Supervised invariant coordinate selection

Hannu Oja (with several coauthors)

20.10. 2010, Aalto University

The plan

- Introduction
- Dimension reduction: PCA, ICA, SIR
- Location and scatter functionals
- Invariant coordinate selection (ICS)
- Supervised location and scatter functionals
- Supervised invariant coordinate selection (SICS)
- Examples, some asymptotics

Main references

Nordhausen, K., Oja, H. and Tyler, D.E. (2008). Tools for exploring multivariate data. The package ICS. *Journal of Statistical Software* 28(6), 1-31.

Tyler, D., Critchley, F., Dümbgen, L. and Oja, H. (2009), Invariant coordinate selection. *Journal of Royal Statistical Society B*, 71, 549-592.

Nordhausen, K., Oja H. and Ollila, E. (2010). Multivariate models and first four moments. *Festschrift for T.P. Hettmansperger*, to appear.

Ilmonen, P., Nevalainen, J. and Oja, H. (2010). Characteristics of multivariate distributions and the invariant coordinate system. *Statistics & Probability Letters*, to appear.

Ilmonen, P., Serfling, R., and Oja, H. (2010). Invariant coordinate selection (ICS) functionals. Submitted.

Liski, E., Nordhausen, K., and Oja, H. (2010). Supervised invariant coordinate selection. Submitted.

Introduction

- Let $\mathbf{x}$ be a p-variate random variable with cumulative distribution F_x.
  We consider multivariate nonparametric/semiparametric models
  with few parameters of interest.

- Example 1: Dimension reduction.
  Find a projection matrix $\mathbf{P}$ such that you do not loose information
  if you transform $\mathbf{x} \to \mathbf{z} = \mathbf{P}\mathbf{x}$:
  (i) $\mathbf{x}|\mathbf{P}\mathbf{x}$ is not "interesting" (unsupervised)
  (ii) $\mathbf{y} \perp\!\!\!\perp (\mathbf{I}_p - \mathbf{P})\mathbf{x} \mid \mathbf{P}\mathbf{x}$ (supervised)

- Example 2: Independent components problem.

$$\mathbf{x} = \mathbf{A}\mathbf{z},$$

where $\mathbf{z}$ is a p-vector with independent components. This is a semiparametric model; note that parameter $\mathbf{A}$ is not well-defined.

Dimension reduction

- The dimension of $\mathbf{x}$ is reduced using a $k \times p$ matrix $\mathbf{B}$.

  Then

  $$\mathbf{x} \;\rightarrow\; \mathbf{z} = \mathbf{B}\mathbf{x}$$

  or

  $$\mathbf{x} \;\rightarrow\; \mathbf{z} = \mathbf{P_B}\mathbf{x} \quad \text{where } \mathbf{P_B} = \mathbf{B}'(\mathbf{B}\mathbf{B}')^{-1}\mathbf{B}.$$

- The idea is that $k << p$ and that “no information is lost” in the transformation.

- Dimension reduction methods (unsupervised and supervised):

  PCA, ICA, SIR, SAVE, etc.

PCA, ICA, SIR

- Assume that $E(\mathbf{x}) = \mathbf{0}$. In PCA, one then finds the $p \times p$ transformation matrix $\mathbf{\Gamma}$ such that

$$\mathbf{\Gamma}\mathbf{\Gamma}' = \mathbf{I}_p \quad \text{and} \quad \mathbf{\Gamma}E(\mathbf{x}\mathbf{x}')\mathbf{\Gamma}' = \mathbf{\Lambda}$$

  where $\mathbf{\Lambda}$ is a diagonal matrix (with diagonal elements in a decreasing order). Decompose $\mathbf{\Gamma} = (\mathbf{\Gamma}_1', \mathbf{\Gamma}_2')'$ and transform $\mathbf{z} = \mathbf{\Gamma}_1\mathbf{x}$.

- In the independent component analysis (ICA), FOBI finds transformation matrix $\mathbf{\Gamma}$ such that

$$\mathbf{\Gamma}E(\mathbf{x}\mathbf{x}')\mathbf{\Gamma}' = \mathbf{I}_p \quad \text{and} \quad \mathbf{\Gamma}E(\mathbf{x}\mathbf{x}'E(\mathbf{x}\mathbf{x}')\mathbf{x}\mathbf{x}')\mathbf{\Gamma}' = \mathbf{\Lambda}$$

  where the diagonal elements $\mathbf{\Lambda}$ are given in a specified order.

- The sliced inverse regression (SIR) uses a dependent variable $\mathbf{y}$, and finds finds a transformation matrix $\mathbf{\Gamma}$ which satisfies

$$\mathbf{\Gamma}E(\mathbf{x}\mathbf{x}')\mathbf{\Gamma}' = \mathbf{I}_p \quad \text{and} \quad \mathbf{\Gamma}E(E(\mathbf{x}|\mathbf{y})E(\mathbf{x}|\mathbf{y})')\mathbf{\Gamma}' = \mathbf{\Lambda}$$

  where the diagonal elements $\mathbf{\Lambda}$ are given in a specified order.

Location and scatter functionals

- A **location vector** $\mathbf{T}(F)$ is a p-vector valued functional which is affine equivariant in the sense that

$$\mathbf{T}(F_{\mathbf{Ax+b}}) = \mathbf{AT}(F_{\mathbf{x}}) + \mathbf{b}$$

for all nonsingular $\mathbf{A}$ and vector $\mathbf{b}$.

- A **scatter matrix** $\mathbf{S}(F)$ is a $p \times p$ matrix valued functional which is PDS and affine equivariant in the sense that

$$\mathbf{S}(F_{\mathbf{Ax+b}}) = \mathbf{AS}(F_{\mathbf{x}})\mathbf{A}'$$

for all nonsingular $\mathbf{A}$ and vector $\mathbf{b}$.

- Examples: Mean vector, covariance matrix, M-functionals, S-functionals, and so on.
- A scatter matrix functional $\mathbf{S}(F)$ has the **independent property** if

$\mathbf{x}$ has independent components $\Rightarrow \mathbf{S}(F_{\mathbf{x}})$ is a diagonal matrix.

