Invariant coordinate selection for multivariate data analysis - the package ICS

Klaus Nordhausen[1] Hannu Oja[1] David E. Tyler[2]

[1]Tampere School of Public Health
University of Tampere

[2]Department of Statistics
Rutgers - The State University of New Jersey

Outline

Transformations I

Let $X_1, X_2, \ldots, X_n$ be independent p-variate observations and write $X = (X_1 X_2 \ldots X_n)$ for the corresponding $p \times n$ *data matrix*.

- Affine transformation

$$X \quad \rightarrow \quad AX + b1',$$

  where A is a full-rank $p \times p$ matrix, b a p-vector and 1 a n-vector full of ones.

- Orthogonal transformation

$$X \rightarrow UX$$

  with $U'U = UU' = I$.

Transformations II

- Sign-change transformation

$$X \to JX$$

where J is a $p \times p$ diagonal matrix with diagonal elements ± 1.

- Permutation

$$X \to PX$$

where P is a $p \times p$ permutation matrix.

Location and scatter statistics

A p-vector valued statistic $T = T(X)$ is called a location statistic if it is affine equivariant, that is,

$$T(AX + b1') = AT(X) + b$$

for all full-rank $p \times p$-matrices A and for all p-vectors b.

A $p \times p$ matrix $S = S(X)$ is a scatter statistic if it is affine equivariant in the sense that

$$S(AX + b1') = AS(X)A'$$

for all full-rank $p \times p$-matrices A and for all p-vectors b.

Special scatter statistics

A scatter statistic with respect to the origin is affine equivariant in the sense that

$$S(AXJ) = AS(X)A'$$

for all full-rank $p \times p$-matrices A and for all $n \times n$ sign change matrices J.

A symmetrized scatter statistic version of a scatter statistic S is defined as

$$S_{sym}(X) := S(X_{sym}),$$

where X_{sym} is the matrix of all pairwise differences of the original observation vectors.
A shape matrix is only affine equivariant in the sense that

$$S(AX + b1') \propto AS(X)A'.$$

Independence property

A scatter functional S has the independence property if it is a diagonal matrix for all random vectors with independent margins.

Note that in general scatter statistics do not have the independence property. Only the covariance matrix, the matrix of fourth moments and symmetrized scatter matrices have this property.
If, however, X has independent and at least $p - 1$ symmetric components all scatter matrices will be diagonal matrices.

Examples of scatter matrices

The most common location and scatter statistics are the vector of means and the regular covariance matrix COV.

A so called 1 step M-estimator one is for example the matrix of fourth moments.

$$COV_4(X) = \frac{1}{p+2}\text{ave}[||X_i - \bar{X}||^2_{COV}(X_i - \bar{X})(X_i - \bar{X})']$$

A regular M-estimator is for instance Tyler's shape matrix .

$$S_{Tyl}(X) = p\ \text{ave}\left[\frac{(X_i - T(X))(X_i - T(X))'}{||X_i - T(X)||^2_{S_{Tyl}}}\right]$$

The symmetrized version of Tyler's shape matrix is known as Dümbgens's shape matrix.

Scatter matrices in R

R offers a lot functions for estimating different scatter matrices.

A most likely not complete list:

covRobust:	`cov.nnve`
ICS:	`covOrigin`, `cov4`, `covAxis`, `tM`
ICSNP:	`tyler.shape`, `duembgen.shape`, `HR.Mest`, `HP1.shape`
MASS:	`cov.rob`, `cov.trob`
robustbase:	`covMcd`, `covOGK`
rrcov:	`covMcd`, `covMest`, `covOgk`

Two scatter matrices for ICS

Tyler et al. (2008) showed that two different scatter matrices $S_1 = S_1(X)$ and $S_2 = S_2(X)$ can be used to find an invariant coordinate system as follows:
Starting with S_1 and S_2, define a $p \times p$ transformation matrix $B = B(X)$ and a diagonal matrix $D = D(X)$ by

$$S_2^{-1} S_1 B' = B' D$$

that is, B gives the eigenvectors of $S_2^{-1} S_1$. The following result can then be shown to hold.

The transformation $X \to Z = B(X)X$ yields an invariant coordinate system in the sense that

$$B(AX)(AX) = JB(X)X$$

for some $p \times p$ sign change matrix J.

