Multivariate Nonparametrical Methods Based on Spatial Signs and Ranks

Hannu Oja

Tampere, November 2008

Rough plan

- Data: $(\mathbf{X}, \mathbf{Y})$
- L_1 criterion functions
- Spatial sign, rank, signed-rank: $\mathbf{U}$, $\mathbf{R}$ and $\mathbf{Q}$
- One sample case: Tests and estimates
- Several samples case: Tests
- Multivariate linear regression case: Tests and estimates
- Other approaches

Data and problem

- Data:

$$(\mathbf{X}, \mathbf{Y})$$

  where $\mathbf{X}$ is a $n \times q$ matrix of explaining variables and $\mathbf{Y}$ is a $n \times p$ matrix of response variables.

- The linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

  where $\boldsymbol{\beta}$ is a $q \times p$ matrix of regression coefficients and $\mathbf{E}$ is a $n \times p$ matrix of unobserved residuals.

- We assume that $\mathbf{E}$ is a random sample from a p-variate distribution with density function $f(\mathbf{e})$ centered/symmetric around the origin.

- Special cases: One sample, several samples case

Some important matrices

- Projection matrix

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- Replace observations by their mean vector: Use $\mathbf{P}_{\mathbf{1}_n} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ and transform

$$\mathbf{Y} \;\rightarrow\; \mathbf{P}_{\mathbf{1}_n}\mathbf{Y}$$

- Sign-change matrix $\mathbf{J}$: A diagonal matrix with diagonal elements ± 1:
  If $\mathbf{E}$ is a random sample form a symmetric distribution then $\mathbf{JE} \sim \mathbf{E}$.

- Permutation matrix $\mathbf{P}$ is obtained by the identity matrix by permuting its rows and columns:
  If $\mathbf{E}$ is a random sample then $\mathbf{PE} \sim \mathbf{E}$.

L_1 criterion functions

- One can then also write

$$\mathbf{y}_i = \boldsymbol{\beta}'\mathbf{x}_i + \mathbf{e}_i, \;\; i = 1, ..., n.$$

- If we write $\mathbf{e}_i = \mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i$, $i = 1, ..., n$, then the regular LS estimate minimizes the L_2 criterion function

$$\text{ave}\{||\mathbf{e}_i||^2\} = \text{ave}\{\mathbf{e}_i'\mathbf{e}_i\}$$

- Consider the L_1 criterion functions

$$\text{ave}\{||\mathbf{e}_i||\},$$

$$\text{ave}\{||\mathbf{e}_i - \mathbf{e}_j||\}, \;\; \text{and}$$

$$\text{ave}\{||\mathbf{e}_i - \mathbf{e}_j|| + ||\mathbf{e}_i + \mathbf{e}_j||\}.$$

  See Hettmansperger and Aubuchon (1988).

Spatial sign, rank and signed-rank

- The multivariate spatial sign $\mathbf{U}_i$, multivariate spatial (centered) rank $\mathbf{R}_i$, and multivariate spatial signed-rank $\mathbf{Q}_i$, $i = 1, ..., n$ are implicitly defined using the above three L_1 criterion functions:

$$\begin{aligned} \text{ave}\{||\mathbf{e}_i||\} &= \text{ave}\{\mathbf{U}_i'\mathbf{e}_i\}, \\ \frac{1}{2}\text{ave}\{||\mathbf{e}_i - \mathbf{e}_j||\} &= \text{ave}\{\mathbf{R}_i'\mathbf{e}_i\}, \text{ and} \\ \frac{1}{4}\text{ave}\{||\mathbf{e}_i - \mathbf{e}_j|| + ||\mathbf{e}_i + \mathbf{e}_j||\} &= \text{ave}\{\mathbf{Q}_i'\mathbf{e}_i\}. \end{aligned}$$

- If $\mathbf{U}(\mathbf{e}) = ||\mathbf{e}||^{-1}\mathbf{e}$ if $\mathbf{e} \neq \mathbf{0}$ and $\mathbf{0}$ if $\mathbf{e} = \mathbf{0}$ then

$$\begin{aligned} \mathbf{U}_i &= \mathbf{U}(\mathbf{e}_i), \\ \mathbf{R}_i &= \text{ave}_j\{\mathbf{U}(\mathbf{e}_i - \mathbf{e}_j)\}, \text{ and} \\ \mathbf{Q}_i &= \frac{1}{2}\text{ave}_j\{\mathbf{U}(\mathbf{e}_i - \mathbf{e}_j) + \mathbf{U}(\mathbf{e}_i + \mathbf{e}_j)\} \end{aligned}$$

- Univariate case: regular sign, rank and signed rank.