Invariant coordinate selection (ICS)

- Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two different scatter functionals.
- Define transformation matrix functional $\mathbf{\Gamma} = \mathbf{\Gamma}(F)$ (and an auxiliary diagonal matrix functional $\mathbf{\Lambda} = \mathbf{\Lambda}(F)$) as a solution of

$$\mathbf{\Gamma}\mathbf{S}_1\mathbf{\Gamma}' = \mathbf{I}_p \quad \text{and} \quad \mathbf{\Gamma}\mathbf{S}_2\mathbf{\Gamma}' = \mathbf{\Lambda}$$

  where the elements of $\mathbf{\Lambda}$ are in a prespecified order.
- $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ give the eigenvectors and eigenvalues of $\mathbf{S}_1^{-1}\mathbf{S}_2$. If the eigenvalues are distinct then the eigenvectors are uniquely defined up to their signs.
- Invariant coordinate system (ICS): If the eigenvalues in $\mathbf{\Lambda}$ are distinct, then

$$\mathbf{\Gamma}(F_{\mathbf{A}\mathbf{x}})\mathbf{A}\mathbf{x} = \mathbf{\Gamma}(F_{\mathbf{x}})\mathbf{x}, \quad \text{for all nonsingular } \mathbf{A}.$$

- If $\mathbf{S}_1$ and $\mathbf{S}_2$ both have the independence property then $\mathbf{\Gamma}\mathbf{x}$ solves the ICA problem.