On the choice of S_1 and S_2

As shown previously there are a lot of possibilities for S_1 and S_2 to choose from. Since all the scatter matrices have different properties, these can yield different invariant coordinate systems.
Unfortunately, there are so far no theoretic results about the optimal choice. This is still an open research question.

Some comments are however already possible:

- For two given scatter matrices, the order has no effect
- Depending on the application in mind the scatter matrices should fulfill some further conditions.
- Practise showed so far that for most data sets different choices yield only minor differences.

Implementation in R

The two scatter transformation for ICS is implemented in R in the package ICS. The main function `ics` can take the name of two functions for S_1 and S_2 or two in advance computed scatter matrices and returns and S4-object. The package ICS offers then furthermore several functions to work with an `ics` object and offers also several scatter matrices and two tests of multinormality.

The function `ics` has options for different standardization methods for B and D.

```
ics(X, S1 = cov, S2 = cov4, S1args = list(),
    S2args = list(), stdB = "Z", stdKurt = TRUE,
    na.action = na.fail)
```

Multivariate nonparametric methods which are meaningful in the context of ICS are implemented in the R package ICSNP.

What is an ICS good for?

So what is an ICS good for? In the following applications an ICS can be of use:

- Descriptive statistics
- Finding outliers, structure and clustering
- Dimension reduction
- Independent component analysis
- Multivariate nonparametrics

The following slides will by means of examples motivate some of the above applications.

Descriptive statistics

The way how the transformation matrix B is obtained can also be seen as taking the ratio of two different scale measures. Therefore the elements of D can be seen as measures of Kurtosis.
Using this interpretation and given the choice of S_1, S_2, T_1 and T_2, one obtains immediately:

- The location: $T_1(X)$
- The scatter: $S_1(X)$
- Measure of skewness: $T_2(Z) - T_1(Z)$
- Measure of Kurtosis: $S_2(Z)$

Note that when S_1 is the regular covariance matrix and S_2 the matrix of fourth moments, the elements of D are based on classical moments

The last two measures can then be used to construct tests for multinormality or ellipticity (see for example Kankainen et al. 2007).

Finding outliers, structure and clusters

In general, that the coordinates are ordered according to their kurtosis is a useful feature. One can assume that interesting directions are the ones with extreme measures of kurtosis.

Outliers for example are usually shown in the first coordinate.

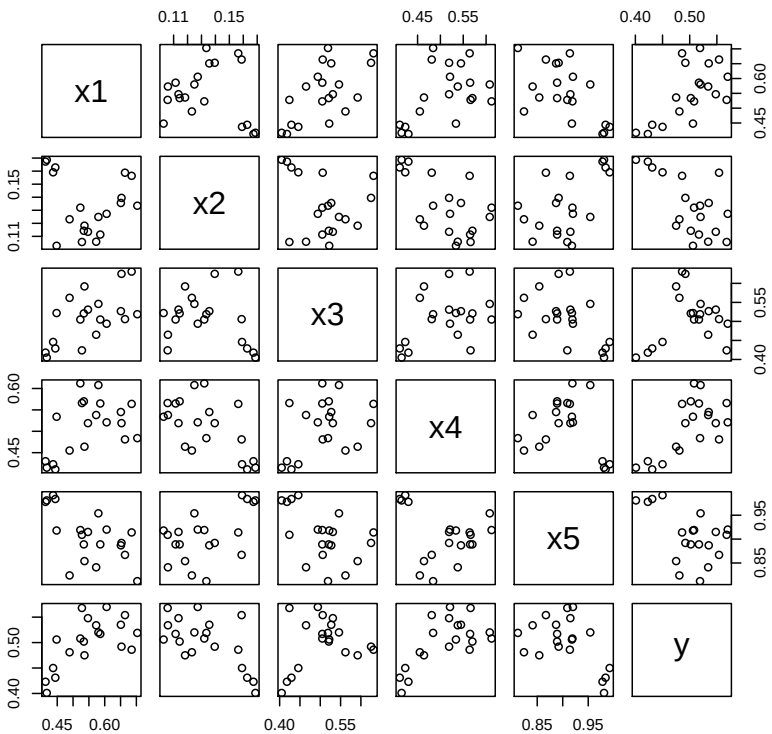
If the original data is coming from an elliptical distribution, S_1 and S_2 measure the same population quantity and therefore the values of D in that case should approximately be all the same. For non-elliptical distributions however they measure different quantities and therefore the coordinates can reveal "hidden" structures.