Figure 1: *Bivariate scores: original data.*

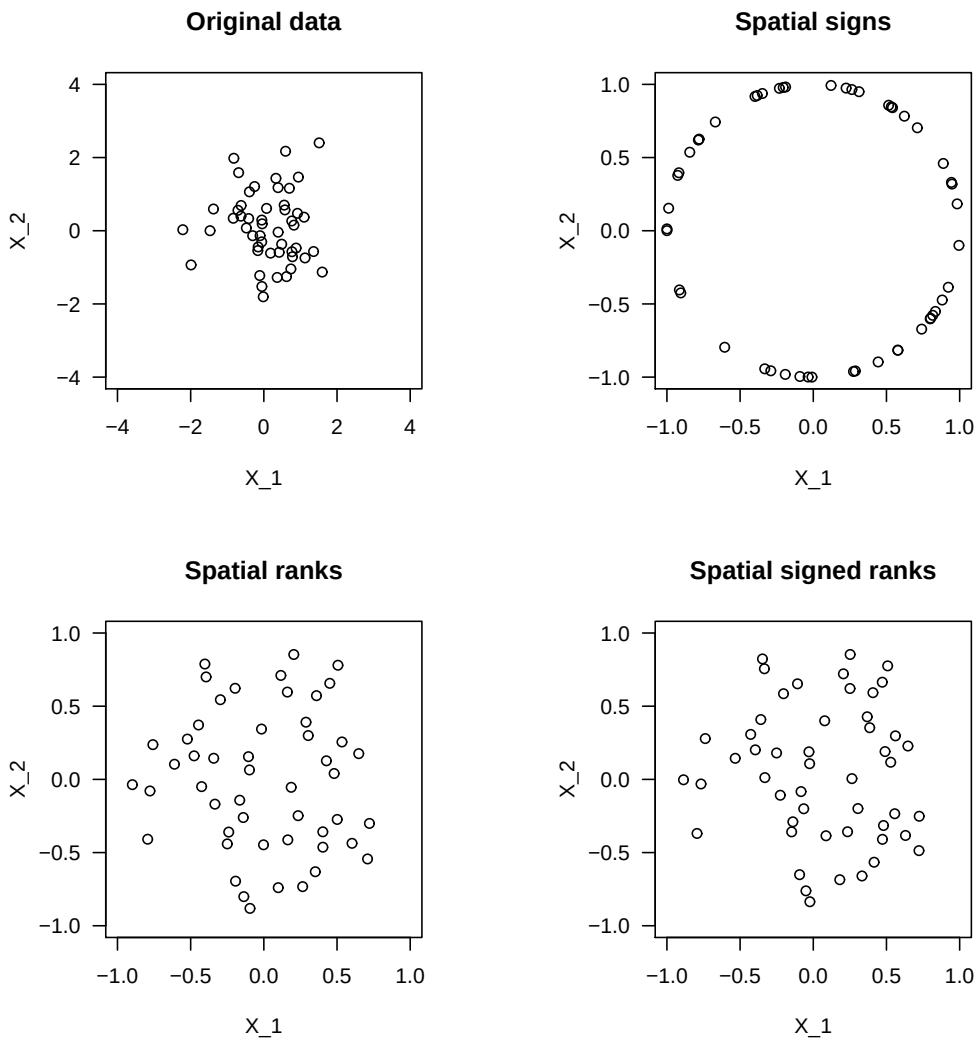


Figure 2: *Bivariate scores: rescaled data.*

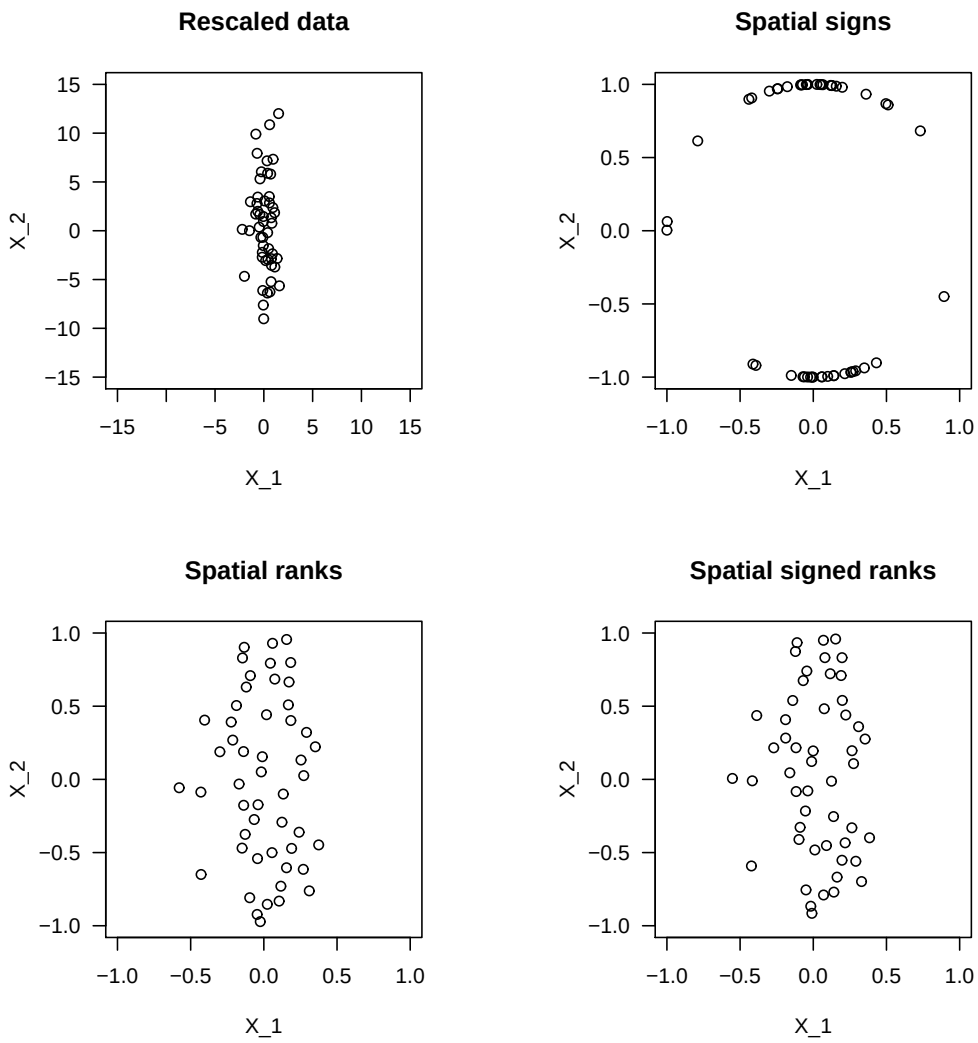
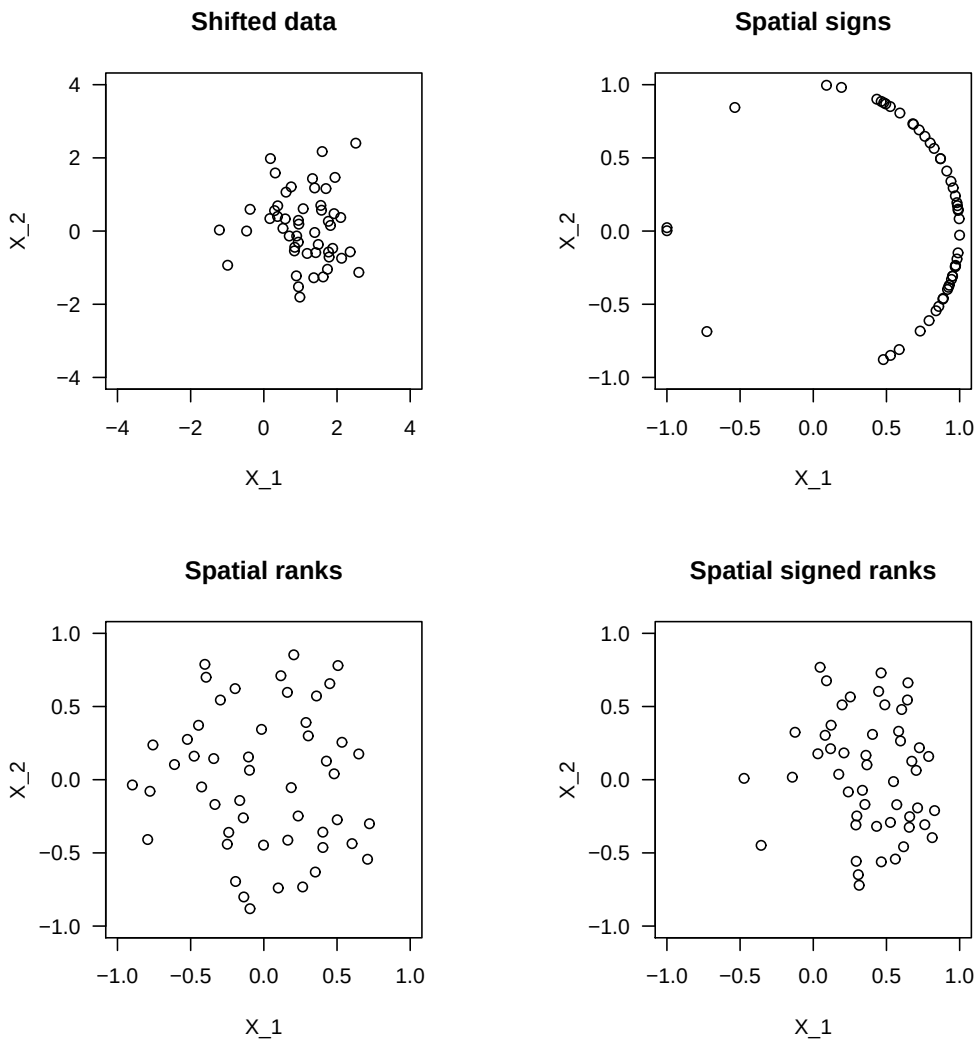


