models of Section 5.2 (see also Nordhausen *et al.* (2009b)), skew elliptical models (see, for example, Genton (2004)), etc. This discussion partly supports this claim by proposing an informal graphical method that allows us to 'test' the null hypothesis $\mathcal{H}_0^{\mathrm{IC}}$ under which IC models are appropriate.

In Fig. 1(c), it is shown how a couple of location–scatter estimates $(\hat{\mu}_l, \hat{V}_l)$, $l = 1, 2$, can be used to detect departures from ellipticity, on the basis of the fact that, for any such couple and under ellipticity, we should have $d_i(\hat{\mu}_2, \hat{V}_2) \approx \lambda\, d_i(\hat{\mu}_1, \hat{V}_1)$ for some $\lambda > 0$. For $\mathcal{H}_0^{\mathrm{IC}}$, we could similarly think of using three—or four—different scatter estimates to derive—typically, via theorem 5—a couple of consistent estimates $\hat{H}_l$, $l = 1, 2$, for the underlying mixing matrix H (clearly, it is crucial to adopt a common normalization for $\hat{H}_1$, $\hat{H}_2$ and H here, such as the Z-standardization in the R package ICS; see Nordhausen *et al.* (2008) for details). Although proper (Frobenius-type) distances between the resulting $\hat{H}_1$ and $\hat{H}_2$ would provide natural test statistics for $\mathcal{H}_0^{\mathrm{IC}}$, a direct graphical tool, in the same spirit as in Fig. 1(c), is the scatter plot of *ICS distances* $(d_i^{\mathrm{ICS}}(\hat{H}_1), d_i^{\mathrm{ICS}}(\hat{H}_2))$, $i = 1, \ldots, n$, with

$$d_i^{\mathrm{ICS}}(\hat{H}_l) := \sqrt{\{(\hat{H}_l' Y_i - \hat{\mu}_l^{\mathrm{ICS}})'(\hat{\Lambda}_l^{\mathrm{ICS}})^{-2}(\hat{H}_l' Y_i - \hat{\mu}_l^{\mathrm{ICS}})\}},$$

where $\hat{\mu}_l^{\mathrm{ICS}}$ is the vector of marginal medians for the lth ICS and $\hat{\Lambda}_l^{\mathrm{ICS}}$ is the diagonal matrix collecting the corresponding marginal median absolute deviations. Under $\mathcal{H}_0^{\mathrm{IC}}$, all points in such scatter plots should roughly sit on the main diagonal, which allows us to detect possible violations of $\mathcal{H}_0^{\mathrm{IC}}$.

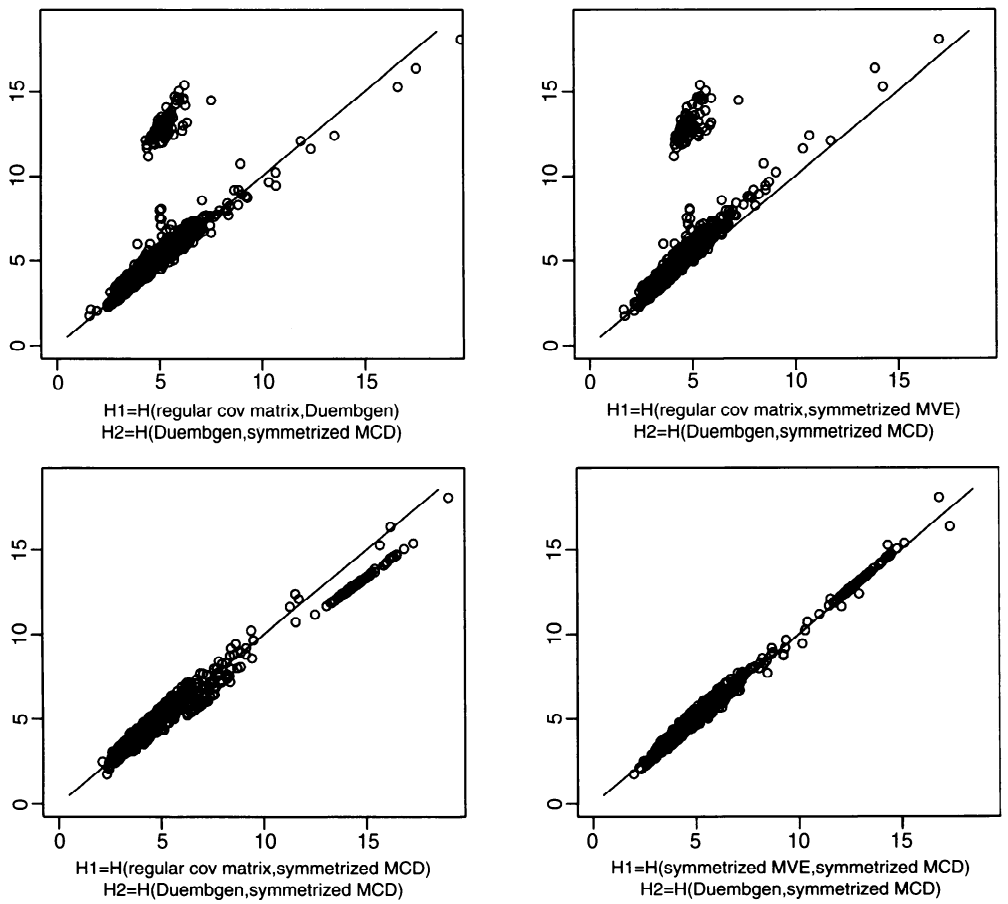

Fig. 6. Scatter plots of ICS distance $(d_i^{\mathrm{ICS}}(\hat{H}_1), d_i^{\mathrm{ICS}}(\hat{H}_2))$, $i = 1,\ldots,n$, with $\hat{H}_l = H(\hat{V}_{l1}, \hat{V}_{l2})$

The choice of the various scatter matrices is, here as well, a delicate issue. But one might still argue that combining scatter matrices with different robustness properties could reveal interesting features. This is illustrated (with the same data as in Section 6.1) in Fig. 6, where, interestingly, only the plot based exclusively on robust scatter matrices seems to be compatible with $\mathcal{H}_0^{\mathrm{IC}}$.

As shown beautifully in the paper, though, the relevance of ICS extends far beyond IC models, and I congratulate the authors for one of the most refreshing and inspiring works of the decade in the field of multivariate statistics.

Mervyn Stone *(University College London)*

This useful paper starts with Cartesian co-ordinates that come with any data, graduates to matrices and ends up with affine invariance—in other words, next door to the open-air geometry of co-ordinate freedom!

I doubt whether the authors depended on the algebra of Sections 2–4 to be confident that that would happen—before writing the computer program that does have to use co-ordinates and matrices. Readers of the paper might have been spared the algebra—if only that great exponent of co-ordinate freedom, Paul Halmos, had gone deeper into probability and statistics to wean us off co-ordinates and matrices wherever and whenever these impede understanding.

It is not too late to supply the alternative thin gruel.

(a) A few concepts and terms from the thinnest and least influential books on multivariate analysis: $\mathcal{V}$ is the vector space of variables (made out of p names) and $\mathcal{E}$ is its dual space of evaluators e whose evaluation of variable v (a possible 'observation' if v is a name) is the bilinear product $[e, v]$. V_1 and

V_2 are inner products on $\mathcal{V}$ and also so-called 'covariance operators' (linear $\mathcal{V} \to \mathcal{E}$).

(b) Realization that fixed point theory can open the door to a simplified equivalent eigenanalysis for V_1 and V_2: $\mathcal{S} = \{v : (V_1 + V_2)(v, u) = 1$ and $(V_1 + V_2)(v, u) \geqslant 0$ for some fixed $u\}$ is the closed surface of a $(V_1 + V_2)$-hemisphere in $\mathcal{V}$. The transformation $T : \mathcal{S} \to \mathcal{S}$ that is defined by $s \to \rho(s) V_2^{-1} V_1 s$ is continuous. So $\mathcal{S}$ has a fixed point h with $V_2 h = \rho(h) V_1 h$ and, as a consequence, you can take it from here with a willingness to 'go to the pictures'.

(c) The pictures that I refer to here are downloadable and are more fully explained in Stone (2008). Their reassuring features are affine invariants as obvious as three lines meeting in a point—and simply discovering them can be a more rewarding and liberating activity for a statistician than sudoku.

Christian Hennig (*University College London*)

The authors did a good job in providing a framework for a class of projection methods to visualize multivariate data sets. The comments on the choice of shape matrices mainly focus on robustness aspects. I think that other considerations are important as well, and in many situations the choice matters more than the paper suggests.