In the case of X coming from an mixture of elliptical distributions, the first or the last coordinate corresponds to Fisher's linear discriminant function (Without knowing the class labels!)

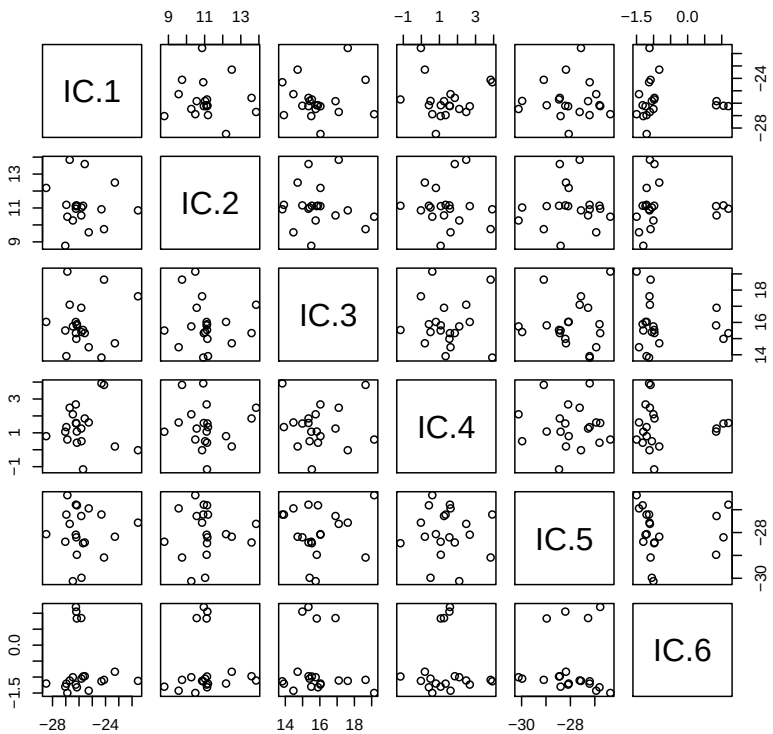
Finding the outliers

The modified wood gravity data is a classical data set for outlier detection methods. It has 6 measurements on 20 observations containing four outliers. Here S_1 is a M-estimator based on t_1 and S_2 based on t_2.