Figure 3: *Bivariate scores: shifted data.*



The scores which we use

- Let $\mathbf{T}(\mathbf{y})$ be a p-vector valued score function: For the statistical analysis of the data, we often transform

$$\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)' \quad \rightarrow \quad \mathbf{T} = (\mathbf{T}_1, ..., \mathbf{T}_n)' \text{ or}$$
$$\rightarrow \quad \hat{\mathbf{T}} = (\hat{\mathbf{T}}_1, ..., \hat{\mathbf{T}}_n)'$$

- Identity score $\mathbf{T}(\mathbf{y}) = \mathbf{y}$
- Spatial sign score $\mathbf{U}(\mathbf{y})$
- Spatial rank score $\mathbf{R}(\mathbf{y}) = \text{ave}_j\{\mathbf{U}(\mathbf{y} - \mathbf{y}_j)\}$
- Spatial signed-rank score $\mathbf{Q}(\mathbf{y}) = \frac{1}{2}\text{ave}_j\{\mathbf{U}(\mathbf{y} - \mathbf{y}_j) + \mathbf{U}(\mathbf{y} + \mathbf{y}_j)\}$
- Optimal score function $\mathbf{L}(\mathbf{y}) = \nabla \log f(\mathbf{y})$

One-sample testing problem: General strategy

- We assume that $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)'$ is a random sample from a distribution generated by

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{e}_i, \;\; i = 1, ..., n,$$

  where $\mathbf{e}_i \sim -\mathbf{e}_i$. We wish to test the null hypothesis $H_0 : \; \boldsymbol{\mu} = \mathbf{0}$ and estimate the unknown $\boldsymbol{\mu}$. The test and estimate is based on an odd score function $\mathbf{T}(\mathbf{y})$.

- The test statistic is obtained as follows:

$$\mathbf{Y} \;\; \rightarrow \;\; \mathbf{T} \;\; \rightarrow \;\; Q^2(\mathbf{Y}) = \mathbf{1}_n' \mathbf{T} (\mathbf{T}'\mathbf{T})^{-1} \mathbf{T}' \mathbf{1}_n$$

- The test statistic is not affine invariant in general: It is not necessarily true that

$$Q^2(\mathbf{YA}) = Q^2(\mathbf{Y}), \quad \text{for all } \mathbf{A}$$

- The p value obtained from an asymptotic distribution or using an exact sign-change version of the test.
- Under the sequence of alternatives $H_n : \boldsymbol{\mu} = n^{-1/2}\boldsymbol{\delta}$,

$$Q^2 \rightarrow_d \chi^2_p(\boldsymbol{\delta}'\mathbf{A}\mathbf{B}^{-1}\mathbf{A}\boldsymbol{\delta})$$

  where

$$\mathbf{A} = E\{\mathbf{T}(\mathbf{e}_i)\mathbf{L}(\mathbf{e}_i)'\} \text{ and } \mathbf{B} = E\left\{\mathbf{T}(\mathbf{e}_i)\mathbf{T}(\mathbf{e}_i)'\right\}.$$

- Let $\mathbf{J}$ be a sign-change matrix (a diagonal matrix with diagonal elements ± 1). Then the p value from an exact test is obtained as

$$E_{\mathbf{J}}\left[I\left(Q^2(\mathbf{J}\mathbf{Y}) \geq Q^2(\mathbf{Y})\right)\right]$$

  where $\mathbf{J}$ is uniformly distributed over all 2^n different cases.