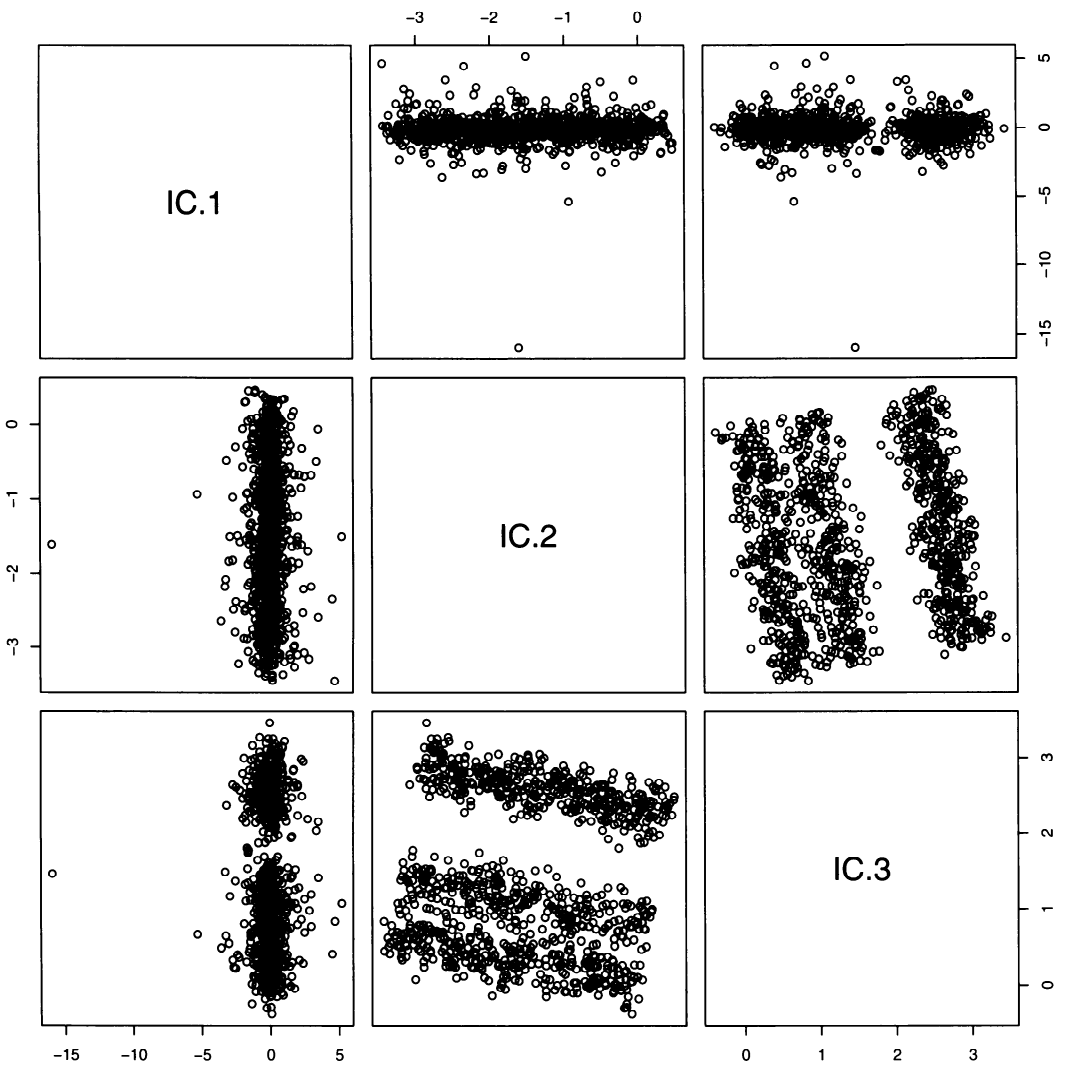

Fig. 7. ICS plot with S_n and $\mathcal{K}_n$

I show a situation in which the choices that are suggested in the paper and the ICS software package do not work well, and an alternative shape matrix does better.

This needs a definition of quality, depending on the patterns of interest, which are clusterings here. What should a good projection method deliver in a benchmark situation with a one-dimensional interesting pattern in a three-dimensional data set? The analogue of what is expected in a high dimensional situation is that the pattern should appear along either the first or the last invariant co-ordinate.

The three variables of the example data set have been generated independently from a t_2-distribution, a uniform distribution and a mixture with 300 points from $\mathcal{N}(0, 1)$, 300 points from $\mathcal{N}(4, 2.25)$ and 400 points from $\mathcal{N}(12, 1)$. Fig. 7 shows the solution with the default of the ICS software ($V_1 = S_n$ and $V_2 = K_n$; this is similar to the solution with V_1 maximum likelihood (ML) for Cauchy, V_2 ML for t_2). The cluster pattern is not optimally visible along the third co-ordinate. In Fig. 8, ML for t_2 and the minimum covariance determinant (MCD) have been used. This shows the pattern along the second co-ordinate.

Fig. 9 shows the best solution, which stems from MCD as V_2 and V_1 ('local shape') defined as follows.

(a) Compute a matrix of Mahalanobis distances between points (based on the MCD with 20% breakdown point, say).