Original data

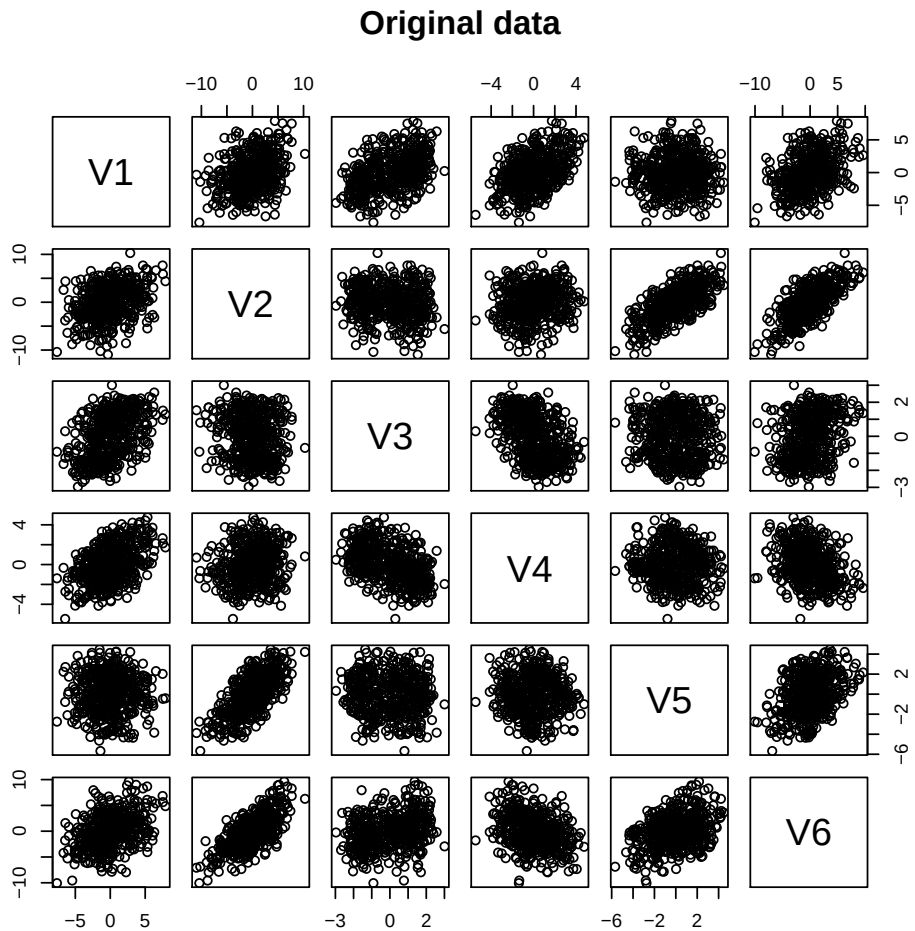


Invariant coordinates

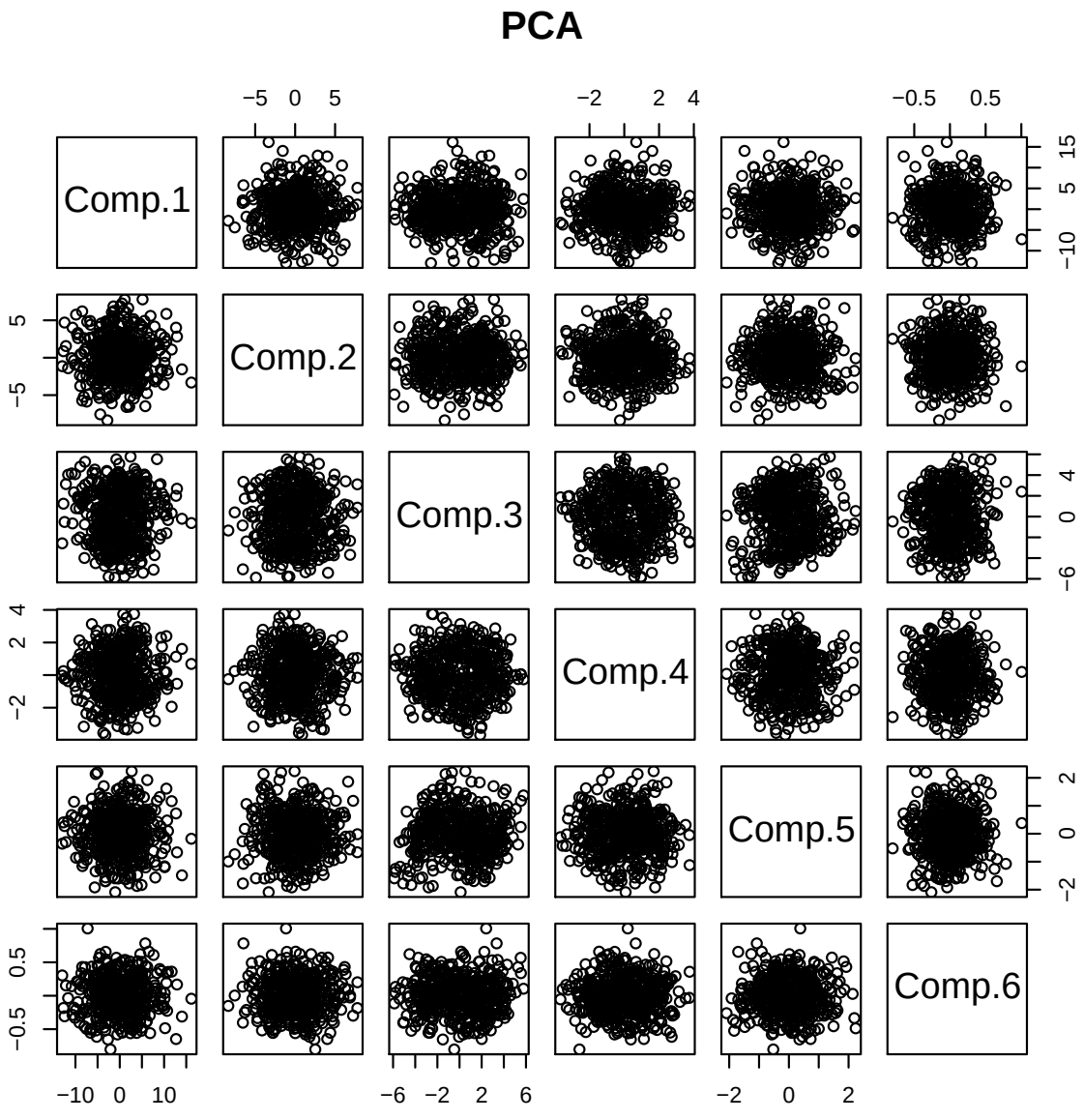


Finding the structure I