One-sample testing problem: Inner standardization

- Find positive definite symmetric (scatter matrix) $\mathbf{S}$ such that if

$$\hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{S}^{-1/2}\mathbf{y}_i),\ i = 1, ..., n,$$

and

$$\hat{\mathbf{T}} = (\hat{\mathbf{T}}_1, ..., \hat{\mathbf{T}}_n)'$$

then

$$\hat{\mathbf{T}}'\hat{\mathbf{T}} \propto \mathbf{I}_p$$

- The test statistic is obtained as follows:

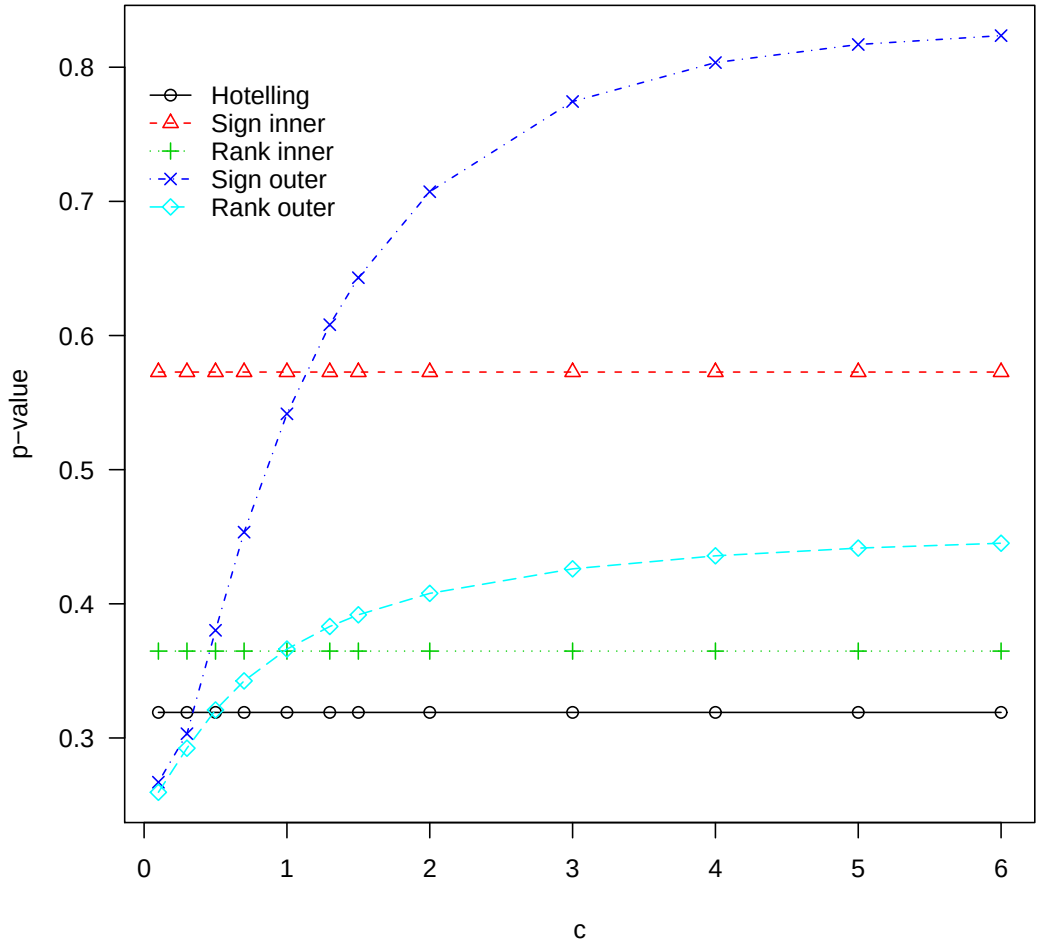
$$\mathbf{Y} \rightarrow \hat{\mathbf{T}} \rightarrow Q^2(\mathbf{Y}\mathbf{S}^{-1/2}) = \mathbf{1}_n'\hat{\mathbf{T}}(\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}'\mathbf{1}_n$$

- $Q^2(\mathbf{Y}\mathbf{S}^{-1/2})$ is an affine invariant version of the test statistic

One-sample testing problem: Different tests

- Identity score: Regular Hotelling's T^2
- Spatial sign score: Multivariate extension of univariate sign test
- Spatial rank score: Multivariate extension of univariate Wilcoxon signed-rank test.
- Inner centering is needed for the affine invariance of the sign and signed-rank tests

Figure 4: *150 observations from* $N_2((.2, 0), \mathbf{I}_2)$*. The second component is multiplied by c.*



One-sample testing problem: Liming efficiencies

Table 1: Asymptotic efficiencies of the sign test and the signed-rank test relative to Hotelling's T^2 under p-variate t distributions with ν degrees of freedom for selected values of p and ν.

	Sign test			Signed-rank test		
dimension p	$\nu = 3$	$\nu = 6$	$\nu = \infty$	$\nu = 3$	$\nu = 6$	$\nu = \infty$
1	1.62	0.88	0.64	1.90	1.16	0.95
2	2.00	1.08	0.78	1.95	1.19	0.97
4	2.25	1.22	0.88	2.02	1.21	0.98
10	2.42	1.31	0.95	2.09	1.22	0.99

One-sample estimation problem: General strategy

- We assume that $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)'$ is a random sample from a distribution generated by

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{e}_i, \;\; i = 1, ..., n,$$

  where $\mathbf{e}_i \sim -\mathbf{e}_i$. We wish to estimate unknown $\boldsymbol{\mu}$.

- The location estimate $\hat{\boldsymbol{\mu}}$ is determined by the estimating equation

$$\sum_{i=1}^{n} \mathbf{T}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}) = \mathbf{0}.$$

- Under general assumptions

$$\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \;=\; \mathbf{A}^{-1}\, \sqrt{n}\, \text{ave}\{\mathbf{T}(\mathbf{e}_i)\} + o_P(1) \;\rightarrow_d\; N_p(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$$

- One sample case: $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ is reduced to σ^2 (identity), $[4f^2(\mu)]^{-1}$ (sign) and $[12(\int f^2(y)dy)^2]^{-1}$ (signed-rank).

One-sample location estimates

- Identity score: The sample mean vector which minimizes

$$\sum_{i=1}^{n} ||\mathbf{y}_i - \boldsymbol{\mu}||^2$$

- Spatial sign score: The sample spatial median which minimizes

$$\sum_{i=1}^{n} ||\mathbf{y}_i - \boldsymbol{\mu}||$$

- Spatial signed-rank score: The sample spatial Hodges-Lehmann (HL) estimator which minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{y}_i + \mathbf{y}_j - 2\boldsymbol{\mu}||$$

- The spatial median and HL estimate can be made affine equivariant using transformation-retransformation (TR) technique - that is - simultaneous estimation of location and scatter

Simultaneous estimates for location and scatter

- Find $\hat{\boldsymbol{\mu}}$ and $\mathbf{S}$ such that if

$$\hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{S}^{-1/2}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})), \;\; i = 1, ..., n,$$

and

$$\hat{\mathbf{T}} = (\hat{\mathbf{T}}_1, ..., \hat{\mathbf{T}}_n)'$$

then

$$\hat{\mathbf{T}}'\mathbf{1}_n = \mathbf{0} \text{ and } \hat{\mathbf{T}}'\hat{\mathbf{T}} \propto \mathbf{I}_p$$

- Then $\hat{\boldsymbol{\mu}}$ and $\mathbf{S}$ are location and scatter estimates corresponding to the score function $\mathbf{T}$.