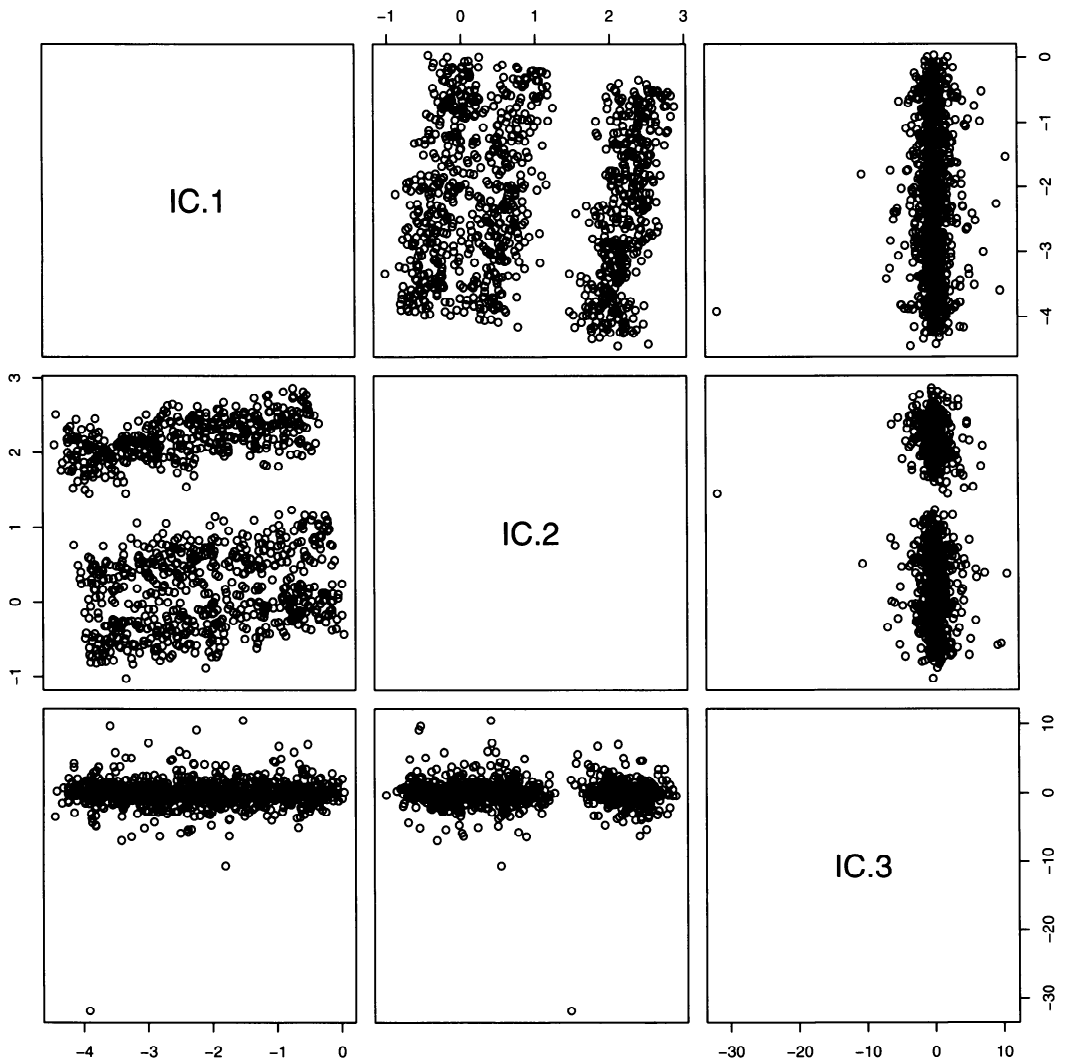

Fig. 8. ICS plot with ML for t_2 and MCD

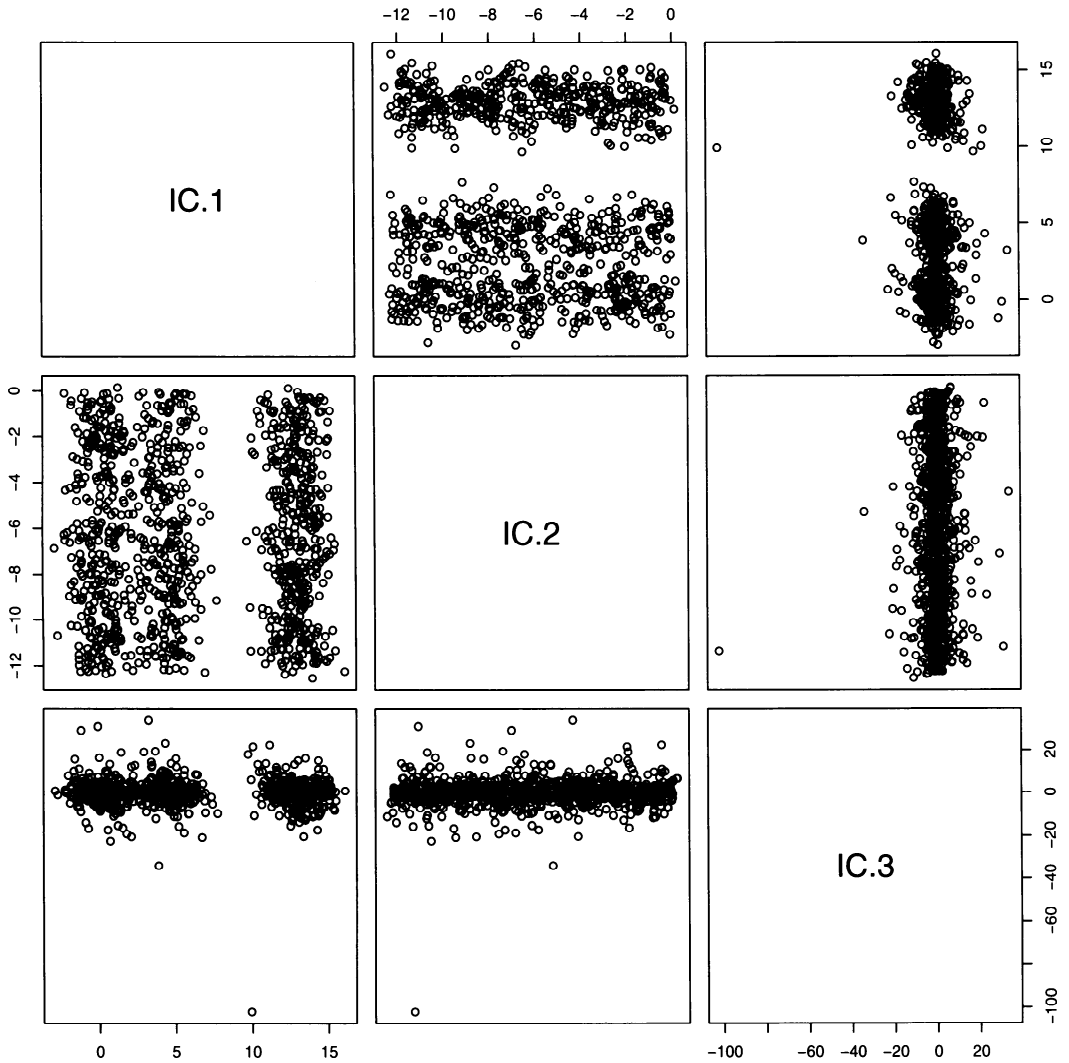

Fig. 9. ICS plot with local shape and MCD

(b) For every point, compute the covariance matrix of its 10% nearest neighbours.
(c) Standardize all these matrices by their traces to unify the influence of every point.
(d) Pool the covariance matrices.

Using this matrix together with a global covariance matrix brings forth those co-ordinates along which the local structure differs from the global structure.

Here is another idea.

(a) Compute an affine invariant clustering of the data.
(b) Use the pooled within-cluster covariance matrix.

Conlusion: if clustering is of interest, it is advantageous to choose scatter matrices to explore global *versus* within-cluster structure.

A. P. Dawid (*University of Cambridge*)

The central idea of this paper is very neat: that, in the presence of two different measures of scatter, defining

two different inner products over the variables, we can apply simultaneous diagonalization to define a 'natural' set of basic variables for further analysis and display of the data. However, this set is natural only to the extent that the chosen pair of scatter measures can be considered natural. But, even in this case, why stop at two such measures?—in many problems there will be a wide variety of interesting scatter measures. Unfortunately the theory as presented requires, not one, not three, but exactly two scatter measures.

Is there anything useful that can be said about an appropriate treatment (in a symmetrical fashion) of more than two?

The following contributions were received in writing after the meeting.

Henri Caussinus (*Institut de Mathématiques de Toulouse*) **and Anne Ruiz-Gazen** (*Toulouse School of Economics*)

We congratulate the authors for their very interesting paper which brings significant improvements in the theoretical knowledge of scatter matrices comparison. From our perspective several issues deserve further attention. The first issue is the choice of the dimension of the graphical display, which has been a crucial concern since the earlier time of the projection pursuit approach (Sun, 1991): which projections are significant, i.e. which projections contain a genuine structure rather than merely random variation corresponding to elliptical distributions? For example, within the framework of theorem 4, what is the value of k or, more precisely, what is the dimension of the subspace containing the $k\mu_j$? The answer to this practical question rests on the distribution of eigenvalues of the matrix product involved. We gave very preliminary theoretical results for specific scatter matrices in Caussinus *et al.* (2003a) for the detection of outliers and in Caussinus *et al.* (2003b) for the detection of groups. Another issue is the complementary use of invariant co-ordinate selection and classification. The co-ordinates that are selected by invariant co-ordinate selection can be used to visualize possible groups, to suggest their number and to improve the efficiency of clustering algorithms. These various aspects were illustrated in Caussinus and Ruiz-Gazen (2007) as an encouragement for further research. A third issue concerns the choice of the (class of) scatter matrices to be compared with respect to the structure of interest. From our experience, many choices lead to displaying outliers. Since, in practice, outliers are often present in the data sets, they can mask other interesting features. To display groups or special structures like those of example 2 or the RANDU data set, much care is needed in the choice of scatter estimators to be compared, and all the more so in the presence of outliers. It seems that scatter matrices resting on pairwise differences are of special interest. A class of scatter matrices depends on a tuning parameter whose choice is also challenging. As quoted by Tyler and his colleagues, some of our results lead to choosing small values of this parameter; this is basically the case when looking for outliers. However, in other cases of interest, our practice and some limited unpublished results lead to different values, e.g. 2, the value which appears in Caussinus *et al.* (2003a). We hope that the authors will be interested in further investigating these various issues.

Christophe Croux (*Katholieke Universiteit Leuven*)

This paper introduces a new tool for multivariate data analysis, called invariant co-ordinate selection. I consider the ideas in this paper to be new and innovative, and this paper is very likely to result in a new stream of research in multivariate analysis. I congratulate the authors for this fascinating paper, and for the clear exposition of their work.

The method is quite easy to put in practice: you compute eigenvalues and eigenvectors of V_1 and V_2, with V_1 and V_2 two scatter matrices. The idea only works if V_1 and V_2 are different scatter matrices. The reason why this method has not been discovered earlier is probably because most statisticians only use the covariance matrix. Scatter matrices are well known in the robustness literature, but application of the methods here does not require the scatter matrices to be robust. What I consider as most important contributions are as follows.

(a) The introduction of an 'invariant co-ordinate system': an affine transformation of the data is not changing the co-ordinate system. Principal component analysis only has this property for orthogonal transformations. The invariant co-ordinate system depends on the choice of the two scatter matrices and yields arbitrary transformations for elliptical distributions. Also in principal components, one computes eigenvectors of a chosen scatter matrix (most often the covariance matrix), and one obtains arbitrary rotations for spherical distributions.

(b) The result that invariant co-ordinate selection retrieves
   (i) the independent components of the independent component analysis model and
   (ii) Fisher's linear discriminant subspace for mixtures of elliptical distributions.