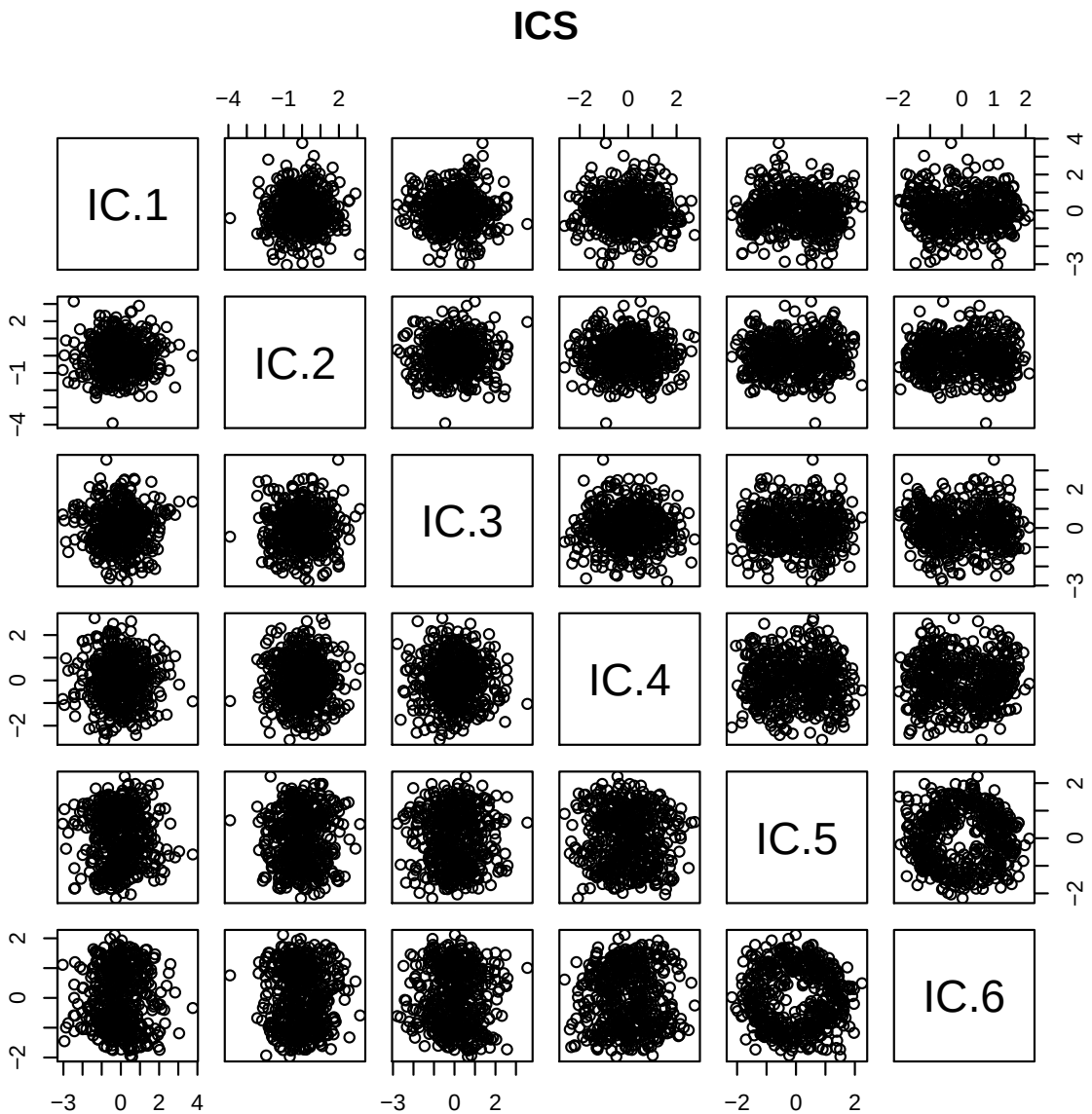
The following data set is simulated and has 400 observations for 6 variables. It looks like an elliptical data set.