Algorithms for the spatial median

- The Weiszfeld algorithm for the spatial median

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \frac{\text{ave}\{\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})\}}{\text{ave}\{||\mathbf{y}_i - \boldsymbol{\mu}||\}}$$

- Hettmansperger-Randles estimate: One iteration step (as in M-estimation) updates first the residuals, then the location center, and finally the scatter matrix as follows.

1.
$$\mathbf{e}_i \leftarrow \mathbf{S}^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}), \;\; i = 1, ..., n$$

2.
$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \frac{\mathbf{S}^{1/2}\,\text{ave}\{\mathbf{U}(\mathbf{e}_i)\}}{\text{ave}\{||\mathbf{e}_i||^{-1}\}}$$

3.
$$\mathbf{S} \leftarrow p\,\mathbf{S}^{1/2}\,\text{ave}\{\mathbf{U}(\mathbf{e}_i)\mathbf{U}(\mathbf{e}_i)'\}\,\mathbf{S}^{1/2}.$$

Figure 5: *A sample of from a bivariate normal distribution: Estimates with 95 % confidence ellipsoids.*

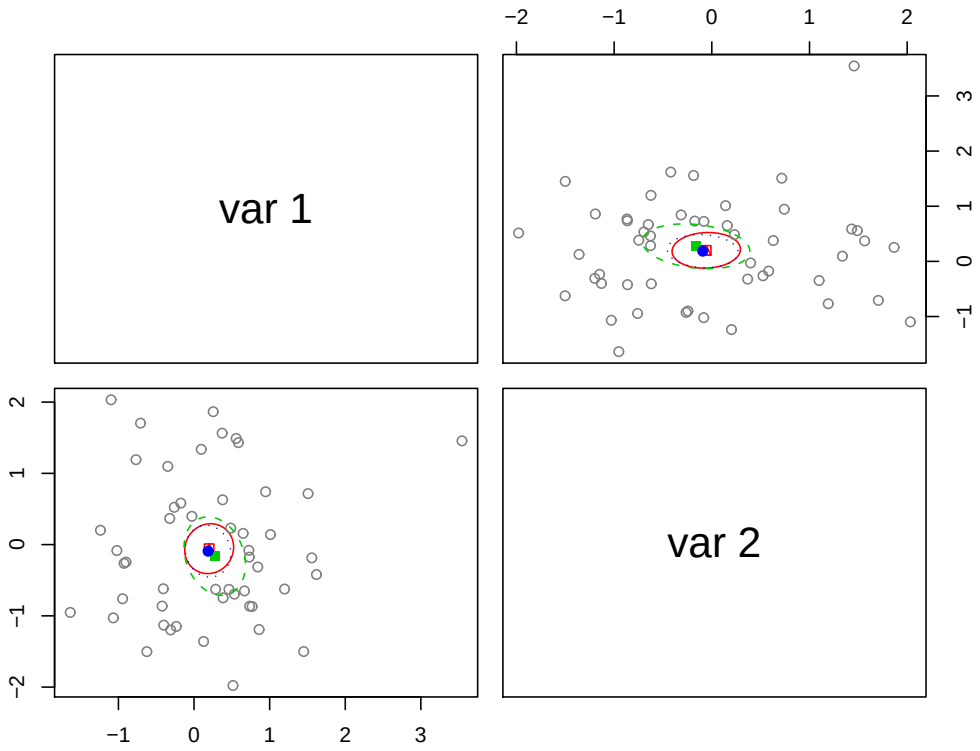


Figure 6: *A sample from a bivariate normal distribution: Estimates with 95 % confidence ellipsoids.*

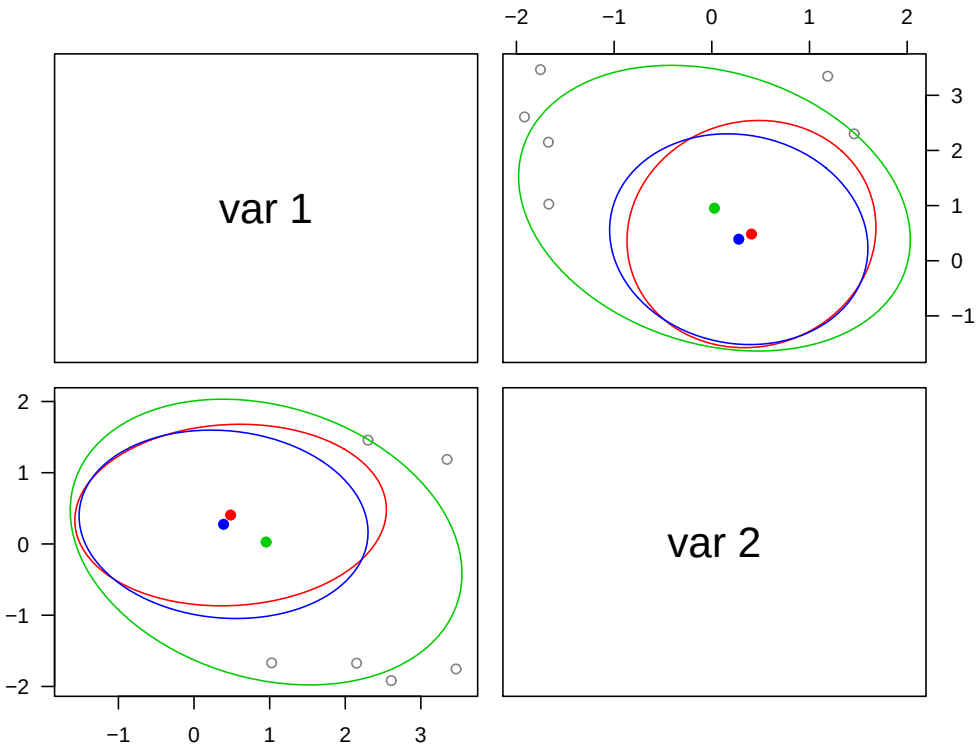


Figure 7: *A sample from a trivariate t_3 distribution: Estimates with 95 % confidence ellipsoids*

.