If we have neither a mixture of elliptical distributions, nor an independent component analysis model, then the interpretation of the selected co-ordinates is relying on a projection pursuit argument, where one generalizes a generalized measure of kurtosis. Note that this measure of kurtosis is defined conditionally on a given multivariate distribution. For an arbitrary univariate distribution, it is not so clear how this generalized measure of kurtosis is defined.

Whereas most of the theory in multivariate statistics relies on elliptical distributions, the authors go one step beyond this and open a whole new area of research. I liked reading the paper, and I congratulate the authors once more.

Peter Filzmoser (*Vienna University of Technology*)

I congratulate the authors for this interesting contribution that combines and generalizes several approaches. The work of Caussinus and Ruiz-Gazen on generalized principal components analysis is generalized, and Fisher's linear discriminant subspace turns out to be a special case. Also independent component analysis and projection pursuit are taken into account. For the latter method, invariant co-ordinate selection (ICS) does not require the pursuit effort. In contrast, one could evaluate the co-ordinate pairs resulting from ICS for their 'interestingness', thereby using standard projection pursuit indices. Moreover, as already indicated by the authors, different pairs of scatter matrices could be used to find interesting projections. A further idea could be to use linear combinations of scatter matrices and to combine them in the same way as is done now with equation (13). Depending on the coefficients for the linear combinations, different insights into the multivariate data structure could be obtained.

An interesting aspect of ICS is that the pairs plots offer the possibility of interpreting the outliers. For instance, in example 1 (Fig. 2) the directions of the first two ICS components refer to the contributions of the nine variables. Thus, by inspecting these 'loadings' it could be possible to interpret the outlier groups in terms of the original variables.

Finally, thanks to the available R package 'ICS' I did some experiments with high dimensional data. I generated two multivariate normally distributed data clouds in 1000 dimensions, the first cloud consisting of 2000 observations, and the second of 200, and both centred at the origin. The covariance matrices are the identity matrix for the first cloud and the identity matrix multiplied by 1.2 for the second cloud. Thus, it is practically impossible to distinguish both groups in any pairs plot. With the default parameters for the 'ics' function we can see slightly different behaviour of both groups in the first and last ICS directions. When taking the classical covariance matrix of the original and of the weighted data, with weights obtained from a multivariate outlier detection method, we can clearly see both groups. Here I used an outlier detection method that is not affine equivariant (Filzmoser *et al.*, 2008) and, although theoretical results for ICS would no longer hold, the practical results are very useful.

Marc Hallin (*Université Libre de Bruxelles*)

This paper, which brings together and unifies fundamental ideas from several statistical areas—principal components, discriminant analysis, robustness, invariance, statistical depth, flexible modelling, independent component analysis, ...—is certainly among the most stimulating and refreshing that I have read for many years.

Focusing on the use of two distinct measures of scatter (or shape) $V_1(F)$ and $V_2(F)$ in detecting departures from ellipticity, one question, which is not examined by the authors, naturally comes to mind: for given non-elliptical F, is there any such thing as a 'most efficient' or 'most contrasting' choice of $F \mapsto V_j(F)$, $j = 1, 2$—maximizing, for instance, some adequate distance between the scaled version of $(\rho_1, \ldots, \rho_p)$ and $(1, \ldots, 1)$? This question, quite presumably, is related to the problem of constructing 'optimal' tests for sphericity (robust alternatives to the traditional Mauchly (1940) and John (1972) tests can be found in Tyler (1982, 1987) and Hallin and Paindaveine (2006)). Answering such a question would be most useful, for instance in the problem of recovering, in an optimal way, independent components in independent components analysis models.

Affine invariance or equivariance, however, is not the only invariance property that we could require for the scatter matrices $F \mapsto V_j(F)$, $j = 1, 2$. Another group of transformations, of equal relevance, is not mentioned, which also preserves ellipticity: the group of *monotone radial transformations*. More precisely, assuming that some location $\theta = \theta(F)$ has been chosen, consider a scatter functional $F \mapsto V(F)$ (in the sense of this paper), and let

$$r_V := \{(Y-\theta)' V^{-1} (Y-\theta)\}^{1/2},$$
$$U_V := V^{-1/2}(Y-\theta)/\|V^{-1/2}(Y-\theta)\| = V^{-1/2}(Y-\theta)/r_V.$$

Then, Y (with distribution function F_Y) is elliptical if and only if $Y_V^g := \theta + g(r_V)V^{1/2}U_V$ (with distribution function $F_{Y_V^g}$) is also elliptical, where $r \mapsto g(r)$ is an arbitrary continuous monotone increasing transformation of $\mathbb{R}^+$ such that $g(0)=0$ and $\lim_{r\to\infty}\{g(r)\}=\infty$. Classical invariance arguments suggest that $V(F_{Y_V^g})$ be proportional (shape equivalent) to $V(F_Y)$ for any g and F_Y—a property that the scatter functionals considered in the paper only have when restricted to the family of elliptical F_Ys. This invariance under radial transformations severely restricts the class of admissible scatter functionals; note that the functional that was proposed by Tyler (1987) satisfies the condition—but other solutions do exist.

In the empirical version (denote by $F_Y^{(n)}$ the empirical distribution function for a sample $Y_1, \ldots, Y_n$ of size n), similar invariance arguments imply that $V^{(n)} := V(F_Y^{(n)})$ should be measurable with respect to $U_{V^{(n)};i} := (V^{(n)})^{-1/2}(Y_i-\theta)/\|(V^{(n)})^{-1/2}(Y_i-\theta)\|$ and the ranks $R_{V^{(n)};i}^{(n)}$ of the distances $r_{V^{(n)};i} := \{(Y_i-\theta)'(V^{(n)})^{-1}(Y_i-\theta)\}^{1/2}, i=1,\ldots,n$. This is not easily achieved for finite n, but it holds for the M-estimator that was proposed by Tyler (1987) and, under asymptotic form, for the R-estimators of shape that were developed in Hallin *et al.* (2006).

Daniel Peña and Júlia Viladomat (*Universidad Carlos III de Madrid*)

The authors present a very general method to generate an affine invariant co-ordinate system by projecting the data onto some eigenvectors of the matrix $V_1^{-1}V_2$, where V_1 and V_2 are any pair of (robust) affine equivariant scatter matrices. These projections are shown to reveal departures from an elliptical distribution and can be seen as a projection pursuit method based on kurtosis (see equation (14)). Projection directions maximizing and minimizing kurtosis were shown to be useful for robust multivariate estimation in Peña and Prieto (2001b), who also proved the optimality properties of these directions for clustering (Peña and Prieto, 2001a). They used numerical optimization to find these optimal directions. An important contribution of this paper is that these directions can also be obtained as eigenvectors of some general class of kurtosis matrices.

Thus, we have two ways of finding extreme directions of kurtosis. The first way is through numerical optimization and the second finds the eigenvectors of some generalized kurtosis matrix. In Peña *et al.* (2008) we have compared these two approaches in a particular case. Given a multivariate random vector X with mean μ and covariance matrix Σ, we propose to compute the eigenvectors of the kurtosis matrix $K = E(Z^{\mathrm{T}}ZZZ^{\mathrm{T}})$, where $Z = \Sigma^{-1/2}(X-\mu)$. Using this matrix is equivalent to choosing $V_1 = \Sigma$, and $V_2 = E\{Z^{\mathrm{T}}Z(X-\mu)(X-\mu)^{\mathrm{T}}\}$ in this paper. We then show that if the ratio n/p is large, where n is the sample size and p the dimension, the estimation of a matrix of dimension p is reliable and estimating its eigenvectors becomes accurate and useful. Also, in this case numerical optimization is computationally intensive. However, when n/p is small, estimating the elements of the matrix has limited precision and the eigenvectors are not useful for showing the clusters. Since the use of the kurtosis matrix K is based on an existent kurtosis-based algorithm, we can use the algorithm in Peña and Prieto (2001a) when n/p is small. An interesting problem is the performance of these two procedures under the more general situation of different scatter matrices. Then the use of just any pair of robust scatter matrices does not guarantee the identification of the clusters, whereas the directions of extreme kurtosis have been found to be effective in this situation.

Werner A. Stahel and Martin Mächler (*Eidgenössiche Technische Hochschule, Zurich*)

The paper introduces an elegant piece of theory and derives a very useful tool for finding patterns in multivariate data. We warmly congratulate the authors for this work.