Finding the structure II

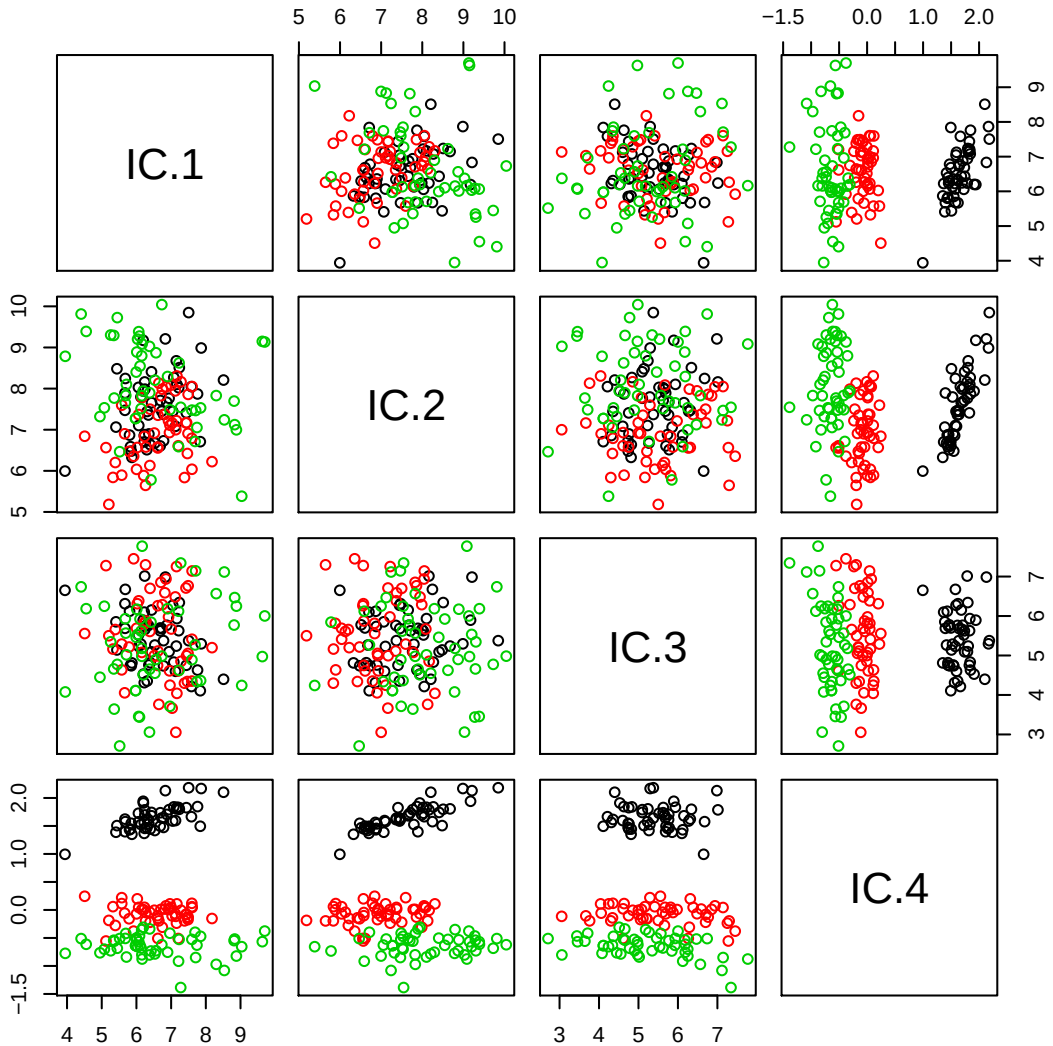


Finding the structure III



Clustering and dimension reduction

To illustrate clustering via an ICS we use the Iris data set.
The different species are colored differently and we can see, that the last component is enough for doing the clustering.