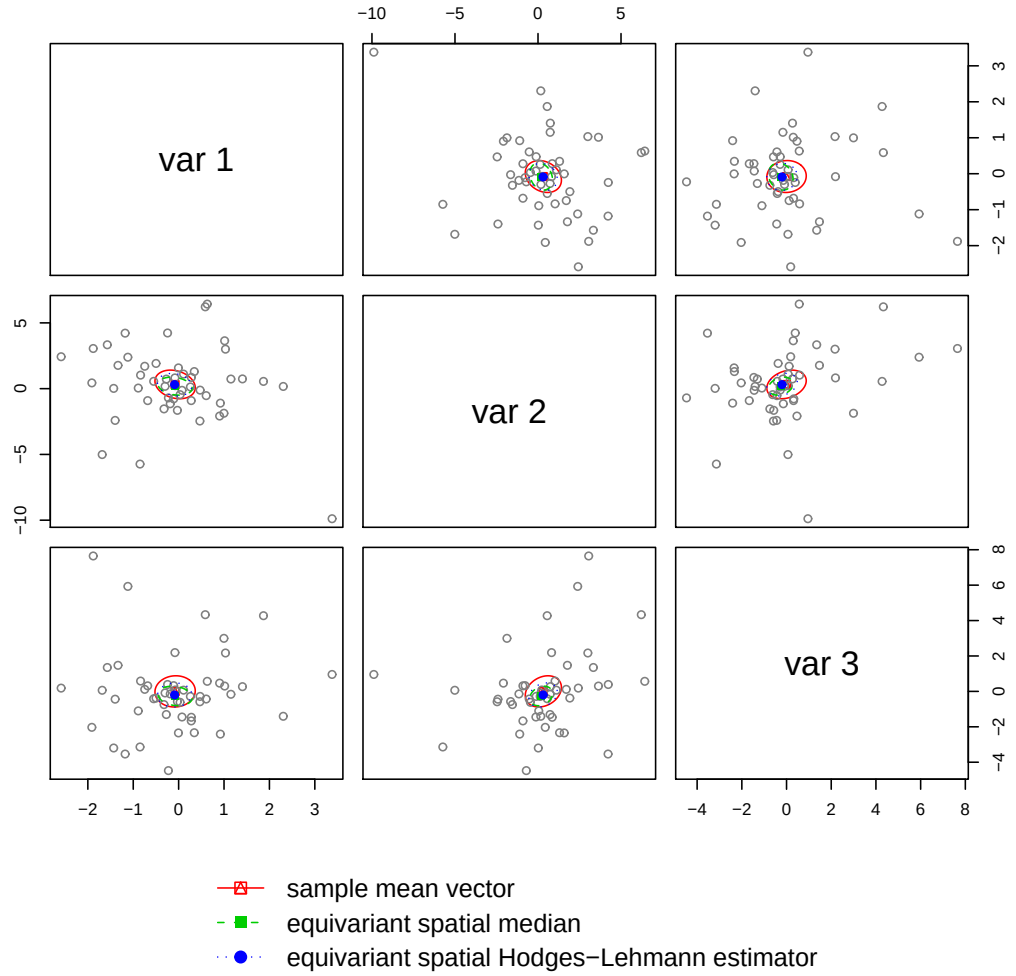
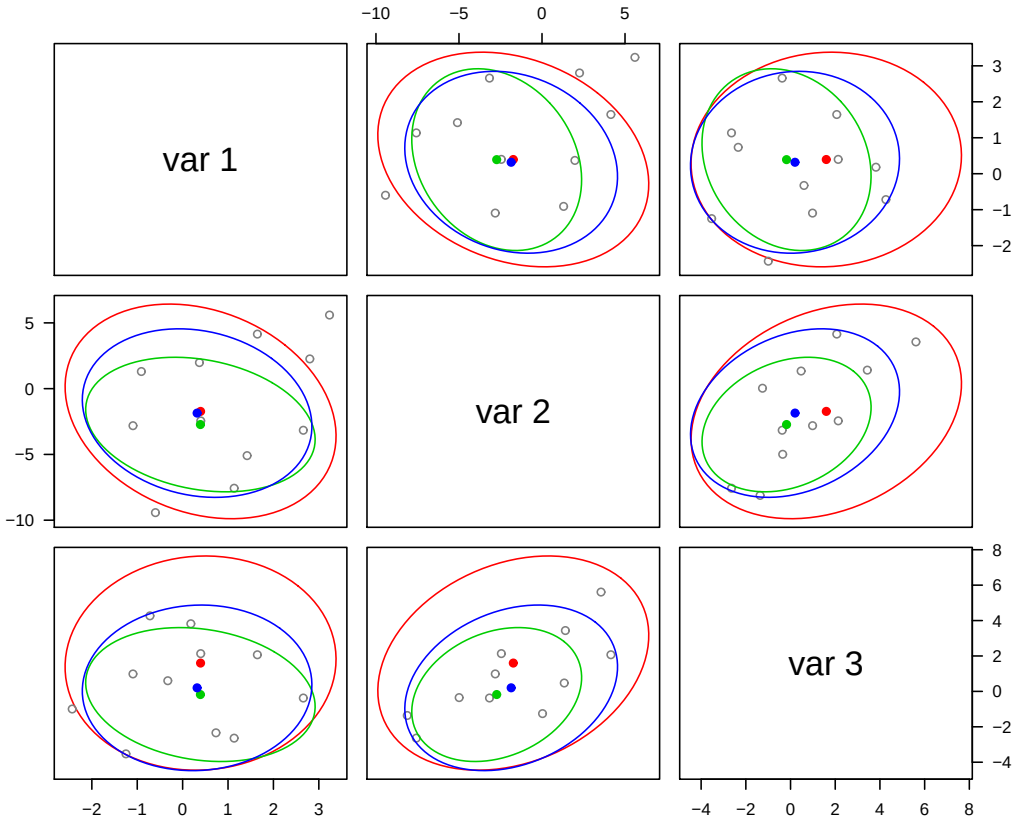


Figure 8: *A sample from a trivariate t_3 distribution: Estimates with 95 % confidence ellipsoids*

.



Several samples testing problem: Model

- The data are given in the form

$$(\mathbf{X}, \mathbf{Y})$$

where $\mathbf{X}$ is an $n \times c$ matrix indicating the group/sample membership.