This comment recalls a benchmark distribution for multivariate tools that aim at good robustness properties, which was introduced in section 5.5a of Hampel *et al.* (1986), which we shall call the 'barrow wheel'. It is a mixture of a flat normal distribution contaminated with a portion $\varepsilon = 1/p$ of gross errors concentrated near a one-dimensional subspace. Let

$$G_0 = \left(1-\frac{1}{p}\right)\mathcal{N}_p\{\mathbf{0}, \mathrm{diag}(\sigma_1^2, 1, \ldots, 1)\} + \frac{1}{p}H,$$

where p is the dimension and H is the distribution of $\mathbf{Y}$, where $Y^{(1)}$ has a symmetric distribution with $(Y^{(1)})^2 \sim \chi_{p-1}^2$ and is independent of $Y^{(2)}, \ldots, Y^{(p)} \sim \mathcal{N}_p(\mathbf{0}, \sigma_2^2 I_{p-1})$ (Fig. 10(b)). Then, this distribution is rotated such that the $X^{(1)}$-axis points in the space diagonal direction (1, 1,...,1), and the components are rescaled to obtain G. Note that the covariance matrix of both G_0 and G will tend to I_p for $\sigma_1 \to 0$ and $\sigma_2 \to 0$, and all known 'cheap alternatives' to high breakdown ('class III') point covariance estimation fail to detect the outlier part H. For more details and R functions, see `http://stat.ethz.ch/research/areas/robustness`.

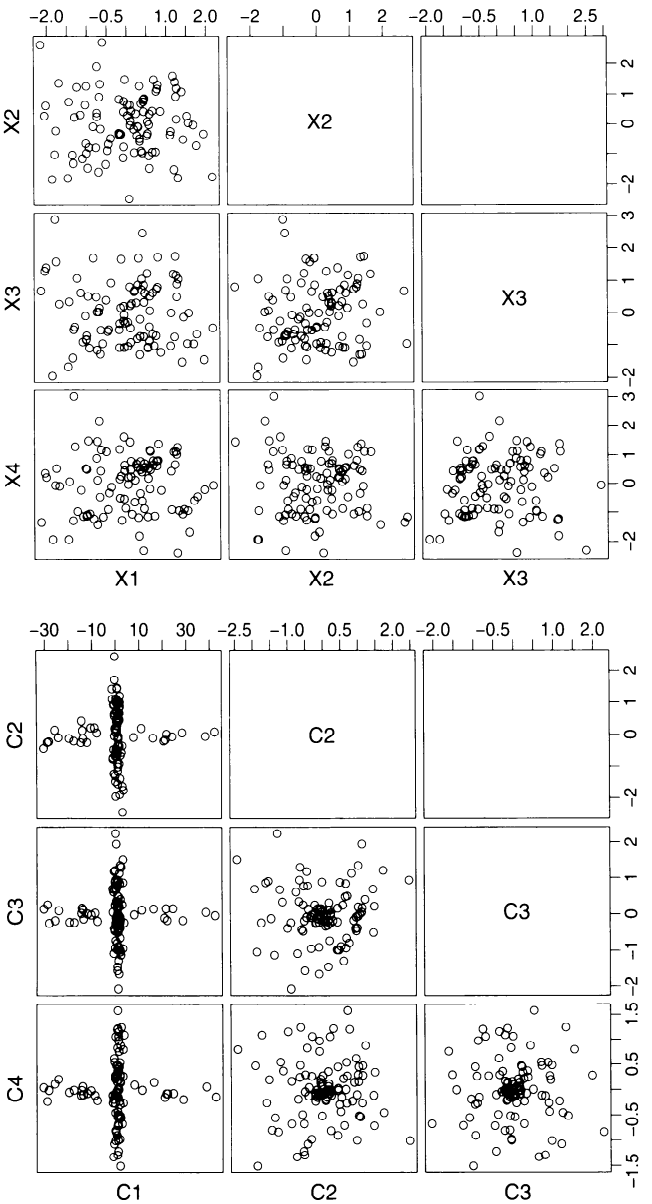

Fig. 10. Scatter plot matrices of a sample from the barrow wheel distribution, $p = 4$, and of the invariant co-ordinates obtained from it

Is the barrow wheel an artificial situation? The 'wheel' describes a multivariate normal distribution with a strong linear relationship between variables—a situation which multivariate statistics searches for. The outliers are 'nasty', but making them more realistic does not render the problem of detecting the structure much easier. Robust multivariate procedures should therefore pass this benchmark.

Fig. 10(a) shows a sample from G for $p = 4$ and $\sigma_1 = 0.1$ and $\sigma_2 = 0.2$. Any structure seems difficult to spot. The invariant co-ordinate selection that is obtained from using the robust MCD covariance as V_2 and the empirical covariance matrix as V_1 shows the structure very clearly (Fig. 10(b)). Note that the outliers would appear in the last co-ordinates if we followed the advice of the authors to use a mildly or non-robust scatter estimate as V_1 and a more robust estimate as V_2.

Thus, invariant co-ordinate selection passes the benchmark—if a high breakdown scatter matrix is used. The cheaper alternative that is based on a class II scatter matrix and a one-step W-estimate applied to it (Section 6 of the paper) will generally miss the structure. If a full class III estimate is too expensive, we recommend simply restricting the number of elemental subsets of the usual resampling algorithm to find such an estimate and using the respective 'unsecure' estimate as V_2.

The **authors** replied later, in writing, as follows.

We thank all the discussants for their insightful and generally encouraging remarks. Many of the points that were made by them have also been major concerns of ours, and we hope that our paper stimulates others to develop this topic further. The discussants have already pointed the way to many important open problems.

Rather than respond to the discussants one by one, we address their main recurring themes.

Choice of scatter and statistical variability

One of the more prominent themes in the contributions is the choice of the scatter matrices. This is certainly a major topic deserving a better understanding. A good choice for the scatter matrices, though, will probably depend on the problem at hand, e.g. whether interest lies in a mixture problem, an independent components analysis (ICA) problem or some other problem.

Much of the discussions tends to focus on the role of invariant co-ordinate selection (ICS) in detecting mixtures or clusters. In this setting, it seems natural that one should try to define one scatter matrix so that it can be viewed as a measure of *within-group* scatter. This is essentially the idea behind Dr Hennig's proposal for a *local shape* matrix. (As defined, this matrix is not affine equivariant but can be made so by replacing $\mathrm{tr}(V_{l_m})$ with $\det(V_{l_m})$ in its definition.) It is also the motivating idea behind the clustering algorithm that was proposed by Art *et al.* (1982), as well as the idea behind the scatter matrices based on downweighting large pairwise differences that were noted in the discussion of Professor Caussinus and Professor Ruiz-Gazen and in Lutz Dümbgen's oral presentation explaining the choice of scatter matrices that were used in the RANDU example.

Nevertheless, if one of the models that are considered in Section 5 holds, the results of our paper imply that the choice of the scatter matrices used in deriving the ICS co-ordinates is theoretically irrelevant for sufficiently large sample sizes. As noted by Professor Hallin and by Dr Ringrose, the main considerations are the theoretical separation of the ICS roots, $\rho_1(F), \ldots, \rho_p(F)$, and the statistical variability of the sample scatter matrices $\hat{V}_1$ and $\hat{V}_2$. If the theoretical roots are not well separated then some modest statistical variability in the scatter matrices may result in the ICS co-ordinates being poorly estimated. The theoretical ICS co-ordinates, however, do not depend on the choice of the scatter matrices, at least within the context of the theorems of Section 5. Studying the statistical variability of the ICS roots and co-ordinates is a reasonably straightforward problem, at least asymptotically. However, the more important problem of understanding the theoretical separation of the roots based on two given scatter functionals for a specific underlying model appears to be very challenging, and any results on this topic are greatly welcomed.

To illustrate these points further, consider the example that was presented by Dr Hennig. This interesting example is presented as a clustering problem but it does not fall under the mixture models that were considered in Section 5.1. Rather, it provides a nice example of an ICA model. Theorem 5 states that essentially any two scatter matrices should uncover the structure. Furthermore, as noted in the discussion after theorem 5, the independence property or symmetrization is not needed here since two of the three marginals are symmetric. The scatter matrices that were first considered by Hennig, i.e. S_n and K_n, may not be appropriate since neither one is defined at the population model owing to the t_2-distribution. (Curiously, the t_2-component is easily found and the difficulty appears to be in separating the mixture component from the uniform component.) Otherwise, any well-defined pair of scatter matrices should find the independent components for a sufficiently large sample size, even if they are not specialized to this particular problem.