Independent component analysis

Independent component analysis (ICA) is a method often applied in signal processing or medical image analysis.

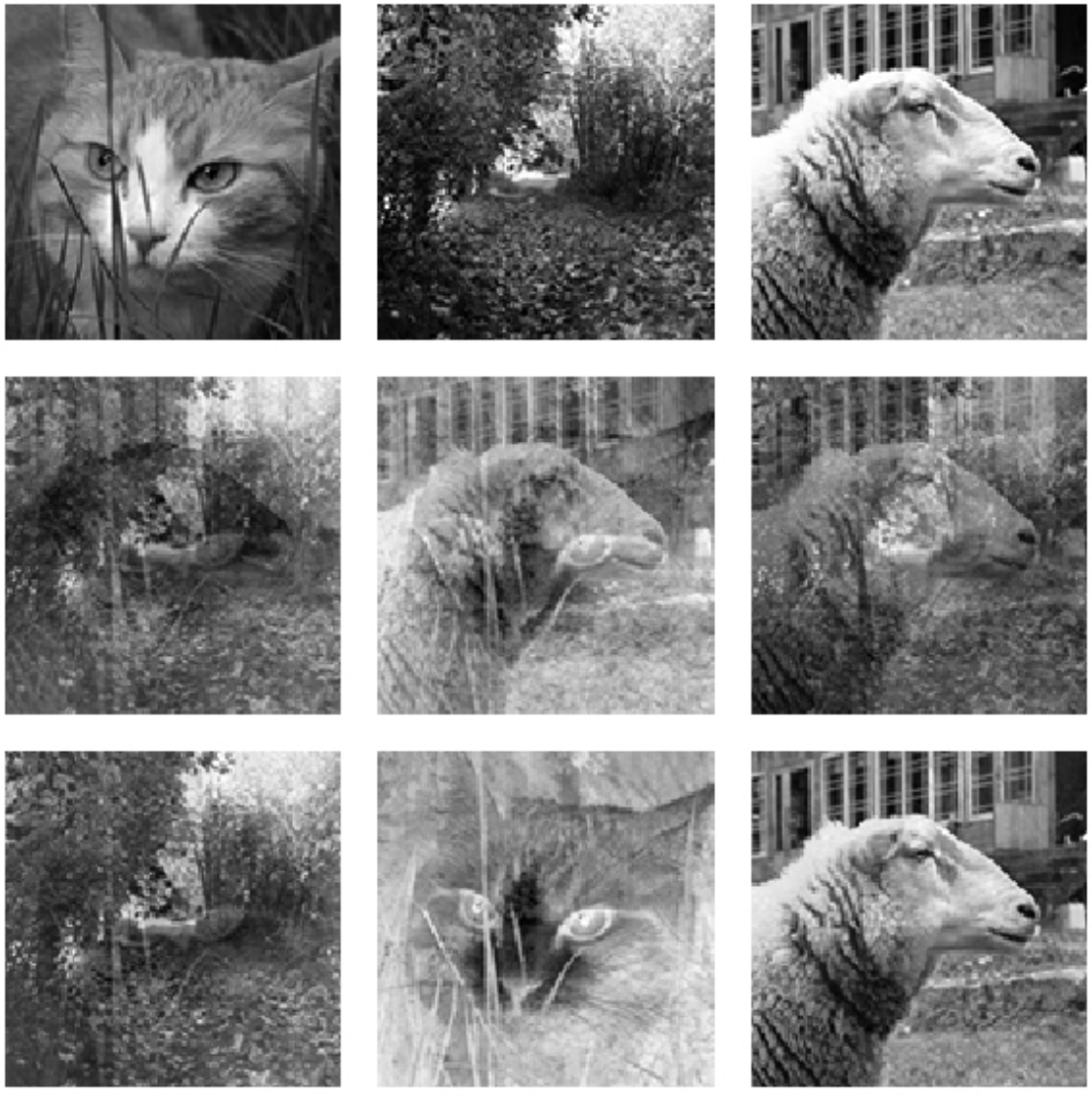
The most basic ICA model is of the form

$$X_i = AZ_i, \quad i = 1, \ldots, n$$

where Z_i has independent components and A is a full rank mixing matrix. The goal is to find an unmixing matrix B to recover the independent components.