- We assume that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{E}$$

where $\boldsymbol{\mu}$ is the $c \times p$ matrix giving the group centers $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_c$. We wish to test the null hypothesis $H_0 : \boldsymbol{\mu}_1 = ... = \boldsymbol{\mu}_c$.

Several samples testing problem: General strategy

- Choose $\hat{\boldsymbol{\mu}}$ so that $\sum \mathbf{T}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}) = \mathbf{0}$ and write $\hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}), i = 1, ..., n.$
- Transform

$$\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)' \;\rightarrow\; \hat{\mathbf{T}} \;\rightarrow\; Q^2 = n \cdot tr\left((\hat{\mathbf{T}}'\mathbf{P}_{\mathbf{X}}\hat{\mathbf{T}})(\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\right)$$

  where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

- Assume that $\frac{1}{n}\mathbf{X}'\mathbf{X} \rightarrow \mathbf{D}$. Under the null hypothesis Pillai's trace

$$Q^2 \;\rightarrow_d\; \chi^2_{(c-1)p}$$

- For permuted observations, we obtain

$$Q^2(\mathbf{PY}) = n \cdot tr\left((\hat{\mathbf{T}}'\mathbf{P}'\mathbf{P}_{\mathbf{X}}\mathbf{P}\hat{\mathbf{T}})(\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\right)$$

  and the p value from an exact (conditionally distribution-free test) is

$$E_{\mathbf{P}}\left[I\left(Q^2(\mathbf{PY}) \geq Q^2(\mathbf{Y})\right)\right]$$

Several samples testing problem: Approximate power

- Write $\lambda_i = \lim(n_i/n)$, $i = 1, ..., c$. Consider the alternative sequence

$$H_n : \boldsymbol{\mu}_i = \boldsymbol{\mu} + n^{-1/2}\boldsymbol{\delta}_i, \quad i = 1, ..., c$$

where $\sum \lambda_i \boldsymbol{\delta}_i = \mathbf{0}$. Then the test statistic Q^2 has a limiting noncentral chi squared distribution with $(c-1)p$ degrees of freedom and noncentrality parameter

$$\sum_{i=1}^{c} \lambda_i \boldsymbol{\delta}_i' (\mathbf{A}\mathbf{B}^{-1}\mathbf{A}) \boldsymbol{\delta}_i.$$

- If

$$\mathbf{Y} = \boldsymbol{\Delta} + \mathbf{E}, \quad \text{where } \boldsymbol{\Delta}'\mathbf{1}_n = \mathbf{0}$$

then the approximate distribution of Q^2 is a noncentral chi squared distribution with $(c-1)p$ degrees of freedom and noncentrality parameter

$$tr\left((\boldsymbol{\Delta}'\mathbf{P}_{\mathbf{X}}\boldsymbol{\Delta})(\mathbf{A}\mathbf{B}^{-1}\mathbf{A})\right)$$

Several samples testing problem: Different tests

- Identity score: Regular MANOVA (Pillai's trace)
- Spatial sign score: Multivariate extension of classical Mood's test
- Spatial rank score: Multivariate extension of classical Kruskal-Wallis test (Wilcoxon rank-sum test).
- Note that no inner centering is needed for the rank scores. The tests based on spatial signs and ranks are not invariant - they can again made invariant by using inner standardization.

Multivariate regression: Testing problem I

- Consider the linear regression problem where

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

  where $\mathbf{X}$ is a full-rank matrix of q explaining variables and $\boldsymbol{\beta}$ is the $q \times p$ matrix of regression coefficients. We first wish to test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, that is, IID vs given structure.

- We assume that $\frac{1}{n}\mathbf{X}'\mathbf{X} \to \mathbf{D}$.

- Write

$$\mathbf{T}_i = \mathbf{T}(\mathbf{y}_i) \text{ and } \mathbf{T}_i(\boldsymbol{\beta}) = \mathbf{T}(\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i), \quad i = 1, ..., n$$

  and

$$\mathbf{T} = (\mathbf{T}_1, ..., \mathbf{T}_n)' \text{ and } \mathbf{T}(\boldsymbol{\beta}) = (\mathbf{T}_1(\boldsymbol{\beta}), ..., \mathbf{T}_n(\boldsymbol{\beta}))'.$$

- The test statistic (starting from $n^{-1/2}\mathbf{T}'\mathbf{X}$)

$$Q^2 = n \cdot tr\left((\mathbf{T}'\mathbf{P}_\mathbf{X}\mathbf{T})(\mathbf{T}'\mathbf{T})^{-1}\right)$$

  has, under the null hypothesis a limiting χ^2_{pq} distribution.

Multivariate regression: Estimation problem

- Consider the linear regression problem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}.$$

  We first wish to estimate unknown $q \times p$ matrix $\boldsymbol{\beta}$,

- The estimate $\hat{\boldsymbol{\beta}}$ based on score function $\mathbf{T}$ solves

$$\mathbf{T}(\hat{\boldsymbol{\beta}})'\mathbf{X} = \mathbf{0}$$

- The connection between the estimate and test:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{A}^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{T}(\boldsymbol{\beta})'\mathbf{X}\right)\mathbf{D}^{-1} + o_P(1)$$

- Then, under general assumptions,

$$\sqrt{n}vec(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d N_{qp}(\mathbf{0}, \mathbf{D}^{-1} \otimes \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$$

  where $\mathbf{A}$ and $\mathbf{B}$ are as before

Multivariate regression: Different estimates

- Identity score: The regular LS estimate which minimizes

$$\sum_{i=1}^{n} ||\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i||^2$$

- Spatial sign score: The multivariate least absolute deviation (LAD) estimate which minimizes

$$\sum_{i=1}^{n} ||\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i||$$

- Spatial rank score: The multivariate mean difference (MD) estimate which minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{n} ||(\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i) - (\mathbf{y}_j - \boldsymbol{\beta}'\mathbf{x}_j)||$$

Multivariate regression: Equivariance of the estimate

- The estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y})$ should have the following equivariance properties.