Figs 11(a) and 11(b) show the results for Dr Hennig's example when Dümbgen's scatter matrix is chosen for $\hat{V}_2$ in both figures, and with the t_2 M-estimate of scatter and its symmetrized version chosen respectively as $\hat{V}_1$. From Fig. 11 the symmetrized version appears to give a slightly better recovery of the independent components, with the sample ICS roots being more widely separated in the symmetrized version, i.e. (1.26, 0.98, 0.80) for Fig. 11(a) *versus* (1.53, 0.87, 0.75) for Fig. 11(b). This suggests, as commented on by Professor Kent, and Professor Caussinus and Professor Ruiz-Gazen, that there may be advantages to symmetrization, at least for moderate sample sizes. It also seems that using a common centre for both scatter matrices may be advantageous. The ICS plots using a t_2 M-estimate of scatter and Tyler's shape

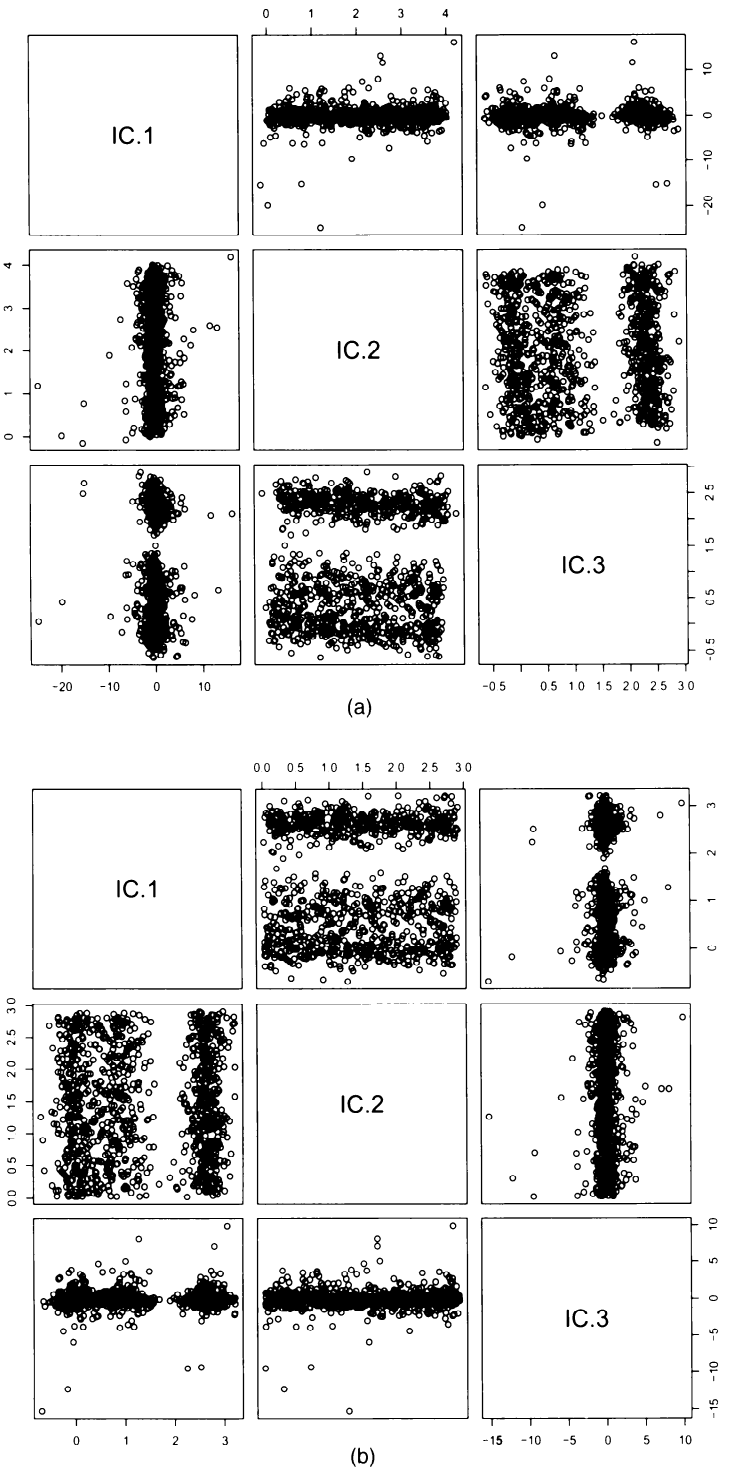

Fig. 11. ICS for Hennig's example: (a) $\hat{V}_1$, the t_2 M-estimate, and $\hat{V}_2$, Dümbgen's scatter; (b) $\hat{V}_1$, the symmetrized t_2 M-estimate, and $\hat{V}_2$, Dümbgen's scatter

matrix (which is the unsymmetrized version of Dümbgen's scatter matrix) centred at the t_2 M-estimate of location gives a plot similar to Fig. 11(b). Using a common centre avoids the additional computations that are needed in working with symmetrized data.

This example also sheds some light on Professor Kent's question regarding the distinction between larger and smaller ICS roots. Dümbgen's matrix may be viewed in a loose sense as being more 'robust' than both the t_2 M-estimate and its symmetrized version. In Fig. 11(a), though, the t_2-component is related to the largest root whereas in Fig. 11(b) it is related to the smallest root.

Statistical inference and general distributions

Our paper does not give any results on statistical inference, but rather leaves this topic open for further research. As noted by Professor Caussinus and Professor Ruiz-Gazen, there are some interesting open inferential problems when we assume a mixture model. Perhaps the most fundamental question, though, is first to determine whether ICS roots significantly differ from each other. Otherwise, the ICS method is simply exploring noise.

Some work on using two scatter matrices to test for multivariate normality can be found in Kankainen *et al.* (2007). In Wang (2008), the sample ICS roots are used to develop tests for the hypothesis that the data come from an elliptical distribution. The local power function of these tests under mixtures of elliptical distributions and under skewed elliptical distributions are also obtained.

The question that is posed by Professor Hallin regarding the optimal choice of scatter matrices for such tests again depends on the problem at hand, i.e. on the alternative model. The problem of testing the hypothesis of ellipticity against a general multivariate distribution is far more complex than testing the hypothesis of sphericity within the class of elliptical distributions. Even when considering a mean mixture of two multivariate normal distributions, we have noted in some preliminary work that one pair of scatter matrices may be more powerful than another pair for some mixtures, whereas the reverse may hold for other mixtures.

For distributions other than the models that are discussed in Section 5, the theoretical ICS co-ordinates themselves can be heavily dependent on the scatter functionals. In this case, the use of more than two scatter functionals, as pondered by Professor Dawid, may be helpful for exploring these more complex non-elliptical structures. Generating a new co-ordinate system based on the comparison of more than two scatter matrices is more problematic since in general three or more scatter functionals cannot be simultaneously diagonalized. (Note that theorem 5 states that all scatter functionals can be simultaneously diagonalized for the ICA model that is considered in the theorem.) Perhaps some approximate simultaneous diagonalization as suggested by Professor Filzmoser can be developed. Approximate simultaneous diagonalization techniques have been developed in another context in ICA; see for example Cardoso and Souloumiac (1996).

The contribution by Dr Paindaveine presents a clever application which uses the information that is contained in more than two scatter matrices, namely a graphical method for accessing how varied different ICS co-ordinate systems may be. The insight that is obtained from this can be used to diagnose whether one of the models considered in Section 5 is appropriate. The graphs that he presents suggest that a wide range of scatter matrices should be considered in practice, since some pairs of scatter matrices may give similar ICS results, whereas others may give differing results.

The hypothesis of the equality of the theoretical ICS roots is equivalent to the hypothesis $V_1 \propto V_2$. Information coming from several scatter matrices can also be used to develop alternative and perhaps more powerful tests for ellipticity by considering the hypothesis $V_1 \propto V_2 \ldots \propto V_k$. Expanding on Dr Paindaveine's idea, several scatter matrices can also be used to test whether there is any significant deviation from one of the models considered in Section 5. For such a test, rather than testing whether the scatter matrices are proportional to each other, we would be interested in testing whether the scatter matrices can be simultaneously diagonalized. These are challenging topics for future research.

High dimensional data and projection pursuit

Several discussants bring up the topic of high dimensional data, and in particular when the sample size n is small relative to the dimension p. For $n \leqslant p+1$, all affine equivariant sample scatter matrices are proportional to each other (see for example Tyler (2009)), and so the ICS method is not applicable in this case. When n is not too large relative to p, as noted by Professor Peña and Dr Viladomat, ICS is unlikely to be successful at finding underlying structures because of the statistical variability of the scatter matrices, unless the structures are extreme.

As p increases as a multiple of n, it is not clear whether ICS roots and co-ordinates will converge to anything. In an infinite dimensional space, how do we define an affine equivariant scatter operator other

than the covariance operator? In this setting, and in the setting for which n/p is not very large, we may need to relax the requirement of affine equivariance and, as suggested by Professor Kent, introduce some type of regularization.