Figure 1: *Iris data; original variables.*

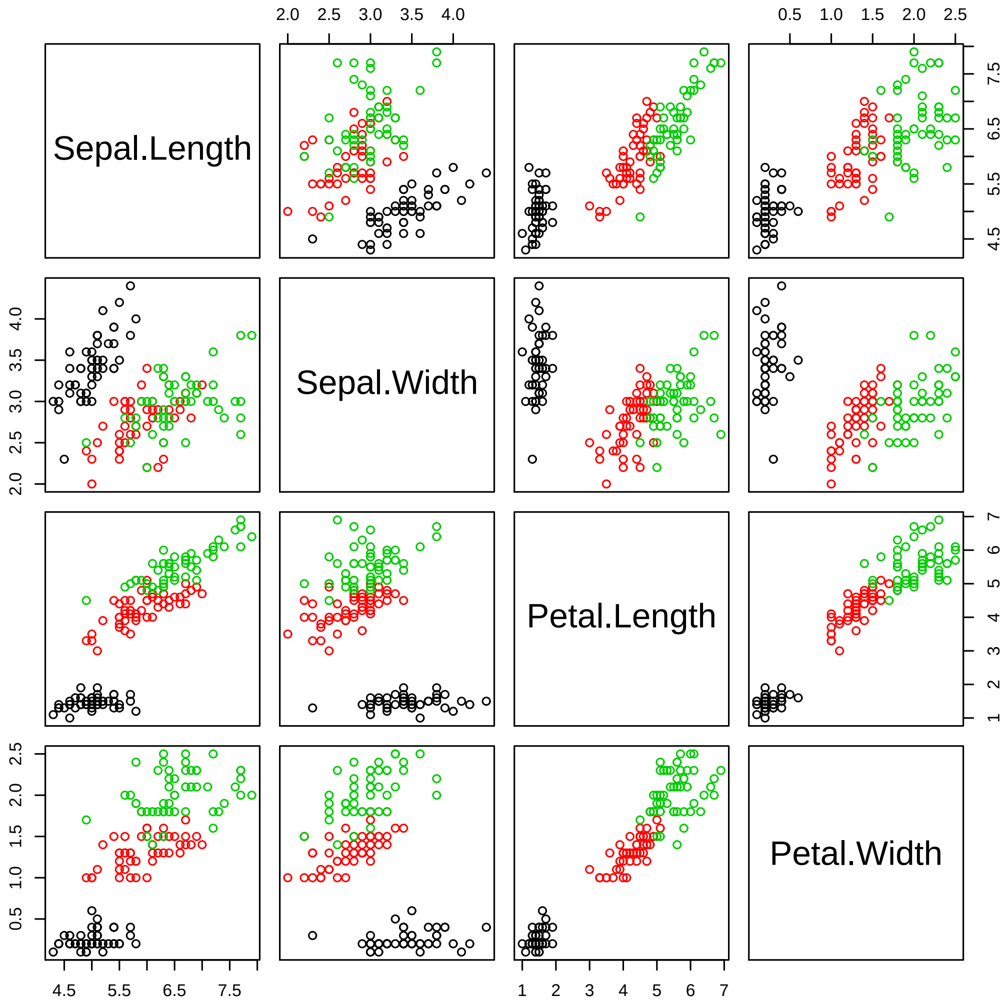


Figure 2: *Iris data; principal components.*

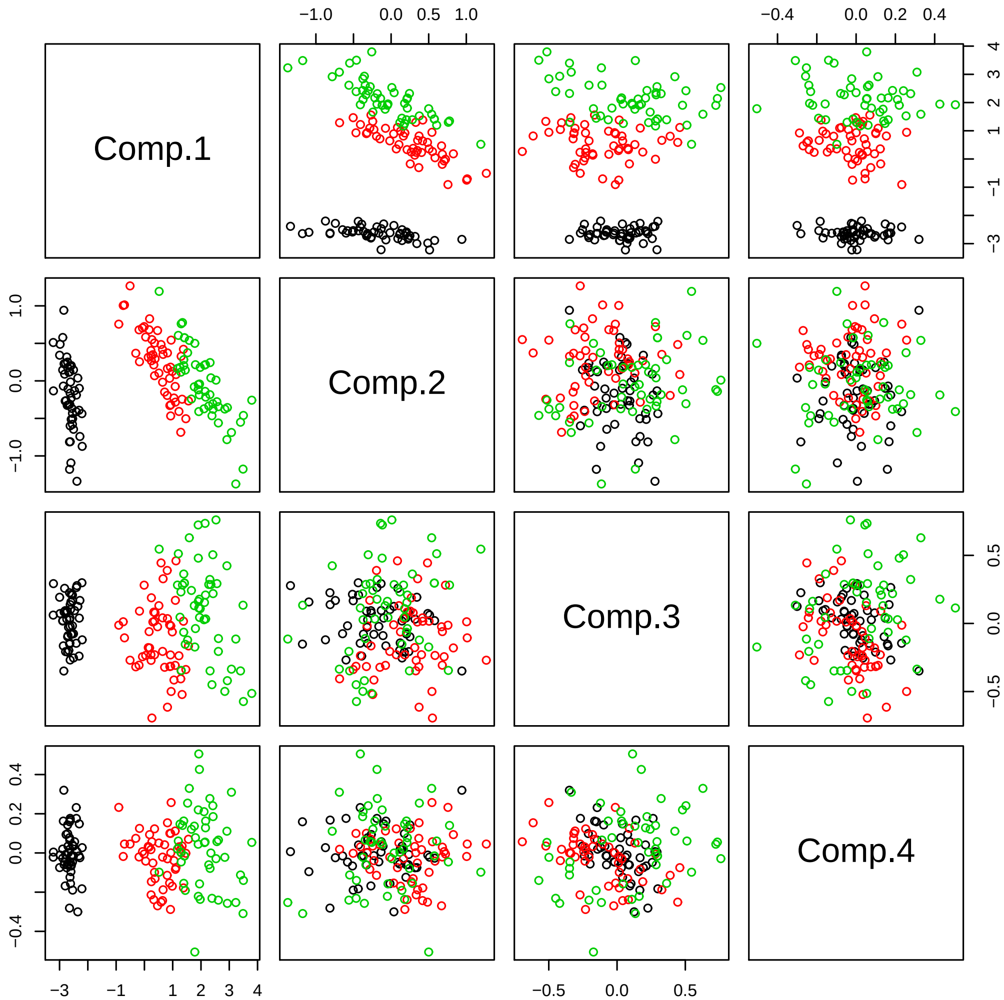


Figure 3: *Iris data; invariant coordinates.*

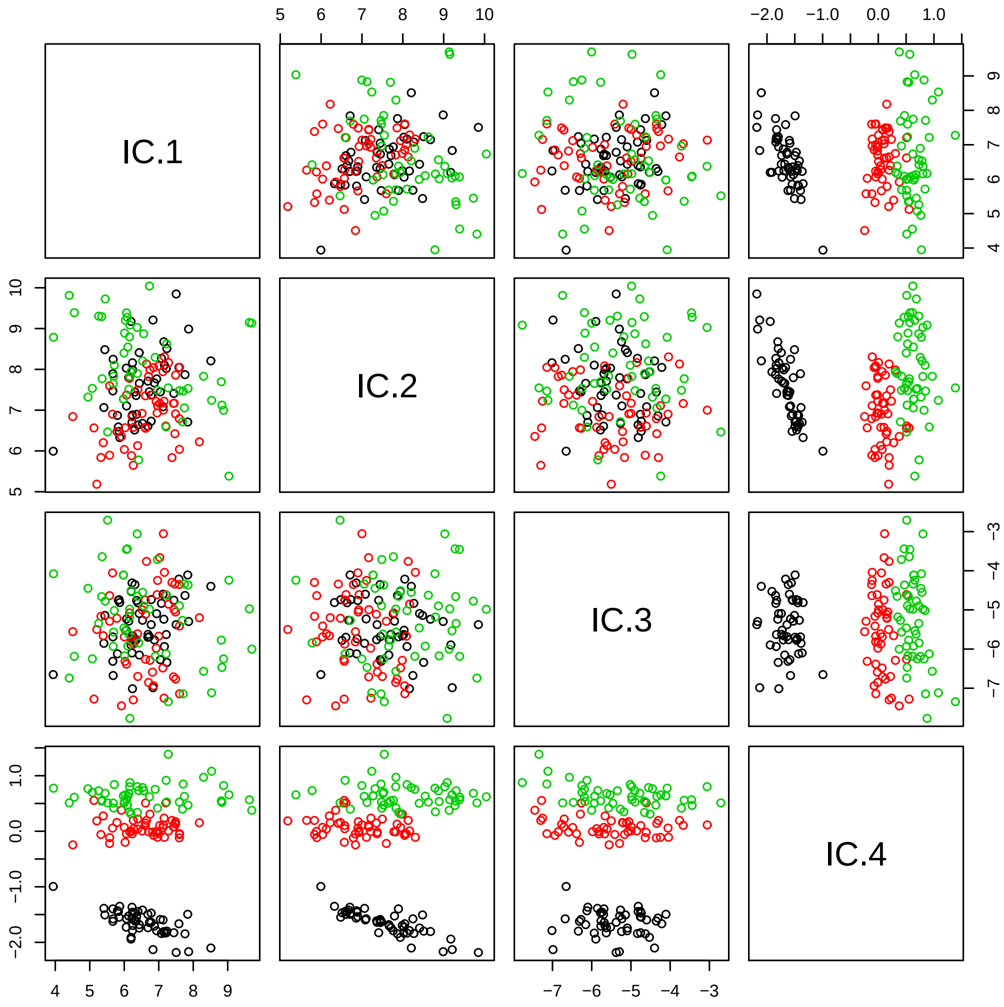


Figure 4: *Dataset 2: Original variables.*

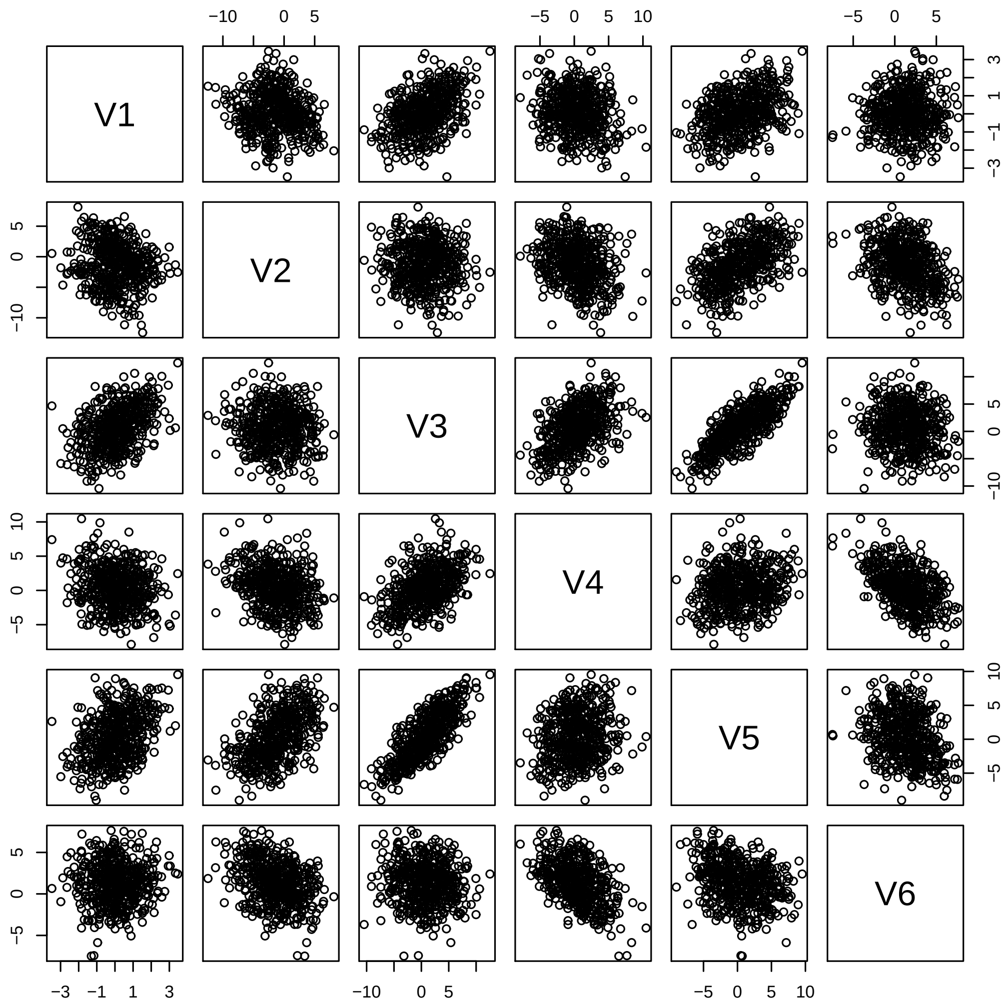


Figure 5: *Dataset 2: Principal components.*

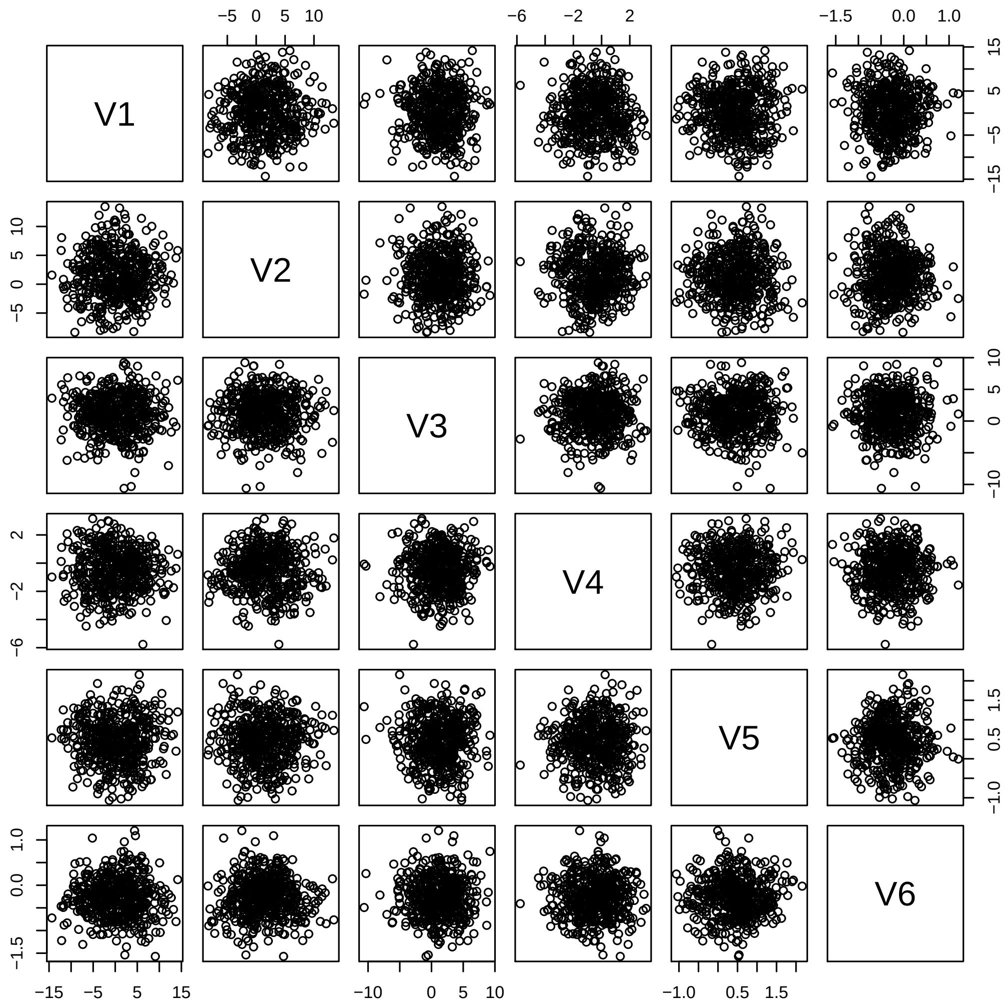