Oja et al. (2006) showed that the two scatter matrix transformation recovers in such a model the independent components if S_1 and S_2 have the independence property and the independent components have different kurtosis values.
Using two robust scatters hence provides a robust ICA method.

Independent component analysis



Multivariate nonparametrics

A lot of multivariate nonparametric methods are not affine equivariant by nature. Applying those methods in an invariant coordinate system is therefore an important improvement.
The method introduced here is easier to apply than the so called transformation re-transformation technique of Chakraborty and Chaudhuri (1996) and has also further properties which can be used in the analysis.

Marginal nonparametric methods

Puri and Sen describe very detailed how to use marginal signs and ranks in multivariate nonparametrics.
However the tests based on these are not invariant under affine transformations, i.e. the test decision will be based on the coordinate system used.
Such tests can now for example be made affine invariant by performing the test in the invariant coordinate system.
For testing purposes the coordinate system should be constructed under the H_0. For example in the one sample location test when testing for the origin the scatter functionals used should be taken wrt to the origin and furthermore should be permutation invariant.
Therefore for the scatters should hold

$$S_k(AXPJ) = AS_k(X)A', \quad \forall\ A, P, J \text{ and } k = 1, 2.$$

Why invariance is important

We simulate 60 observations from a $N_4((0, 0, 0, 0.48)', I_4)$ distribution and then rotate the data with random matrix. The test is whether the origin is the center of symmetry.

Test on original data:

```
        Marginal One Sample Normal Scores Test
data:  Y
T = 9.653, df = 4, p-value = 0.04669
alternative hypothesis: true location is not equal
to c(0,0,0,0)
```

Test on transformed data:

```
        Marginal One Sample Normal Scores Test
data:  (Y %*% t(A))
T = 9.387, df = 4, p-value = 0.05212
alternative hypothesis: true location is not equal
to c(0,0,0,0)
```

Why invariance is important II

Now the same using an invariant coordinate system.

Test on ICS based on original data:

```
        Marginal One Sample Normal Scores Test
data:  Z.Y
T = 9.737, df = 4, p-value = 0.04511
alternative hypothesis: true location is not equal
to c(0,0,0,0)
```

Test based on ICS based on transformed data:

```
        Marginal One Sample Normal Scores Test
data:  Z.Y.trans
T = 9.737, df = 4, p-value = 0.04511
alternative hypothesis: true location is not equal
to c(0,0,0,0)
```

Key references I

- Tyler, D. E., Critchley, F., Dümbgen, L. and Oja, H. (2008). Exploring multivariate data via multiple scatter matrices. Conditionally accepted.
- Oja, H., Sirkiä, S. and Eriksson, J. (2006). Scatter matrices and indepedent component analysis. *Austrian Journal of Statistics*, 19, 175–189.
- Nordhausen, K. and Oja, H. and Tyler, D. E. (2008). Two scatter matrices for multivariate data analysis: The package ICS. Submitted.
- Nordhausen, K. and Oja, H. and Tyler, D. E. (2006). On the efficiency of invariant multivariate sign and rank test. *Festschrift for Tarmo Pukkila on his 60th birthday*, 217–231.
- Nordhausen, K., Oja, H. and Paindaveine, D. (2008). Rank-based location tests in the independent component model. Conditionally accepted.

Key references II

- Chakraborty, B. and Chaudhuri, P. (1996). On a transformation retransformation technique for constructing affine equivariant multivariate median. *Proceedings of American Mathematical Society*, 124, 1529–1537.
- Kankainen, A., Taskinen, S. and Oja, H. (2007). Tests of multinormality based on location vectors and scatter matrices. *Statistical Methods & Applications*, 16, 357–379.
- Puri, M. L. and Sen, P. K. (1971). *Nonparametric methods in multivariate analysis*. New York, Wiley & Sons.
- Tyler, D.E. (1987). A distribution-free *M*-estimator of multivariate scatter. *The Annals of Statistics*, 15, 234–251.