The question of the differences between ICS and projection pursuit methods derived by using one-dimensional projection indexes is a natural one. Professor Kent, for example, questions the difference between using $\mathrm{kurt}(a'Y)$ and using $a'\mathcal{K}a/a'\Sigma a$. Professor Peña and Dr Viladomat report on some recent work comparing these two measures, which we are eager to read. They remark that the theory for ICS does not guarantee the identification of clusters when the scatter matrices of the mixture components differ (or, more precisely, when they are not proportional), whereas projections with extreme univariate kurtosis have been effective in this situation. To the best of our knowledge, the theory for projection pursuit in this case only guarantees identifying the components of a normal mixture when the covariance matrices are equal, or when the components are well separated. It is possible to show that the latter case will also hold for ICS. A special case of this has been considered by Critchley *et al.* (2007).

We do not generally recommend using $\mathcal{K}$ and Σ as the scatter matrices in ICS. For this choice, the results of ICS can be too heavily focused on just a few spurious outliers, and statistical variability of the method can be high for longer-tailed distributions, including mixture models. This pair of scatter matrices is nevertheless of interest, not only for their role in one of the earliest ICA algorithms, FOBI, but also since they can be analytically tractable and hence can help to lead to a better theoretical understanding of the method. For example, in the mixture of two multivariate normal distributions that was discussed after theorem 3, it is shown that when the mixing proportion satisfies $\alpha(1-\alpha)=\frac{1}{6}$ then $a'\mathcal{K}a/a'\Sigma a$ is constant and therefore no direction is distinguishable from any other. If we examine the formula for kurtosis for a mixture of two univariate normal distributions (see Preston (1953)), we draw the same conclusion regarding $\mathrm{kurt}(a'Y)$, i.e. it is constant. There is also a relationship between the two measures under the ICA model, namely

$$a_j'\mathcal{K}a_j/a_j'\Sigma a_j=\mathrm{kurt}(a_j'Y)$$

when $a_j'Y$ is one of the independent components; see Nordhausen *et al.* (2008).

Content and style

As observed by Dr Ringrose, the examples that are given in the paper are intended to illustrate clearly the theory given in the paper. Consequently, the structures that are found in some of the examples can also be found by using a simple principal components analysis (PCA). It is a fair question though to ask what ICS can do that cannot be done with PCA. This question can also be asked of discriminant analysis, with some of us having had the experience of consulting with researchers in applied fields who do obtain answers from PCA even though the problem is one of discriminant analysis. It is well known to researchers in multivariate analysis that one can easily construct examples where PCA will fail to uncover the group differences, particularly when the means vary in the direction of the smallest principal component direction relative to the within-group covariance matrix. Such cases, when group identification is unknown, will result in a similar difference between ICS and PCA.

Other examples can be found in the vast ICA literature, where one of the main motivations for its development is that PCA often fails to find important structures in a multivariate data set. If one applied PCA to our example 2, one would not find the underlying structure regardless of the scatter matrix that is used for the PCA. In practice, it is possible for the ICS roots to be significantly different, yet no obvious structure, groups or outliers may be visible in the plots of the ICS co-ordinates. The theory assures us though that the underlying distribution is more complicated than an elliptical distribution, and so a deeper understanding of the data is needed, as opposed to a simple location–scatter summary, and a closer examination of the ICS co-ordinates may be enlightening. Such examples, though, do not make good initial illustrations.

Whether or not a moratorium should be placed on Fisher's iris data is a matter of debate. They can be useful for illustration while taking up minimal space in a paper or minimal time in a presentation since one does not need to explain the data set in detail. Also, if a method does not perform well on the iris data, then the theory may be suspect. Other fields have their pet data sets for illustrating methods and theory (the iris data are a pet rather than a toy), such as the famous Lena image in computer vision.

We appreciate Professor Stone's co-ordinate-free formulation to the ICS variates. The co-ordinate-free approach to multivariate statistics certainly offers a theoretically elegant and concise view of the topic. Professor Kent's comments also hint at the co-ordinate-free approach in his mention of dual metrics. The statement that two scatter matrices can always be simultaneously diagonalized can be stated more elegantly

in a co-ordinate-free manner by simply noting that for any two inner products on a finite dimensional vector space there is a basis (the ICS basis) which is orthogonal (but not necessarily orthonormal) relative to both inner products. Alternatively, rather than present the usual technical gruel of relating the spectral value decomposition of the symmetric matrix $V_1^{-1/2} V_2 V_1^{-1/2}$ to the eigenvalues and eigenvectors of $V_1^{-1} V_2$ we could note that ICS simply corresponds to the usual spectral value decomposition of the symmetric operator $V_1^{-1} V_2$, where symmetry is with respect to the inner product $(x, y) = x^{\mathrm{T}} V_1^{-1} y$. Some of us have mentioned these more abstract concepts in presentations and at times receive the query why confuse the audience with the abstraction? So, a common consideration in presenting results is the intended audience, which for this paper is those who are interested in general multivariate methodology. The interpretation of ICS as a PCA on standardized data may be particularly appealing to practitioners.

A co-ordinate-free approach can be beneficial in any attempt to generalize ICS to infinite dimensional Hilbert spaces. Here the concept of a mutual orthogonal basis relative to two different inner products still holds, as well as the spectral value decomposition (or the Karhunen–Loève decomposition) for symmetric operators. The covariance operator can also be defined within a co-ordinate-free format; see for example

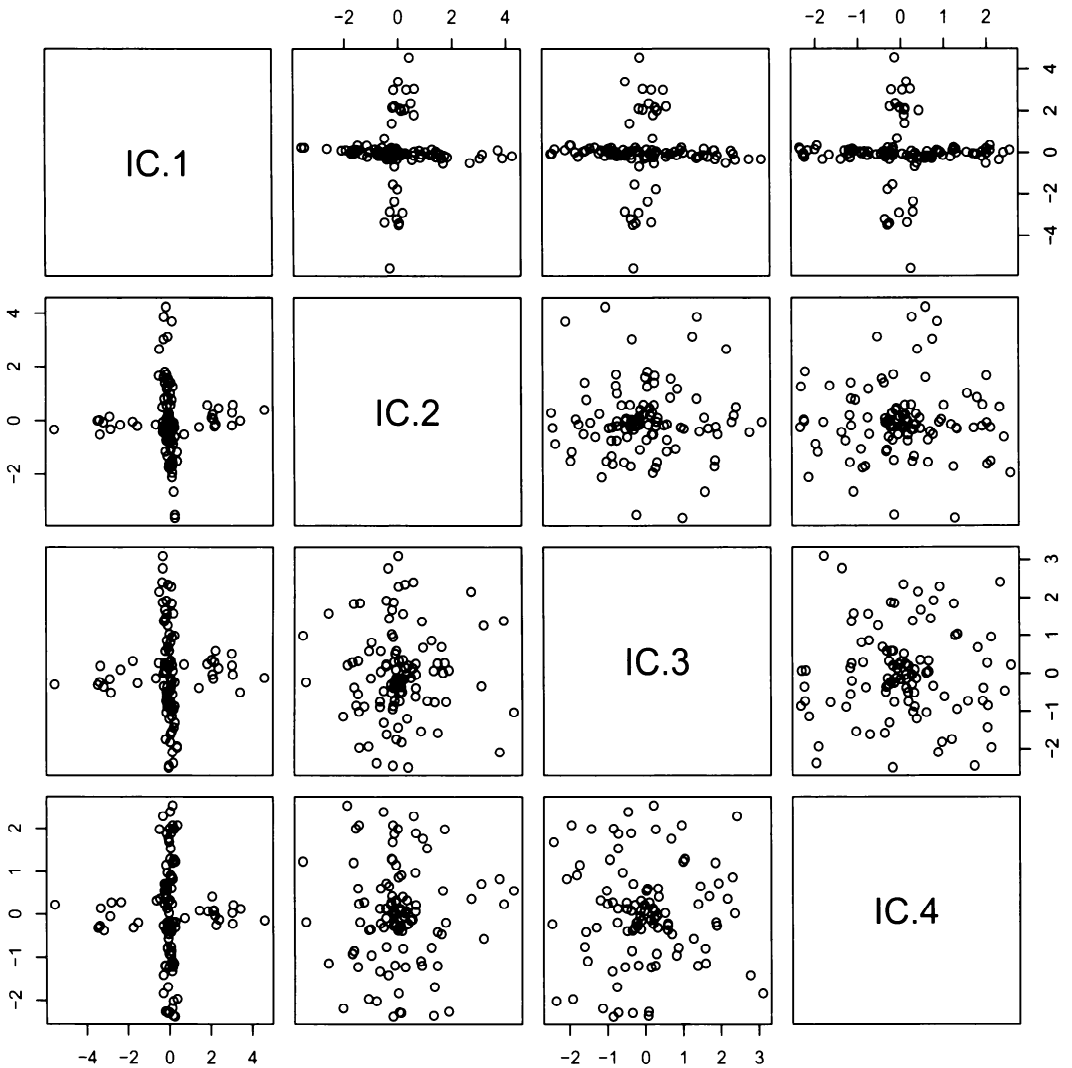

Fig. 12. ICS for Stahel–Mächler's example: $\hat{V}_1$, the t_1 M-estimate, and $\hat{V}_2$, the sample covariance

Eaton (1983). It is not clear, however, how other scatter functionals, whether in finite or infinite dimensional space, can be formulated in a co-ordinate-free manner.