Figure 6: *Dataset 2: Invariant coordinates.*

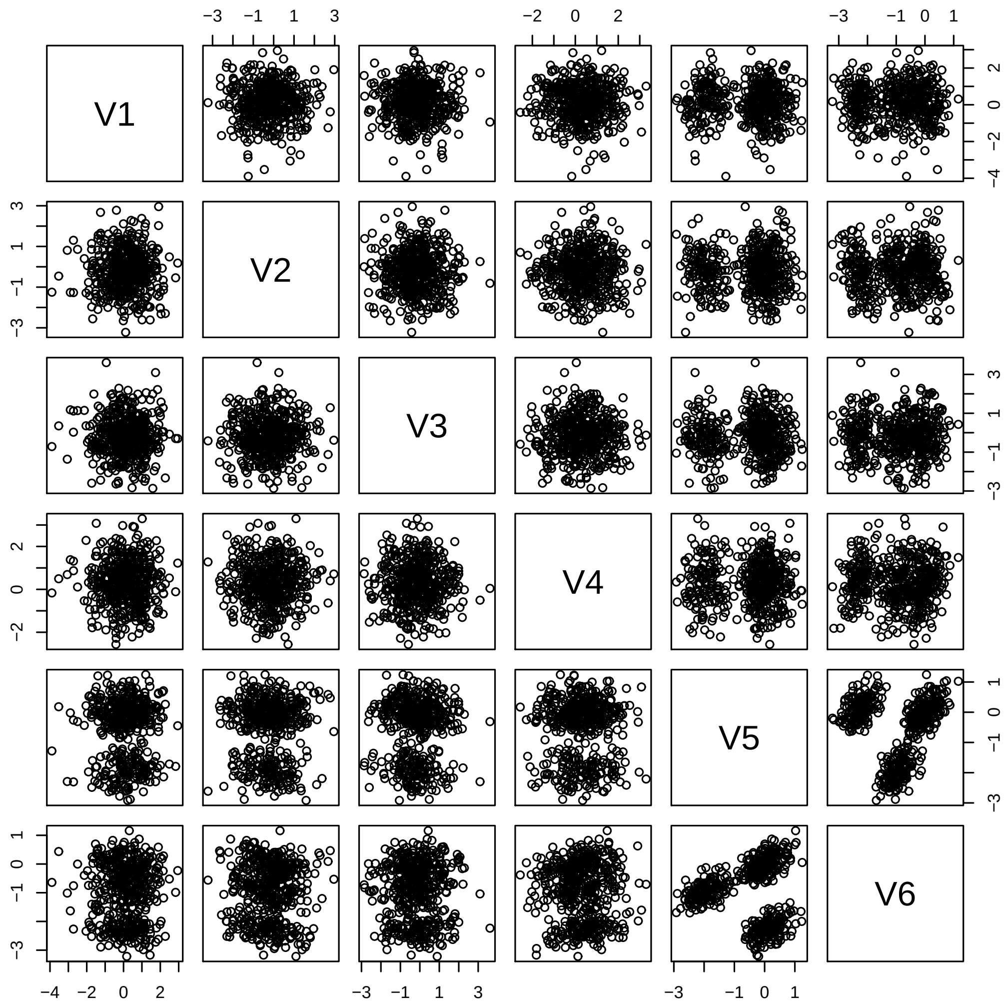


Figure 7: *Dataset 3: Original data.*

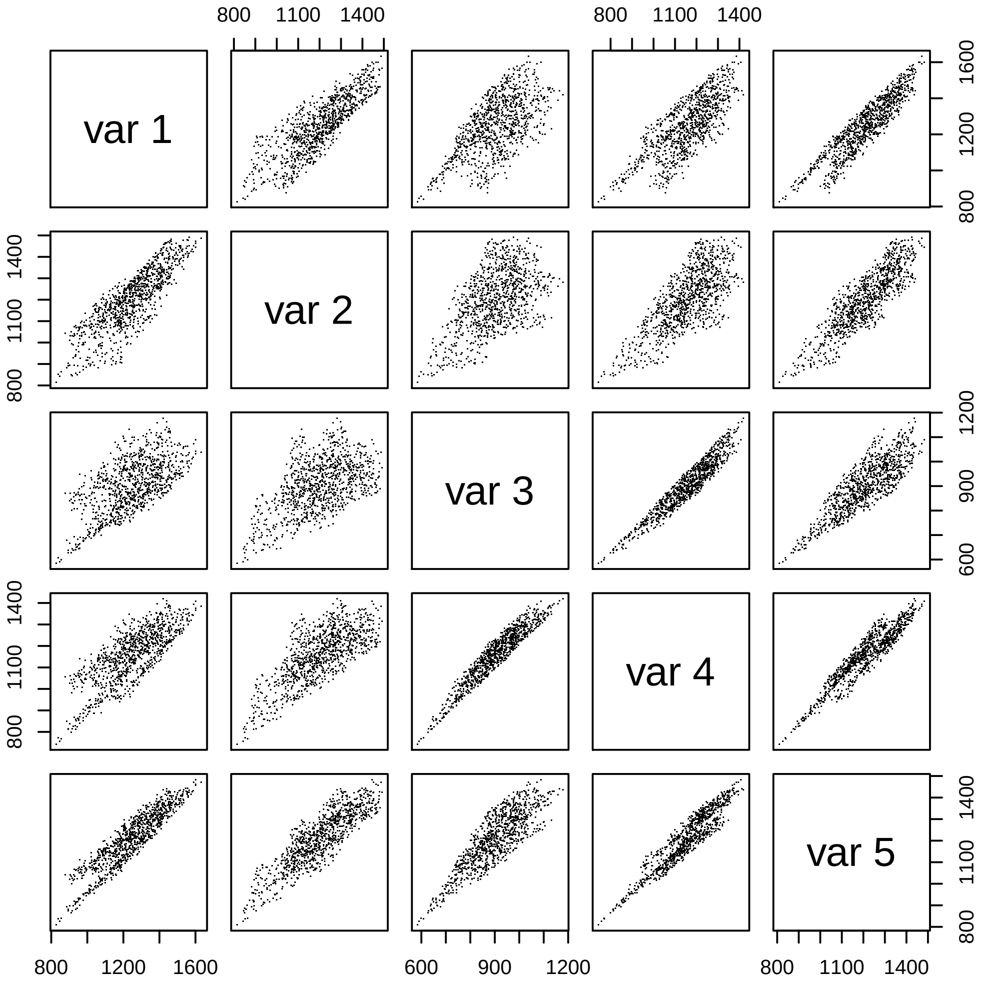


Figure 8: *Dataset 3: Principal components.*

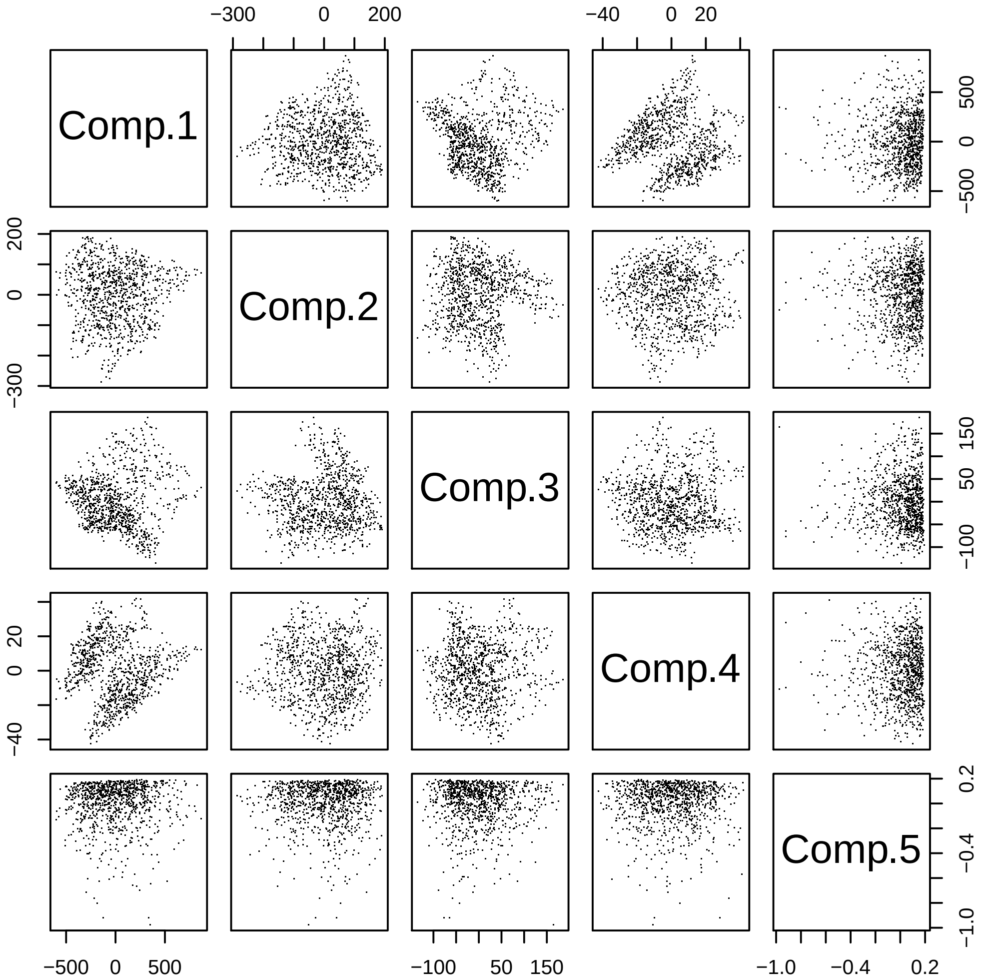
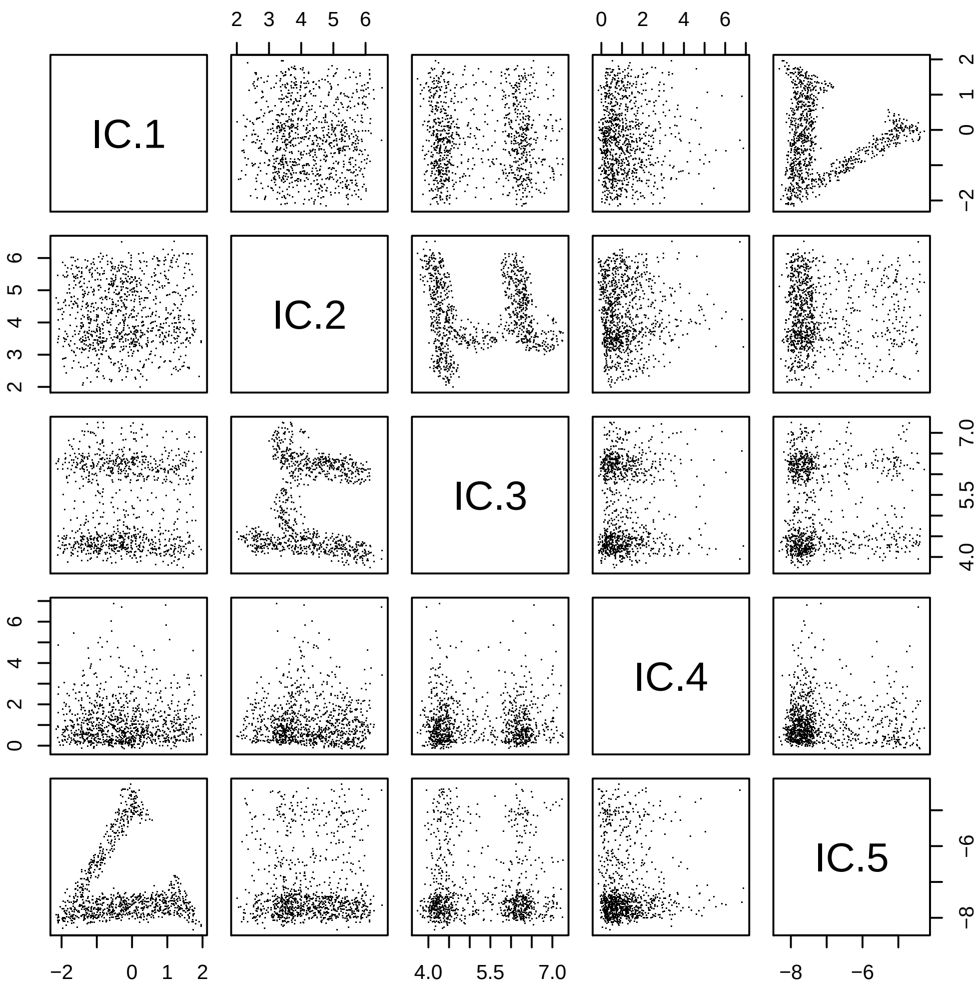


Figure 9: *Dataset 3: Invariant coordinates (using Dümbgen and Huber).*



Use of ICS

- Multivariate invariant/equivariant nonparametric tests and estimates based on transformation and retransformation:
  1. Transform $\mathbf{X} \to \mathbf{Z} = \mathbf{B}\mathbf{X}$
  2. Construct marginal rank tests (Puri-Sen) and corresponding estimates for transformed $\mathbf{Z}$
  3. Retransform estimates back to the original scale
- Optimal rank tests in the IC model - in the spirit of Hallin-Paindaveine tests in the elliptical case
- Hunting for clusters and outliers (using coordinates with high/low kurtosis) - a subset of invariant coordinates can be shown to correspond to Fisher's linear discriminant subspace (under regular assumptions)
- Reduction of dimension - components with high/low kurtosis are often most interesting
- Independent component analysis (ICA): If the two scatter matrices have the independence property then $\mathbf{X} \to \mathbf{B}\mathbf{X}$ transforms to independent components (if the IC model is true)
- R-packages ICS and ICSNP available.