  1. regression equivariance:

$$\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{XH} + \mathbf{Y}) = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}) + \mathbf{H} \text{ for all full-rank matrices } \mathbf{H}$$

  2. $\mathbf{Y}$ equivariance:

$$\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{YW}) = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y})\mathbf{W} \text{ for all full-rank matrices } \mathbf{W}$$

  3. $\mathbf{X}$ equivariance:

$$\hat{\boldsymbol{\beta}}(\mathbf{XV}, \mathbf{Y}) = \mathbf{V}^{-1}\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}) \text{ for all full-rank matrices } \mathbf{V}$$

- The regular LS estimate is equivariant. The LAD and MD estimate can be made affine equivariant using transformation-retransformation (TR) technique.

Algorithms for the LAD estimate

- An iteration step for an algorithm for the LAD estimate updates the residuals and the estimate as follows

  1.
  $$\mathbf{e}_i \;\leftarrow\; y_i - \boldsymbol{\beta}'\mathbf{x}_i, \;\; i = 1, ..., n$$

  2.
  $$\boldsymbol{\beta} \;\leftarrow\; \boldsymbol{\beta} + \left[\text{ave}\{||\mathbf{e}_i||^{-1}\mathbf{x}_i\mathbf{x}_i'\}\right]^{-1} \text{ave}\{\mathbf{x}_i\mathbf{U}(\mathbf{e}_i)'\}$$

- Invariant estimate: First the residuals, second the estimator matrix, and finally the residual scatter matrix are updated as follows.

  1.
  $$\mathbf{e}_i \leftarrow \mathbf{S}^{-1/2}(\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i), \;\; i = 1, ..., n$$

  2.
  $$\boldsymbol{\beta} \;\leftarrow\; \boldsymbol{\beta} + \left[\text{ave}\{||\mathbf{e}_i||^{-1}\mathbf{x}_i\mathbf{x}_i'\}\right]^{-1} \text{ave}\{\mathbf{x}_i\mathbf{U}(\mathbf{e}_i)'\}\mathbf{S}^{1/2}$$

  3.
  $$\mathbf{S} \;\leftarrow\; p\,\mathbf{S}^{1/2}\,\text{ave}\{\mathbf{U}(\mathbf{e}_i)\mathbf{U}(\mathbf{e}_i)'\}\,\mathbf{S}^{1/2}.$$

Multivariate regression: Testing problem II

- Consider next the partitioned model

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

  where $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) is a $n \times q_1$ (resp. $n \times q_2$) matrix. We wish to test the null hypothesis $H_0 : \boldsymbol{\beta}_2 = 0$.

- First inner centered scores

$$\hat{\mathbf{T}} = \mathbf{T}(\hat{\boldsymbol{\beta}}_1, \mathbf{0})$$

  such that $\hat{\mathbf{T}}'\mathbf{X}_1 = \mathbf{0}$. Write also $\hat{\mathbf{X}}_2 = (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$.

- Then the test statistic

$$Q^2 = n \cdot tr\left(\hat{\mathbf{T}}'\mathbf{P}_{\hat{\mathbf{X}}_2}\hat{\mathbf{T}}(\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\right)$$

  has an approximate chi square distribution with $q_2 p$ degrees of freedom.

The use of scatter matrices

- The scatter matrix estimates $\mathbf{S}$ based on the sign and rank scores estimate the covariance matrix (up to a constant) in the case of elliptic distribution
- Estimates $\mathbf{S}$ have good efficiency and robustness properties
- Possible uses: PCA, ICA, CCA, tests for sphericity, tests for ellipticity, etc.

Multivariate nonparametrical methods - other approaches

- Our approach (based spatial signs and ranks) uses

$$\text{ave}\left\{||\mathbf{e}_i||\right\} \quad \text{and} \quad \text{ave}\left\{||\mathbf{e}_i - \mathbf{e}_j||\right\}$$

  and produces orthogonal invariant/equivariant but not scale.

- An approach based on marginal signs and ranks uses

$$\text{ave}\left\{|e_{i1}| + ... + |e_{ip}|\right\} \quad \text{and} \quad \text{ave}\left\{|e_{i1} - e_{j1}| + ... + |e_{ip} - e_{jp}|\right\}$$

  See Puri and Sen (1971).

- Third approach based on affine equivariant signs and ranks uses

$$\text{ave}\left\{V(\mathbf{0}, \mathbf{e}_{i_1}, ..., \mathbf{e}_{i_p})\right\} \quad \text{and} \quad \text{ave}\left\{V(\mathbf{e}_{i_1}, ..., \mathbf{e}_{i_{p+1}})\right\}$$

  where

$$V(\mathbf{e}_{i_1}, ..., \mathbf{e}_{i_{p+1}}) = \text{abs}\left\{\det\begin{pmatrix} 1 & ... & 1 \\ \mathbf{e}_{i_1} & ... & \mathbf{e}_{i_{p+1}} \end{pmatrix}\right\}$$

  is the volume of the simplex with vertices $\mathbf{e}_{i_1}, ..., \mathbf{e}_{i_{p+1}}$ (multiplied by $p!$). See Oja (1999).

Software for sign and rank methods

- MINITAB software
- R-package *exactRankTests: wilcox.exact, perm.test*
- R-packages *quantreg, mblm, glmRob*
- R-functions (www.r) for rank-based analyses of linear models (Terpsta & McKean (2005), J Statist. Software, 14,7):

  `http://www.stat.wmich.edu/mckean/HMC/Rcode/`
- R-packages for multivariate sign and rank methods *ICSNP, SpatialNP, MNM*

Some references

Hettmansperger, T.P. and Aubuchon, J. (1988). Comment on 'Rank-based robust analysis of linear models. I. Exposition and review' by David Draper. *Statistical Science*, 3, 262-263.

Hettmansperger, T.P. and McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*. Great Britain: Arnold.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonp. Statistics*, 5, 201-213

Oja, H. and Randles, R. (2004). Multivariate nonparametric tests. *Statistical Science*, 19, 598605

Sirkiä, S., Taskinen, S., Nevalainen, J. and Oja, H. (2007), Multivariate nonparametrical methods based on spatial signs and ranks: the R package SpatialNP. Conditionally accepted.

Terpstra, J.T. and McKean, J.W. (2005) Rank-Based Analyses of Linear Models Using R. *J.Statist. Software*, 14,7.

Acknowledgement

- The talk is based on joint work with several people: Jyrki Möttönen, Jaakko Nevalainen, Klaus Nordhausen, Ron Randles, Seija Sirkiä, Sara Taskinen, David Tyler among others.

- Looking for new students ... contact Hannu.Oja@uta.fi

THANK YOU !!