Robustness

Several of the discussants noted the role of robustness in ICS. Dr Croux gives a very perceptive discussion on the key points of our paper and in particular notes how ICS is a natural outgrowth of contemplating problems in robust statistics. Researchers within the robustness community are familiar with working with competing functionals (and estimates) measuring (estimating), under general assumptions, presumably the same population parameter. Robust statistics is often focused on automatically accommodating outliers so that they do not influence the interpretation of the majority of the data. It is then natural to ask in the multivariate setting whether a location–scatter summary is reasonable even for the majority of the data, e.g. a 30–30–40 mixture.

The type of outliers that usually cause 'breakdown' for many statistics are not simply spurious outliers but rather outliers that have a pattern of their own, with the most extreme case being point mass contamination. One can argue that such data structures are confounded with mixture models, and as an alternative one could try to identify such a pattern while accommodating spurious outliers. Statistics, other than high breakdown point statistics, tend to blur the majority structure with any outlier structure. However, as shown in the modified wood gravity data example, when two such scatter statistics are used together in ICA, the separate patterns may be more apparent.

A similar phenomenon occurs for the example that was given by Stahel and Mächler, which was originally used in Hampel *et al.* (1986) to illustrate the concept of *breakdown at the edge*. In this example, the majority of the data lie close to some subspace. As correctly noted in their contribution scatter statistics which do not have breakdown points near $\frac{1}{2}$ fail to pick up the near singularity of the majority of the data. This does not imply though that ICS based on two lower breakdown point statistics will also fail to detect this pattern. Although this example does not correspond to one of the mixture models or ICA models that were considered in Section 5, it turns out that the ICS co-ordinates do not depend theoretically on the two scatter matrices used.

To see this, we first note that because of the invariance of ICS it is sufficient to consider the distribution of G_0. The distribution of G_0 is invariant under transformations of the form QX, i.e., if X has distribution G_0, then so does QX when Q is block diagonal with blocks $q_{11} = \pm 1$ and Q_{22} being a 3×3 orthogonal matrix. Consequently, any affine equivariant scatter matrix V at G_0 must be block diagonal with elements v_{11} and $V_{22} = v_{22} I_3$, where I_3 is the 3×3 identify matrix. Consequently, $V_1^{-1} V_2$ has the same block diagonal form, and so either the first or last ICS co-ordinate will correspond to the first variate in G_0 and the other three co-ordinates will correspond to some rotation of the last three co-ordinates in G_0 (except for the idiosyncratic case when V_1 and V_2 are theoretically proportional to each other). Again the question of choice depends on the separation of the theoretical roots along with the statistical variability of the roots. For sufficiently large sample size, the pattern should be detected for any two choices of scatter and neither one needs to have a high breakdown point. Fig. 12 illustrates this for a sample of size $n = 100$ from G_0 using the sample covariance matrix and the Cauchy M-estimate.

Other remarks

We have not been able to respond to all of the comments in the contributions in detail. Dr Ringrose and Professor Filzmoser bring up the topic of biplots for the ICS co-ordinates, and so we wish to note briefly that such biplots have been considered by Caussinus *et al.* (2003) in the context of generalized PCA. Professor Kent raises the question of possible extensions for third moments. Here, we note some recent work on this topic by Nordhausen *et al.* (2009a).

Again we thank all the discussants for their contributions, and we hope that our responses are somewhat enlightening. Overall, though, there is still much work to do and new methodologies are needed to understand better the nature of multivariate data, especially when we move away from the comfort of elliptical distributions.

References in the discussion

Art, D., Gnanadesikan, R. and Kettenring, J. R. (1982) Data-based metrics for cluster analysis. *Util. Math.* A, **21**, 75–99.

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

Cardoso, J.-F. and Souloumiac, A. (1996) Jacobi angles for simultaneous diagonalization. *SIAM J. Math. Anal. Appl.*, **17**, 161–164.
Caussinus, H., Fekri, M., Hakam, S. and Ruiz-Gazen, A. (2003a) A monitoring display of multivariate outliers. *Computnl Statist. Data Anal.*, **44**, 237–252.
Caussinus, H., Hakam, S. and Ruiz-Gazen, A. (2003b) Projections révélatrices contrôlées, groupements et structures diverses. *Rev. Statist. Appl.*, **51**, 37–58.
Caussinus, H. and Ruiz-Gazen, A. (2007) Classification and generalized principal component analysis. In *Selected Contributions in Data Analysis and Classification* (eds P. Brito, P. Bertrand, G. Cucumel and F. De Carvalho), pp. 539–548. Berlin: Springer.
Critchley, F., Pires, A. and Amado, C. (2007) Principal axis analysis. Unpublished manuscript.
Eaton, M. L. (1983) *Multivariate Statistics: a Vector Space Approach*. New York: Wiley.
Filzmoser, P., Maronna, R. and Werner, M. (2008) Outlier identification in high dimensions. *Computnl Statist. Data Anal.*, **52**, 1694–1711.
Genton, M. G. (2004) *Skew-elliptical Distributions and Their Applications: a Journey Beyond Normality*. Boca Raton: Chapman and Hall–CRC.
Hallin, M., Oja, H. and Paindaveine, D. (2006) Semiparametrically efficient rank-based inference for shape: II, optimal *R*-estimation of shape. *Ann. Statist.*, **34**, 2757–2789.
Hallin, H. and Paindaveine, D. (2006) Semiparametrically efficient rank-based inference for shape: I, optimal rank-based tests for sphericity. *Ann. Statist.*, **34**, 2707–2756.
Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics: the Approach based on Influence Functions*. New York: Wiley.
John, S. (1972) The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika*, **59**, 169–174.
Kankainen, A., Taskinen, S. and Oja, H. (2007) Tests of multinormality based on location vectors and scatter matrices. *Statist. Meth. Appl.*, **16**, 357–379.
Mauchly, J. W. (1940) Test for sphericity of a normal *n*-variate distribution. *Ann. Math. Statist.*, **11**, 204–209.
Nordhausen, K., Oja, H. and Ollila, E. (2009a) Multivariate models and the first four moments. In *Festschrift for Thomas P. Hettmansperger* (eds D. R. Hunter, J. L. Rosenberger and D. Richards). To be published.
Nordhausen, K., Oja, H. and Paindaveine, D. (2009b) Signed-rank tests for location in the symmetric independent component model. *J. Multiv. Anal.*, **100**, 821–834.
Nordhausen, K., Oja, H. and Tyler, D. E. (2008) Tools for exploring multivariate data via ICS/ICA. In *R Package, Version 1.1-1* (Available from `http://cran.r-project.org`.)
Peña, D. and Prieto, F. J. (2001a) Cluster identification using projections. *J. Am. Statist. Ass.*, **96**, 1433–1445.
Peña, D. and Prieto, F. J. (2001b) Robust covariance matrix estimation and multivariate outlier detection (with discussion). *Technometrics*, **43**, 286–310.
Peña, D., Prieto, F. J. and Viladomat, J. (2008) Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. To be published.
Preston, E. J. (1953) A graphical method for the analysis of statistical distributions into two normal components. *Biometrika*, **40**, 460–464.
Stone, M. (2008) Going to the pictures: eigenvector as fixed point. *Research Report*. University College London, London. (Available from `www.ucl.stats.uk/research/reports/psfiles/rr299.pdf`.)
Sun, J. (1991) Significance levels in exploratory projection pursuit. *Biometrika*, **78**, 759–769.
Tyler, D. E. (1982) Radial estimates and the test for sphericity. *Biometrika*, **69**, 429–436.
Tyler, D. E. (1987) A distribution-free M-estimator of multivariate scatter. *Ann. Statist.*, **15**, 234–251.
Tyler, D. E. (2009) A note regarding multivariate location and scatter statistics for sparse data sets. *Technical Report*. Department of Statistics, State University of New Jersey, Piscataway.
Wang, J. (2008) Some properties of robust statistics under asymmetric models. *PhD Thesis*. State University of New Jersey, Piscataway.