Some asymptotics for ICS functionals

- Let $\hat{\mathbf{S}}_1$, $\hat{\mathbf{S}}_2$, $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Lambda}}$ be calculated from a random sample with corresponding population values $\mathbf{I}_p$, $\mathbf{\Lambda}$, $\mathbf{I}_p$ and $\mathbf{\Lambda}$. $\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $\lambda_1 \geq ... \geq \lambda_p > 0$.
- Assume that $\sqrt{n}(\hat{\mathbf{S}}_1 - \mathbf{I}_p) = O_p(1)$ and $\sqrt{n}(\hat{\mathbf{S}}_2 - \mathbf{\Lambda}) = O_p(1)$
- Then using $\hat{\mathbf{\Gamma}}\hat{\mathbf{S}}_1\hat{\mathbf{\Gamma}}' = \mathbf{I}_p$ and $\hat{\mathbf{\Gamma}}\hat{\mathbf{S}}_2\hat{\mathbf{\Gamma}}' = \hat{\mathbf{\Lambda}}$ one can show that, if $\lambda_i \neq \lambda_j$ for all $j \neq i$, then

$$
\begin{aligned}
\sqrt{n}(\hat{\mathbf{\Lambda}}_{ii} - \lambda_i) &= \sqrt{n}((\hat{\mathbf{S}}_2)_{ii} - \lambda_i) - \lambda_i\sqrt{n}((\hat{\mathbf{S}}_1)_{ii} - 1) + o_p(1), \\
\sqrt{n}(\hat{\mathbf{\Gamma}}_{ii} - 1) &= -\frac{1}{2}\sqrt{n}((\hat{\mathbf{S}}_1)_{ii} - 1) + o_p(1), \\
(\lambda_i - \lambda_j)\sqrt{n}\hat{\mathbf{\Gamma}}_{ij} &= \sqrt{n}(\hat{\mathbf{S}}_2)_{ij} - \lambda_i\sqrt{n}(\hat{\mathbf{S}}_1)_{ij} + o_p(1).
\end{aligned}
$$

- Regular PCA using $\mathbf{S}$: Choose $\hat{\mathbf{S}}_1 = \mathbf{I}_p$ and $\hat{\mathbf{S}}_2 = \hat{\mathbf{S}}$

Supervised location and scatter functionals

- A **supervised location vector** $\mathbf{T}(F_{\mathbf{x},\mathbf{y}})$ is a p-vector valued functional which is affine equivariant in the sense that

$$\mathbf{T}(F_{\mathbf{Ax}+\mathbf{b},\mathbf{y}}) = \mathbf{A}\mathbf{T}(F_{\mathbf{x},\mathbf{y}}) + \mathbf{b}$$

for all nonsingular $\mathbf{A}$ and vector $\mathbf{b}$.

- A **supervised scatter matrix** $\mathbf{S}(F_{\mathbf{x},\mathbf{y}})$ is a $p \times p$ matrix valued functional which is PDS and affine equivariant in the sense that

$$\mathbf{S}(F_{\mathbf{Ax}+\mathbf{b},\mathbf{y}}) = \mathbf{A}\mathbf{S}(F_{\mathbf{x},\mathbf{y}})\mathbf{A}'$$

for all nonsingular $\mathbf{A}$ and vector $\mathbf{b}$.

Supervised location functionals: Examples

Conditional and weighted mean vectors

- $\mathbf{T}(F_{\mathbf{x},\mathbf{y}}) = E(\mathbf{x}|\mathbf{y} = \mathbf{y}_0)$ for a fixed $\mathbf{y}_0$
- $\mathbf{T}(F_{\mathbf{x},\mathbf{y}}) = E[w(\mathbf{y})E(\mathbf{x}|\mathbf{y})]$
- $\mathbf{T}(F_{\mathbf{x},\mathbf{y}}) = E_w(\mathbf{x}) = E(w(\mathbf{y})\mathbf{x})$

where the weight function satisfies $E(w(\mathbf{y})) = 1$.

Supervised scatter functionals: Examples

Conditional and weighted covariance matrices

- $\mathbf{S}(F_{\mathbf{x},\mathbf{y}}) = Cov(\mathbf{x}|\mathbf{y} = \mathbf{y}_0)$ for a fixed $\mathbf{y}_0$
- $\mathbf{S}(F_{\mathbf{x},\mathbf{y}}) = E[w(\mathbf{y})Cov(\mathbf{x}|\mathbf{y})]$
- $\mathbf{S}(F_{\mathbf{x},\mathbf{y}}) = Cov_w(\mathbf{x}) = E[w(\mathbf{y})(\mathbf{x} - E_w(\mathbf{x}))(\mathbf{x} - E_w(\mathbf{x}))']$

where the weight function satisfies $E(w(\mathbf{y})) = 1$.

Supervised invariant coordinate selection (SICS)

- Let $\mathbf{S}_1$ be a scatter functional and $\mathbf{S}_2$ a supervised scatter functional. ($\mathbf{S}_1 = Cov$ and $\mathbf{S}_2 = Cov_w$, for example.)

- Define transformation matrix functional $\mathbf{\Gamma} = \mathbf{\Gamma}(F_{\mathbf{x},\mathbf{y}})$ (and an auxiliary diagonal matrix functional $\mathbf{\Lambda} = \mathbf{\Lambda}(F_{\mathbf{x},\mathbf{y}}))$ as a solution of

$$\mathbf{\Gamma}\mathbf{S}_1\mathbf{\Gamma}' = \mathbf{I}_p \quad \text{and} \quad \mathbf{\Gamma}\mathbf{S}_2\mathbf{\Gamma}' = \mathbf{\Lambda}$$

where the elements of $\mathbf{\Lambda}$ are in a prespecified order.

- Invariant coordinate system (ICS): If the eigenvalues (listed in $\mathbf{\Lambda}$) are distinct, then

$$\mathbf{\Gamma}(F_{\mathbf{Ax},\mathbf{y}})\mathbf{Ax} = \mathbf{\Gamma}(F_{\mathbf{x},\mathbf{y}})\mathbf{x}, \quad \text{for all nonsingular } \mathbf{A}.$$

- In dimension reduction, one is interested in eigenvectors deviating from zero or deviating from one depending on the choice of $\mathbf{S}_1$ and $\mathbf{S}_2$. (If $\mathbf{S}_1 = Cov$ and $\mathbf{S}_2 = Cov_w$, then eigenvectors corresponding to the eigenvalues deviating from one are of interest.)

An example: Australian athletes data

- The response variable is lean body mass (LBM).
- $p = 10$ explanatory variables: height, weight, red cell count, white cell count, hematocrit, hemoglobin, plasma ferritin concentration, body mass index, sum of skin folds, and percent body fat.
- Supervised ICS procedures were based on the regular covariance matrix $\mathbf{S}_1(F)$ and

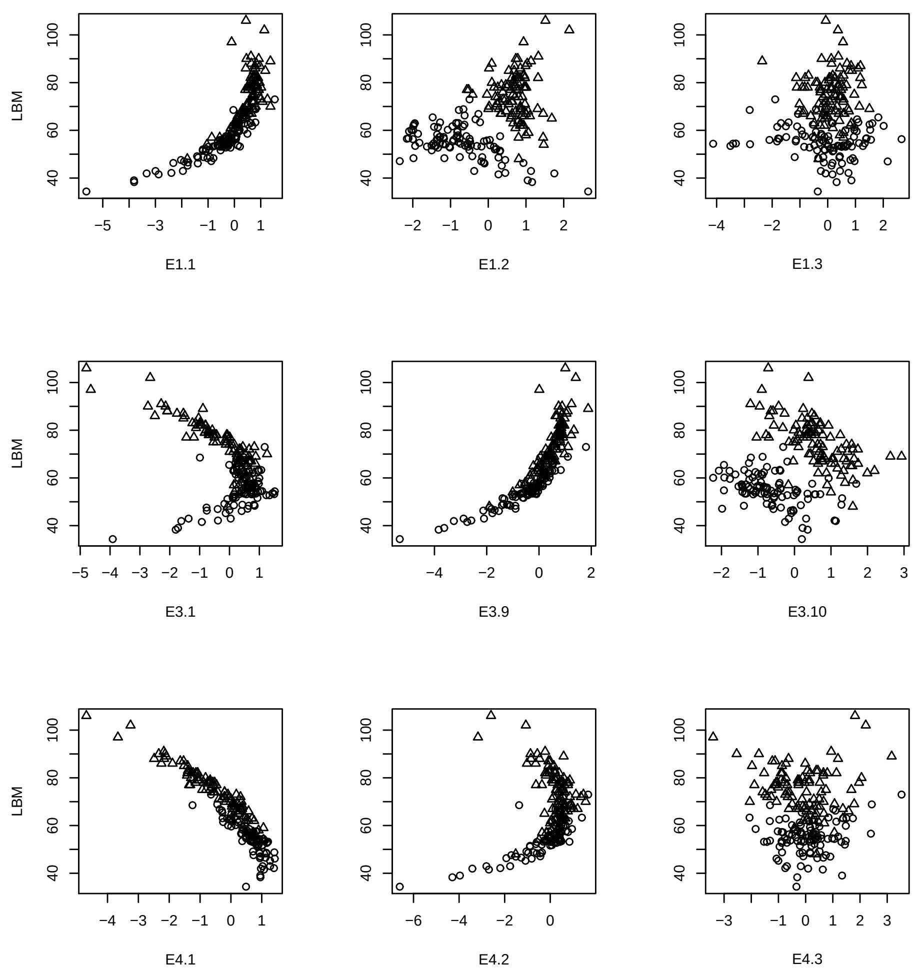
  (E1) $S_2(F_{\mathbf{x},y}) = Cov(\mathbf{x}|y > Q_2(F_y))$

  (E2) $S_2(F_{\mathbf{x},y}) = Cov(\mathbf{x}|Q_1(F_y) < y < Q_3(F_y))$

  (E3) $S_2(F_{\mathbf{x},y}) = Cov\left(\mathbf{x}_i - \mathbf{x}_j \;\middle|\; |y_i - y_j| > F^{-1}_{|y_i-y_j|}(0.9)\right)$,

  where $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$ are two independent copies from the distribution of $(\mathbf{x}, y)$.
- We consider $k = 3$ supervised invariant coordinates with eigenvalues differing most from one.

Figure 10: *Reduced dimension variables vs* LBM. $(\mathbf{E1})$ *first row,* $(\mathbf{E2})$ *second row, and* $(\mathbf{E3})$ *third row.*



Asymptotics for supervised ICS functionals

- Assume that $\sqrt{n}(\hat{\mathbf{S}}_1 - \mathbf{I}_p) = O_p(1)$ and $\sqrt{n}(\hat{\mathbf{S}}_2 - \mathbf{\Lambda}) = O_p(1)$
- Then using $\hat{\mathbf{\Gamma}}\hat{\mathbf{S}}_1\hat{\mathbf{\Gamma}}' = \mathbf{I}_p$ and $\hat{\mathbf{\Gamma}}\hat{\mathbf{S}}_2\hat{\mathbf{\Gamma}}' = \hat{\mathbf{\Lambda}}$ one can show that, if $\lambda_i \neq \lambda_j$ for all $j \neq i$, then

$$\begin{aligned}
\sqrt{n}(\hat{\lambda}_i - \lambda_i) &= \sqrt{n}((\hat{\mathbf{S}}_2)_{ii} - \lambda_i) - \lambda_i\sqrt{n}((\hat{\mathbf{S}}_1)_{ii} - 1) + o_p(1), \\
\sqrt{n}(\hat{\mathbf{\Gamma}}_{ii} - 1) &= -\frac{1}{2}\sqrt{n}((\hat{\mathbf{S}}_1)_{ii} - 1) + o_p(1), \\
(\lambda_i - \lambda_j)\sqrt{n}\hat{\mathbf{\Gamma}}_{ij} &= \sqrt{n}(\hat{\mathbf{S}}_2)_{ij} - \lambda_i\sqrt{n}(\hat{\mathbf{S}}_1)_{ij} + o_p(1).
\end{aligned}$$

- Testing whether exactly $p-k$ eigenvalues are one: Use the test statistic

$$n \cdot \sum_{i=k+1}^{p} (\hat{\lambda}_i - 1)^2.$$

- Testing whether exactly $p-k$ eigenvalues are zero (as in SIR): Use the test statistic

$$n \cdot \sum_{i=k+1}^{p} \hat{\lambda}_i.$$

THANK YOU FOR YOUR ATTENTION !