

**Multivariate componentwise sign and rank tests using an
invariant coordinate selection**

Klaus Nordhausen

(with Hannu Oja and Dave Tyler)

July 2006

OUTLINE

- Motivation
- Preliminaries
- Invariant coordinate selection
- Power simulation results.

Motivation

Let $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ be independent p -variate observations and write

$$\mathbf{Y} = (\mathbf{y}_1 \ \mathbf{y}_2 \ \dots \ \mathbf{y}_n)$$

for the corresponding $p \times n$ *data matrix* in the one sample case.

In the two samples case, write

$$\mathbf{Y} = (\mathbf{Y}_1 \ \mathbf{Y}_2)$$

where $\mathbf{Y}_1, \mathbf{Y}_2$ are independent random samples with sample sizes n_1 and n_2 , $n = n_1 + n_2$, from p -variate distributions.

Problem: Tests based on the marginal signs and ranks in the one and two sample location problem are not invariant under affine transformations.

Transformations

- **Affine transformation**

$$Y \rightarrow AY + b\mathbf{1}',$$

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

- **Orthogonal transformation**

$$Y \rightarrow UY$$

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

- **Sign-change transformation**

$$Y \rightarrow JY$$

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

- **Permutation**

$$Y \rightarrow PY$$

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

Location and Scatter Statistics

A p -vector valued statistic $\mathbf{T} = \mathbf{T}(\mathbf{Y})$ is called a *location statistic* if it is affine equivariant, that is,

$$\mathbf{T}(\mathbf{A}\mathbf{Y} + \mathbf{b}\mathbf{1}') = \mathbf{A}\mathbf{T}(\mathbf{Y}) + \mathbf{b}$$

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

A $p \times p$ matrix $\mathbf{S} = \mathbf{S}(\mathbf{Y}) \geq 0$ is a *scatter statistic* if it is affine equivariant in the sense that

$$\mathbf{S}(\mathbf{A}\mathbf{Y} + \mathbf{b}\mathbf{1}') = \mathbf{A}\mathbf{S}(\mathbf{Y})\mathbf{A}'$$

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

Scatter Statistics w.r.t origin

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

$$S(\mathbf{A}Y\mathbf{J}) = \mathbf{A}S(Y)\mathbf{A}'$$

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

M-estimates of Location and Scatter

M-estimators for Location (\mathbf{T}) and Scatter (\mathbf{S}) satisfy the following two implicit equations:

$$\mathbf{T} = [\text{ave}[w_1(r_i)]]^{-1} \text{ave} [w_1(r_i)\mathbf{y}_i]$$

and

$$\mathbf{S} = \text{ave} [w_2(r_i)(\mathbf{y}_i - \mathbf{T})(\mathbf{y}_i - \mathbf{T})']$$

for some suitably chosen weight functions $w_1(r)$ and $w_2(r)$. The scalar r_i is the Mahalanobis distance between \mathbf{y}_i and \mathbf{T} , that is, $r_i = \|\mathbf{y}_i - \mathbf{T}\|_{\mathbf{S}}$.

Two Scatter Matrices for ICS

Tyler (NPW'05, ICORS'05) showed that two different scatter matrices $S_1 = S_1(\mathbf{Y})$ and $S_2 = S_2(\mathbf{Y})$ 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 $B = B(\mathbf{Y})$ and a diagonal matrix $D = D(\mathbf{Y})$ 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.

Result 1. *The transformation $\mathbf{Y} \rightarrow \mathbf{Z} = B(\mathbf{Y})\mathbf{Y}$ yields an invariant coordinate system in the sense that*

$$B(\mathbf{AY})(\mathbf{AY}) = \mathbf{J}B(\mathbf{Y})\mathbf{Y}$$

for some $p \times p$ sign change matrix \mathbf{J} .

Properties of ICS

- **Uniqueness**

B is unique up to sign changes of its rows.

- **Standardized**

The elements of Z are standardized with respect to S_1 , that is $S_1(Z) = I$

- **Uncorrelated**

The elements of Z are uncorrelated with respect to S_2 , that is $S_2(Z) = D$

- **Kurtosis**

The components of Z are ordered according to their kurtosis

One sample case

Let $\mathbf{Y} = (\mathbf{y}_1 \dots \mathbf{y}_n)$ be a random sample from a p -variate continuous distribution symmetric around unknown $\boldsymbol{\mu}$. We wish to test the null hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$.

For the test, let \mathbf{S}_1 and \mathbf{S}_2 be two scatter matrices with respect to the origin.

Assume also that they are invariant under permutations to the observations. Then, for $k = 1, 2$,

$$\mathbf{S}_k(\mathbf{A}\mathbf{Y}\mathbf{P}\mathbf{J}) = \mathbf{A}\mathbf{S}_k(\mathbf{Y})\mathbf{A}', \quad \forall \mathbf{A}, \mathbf{P}, \mathbf{J},$$

and therefore

$$\mathbf{B}(\mathbf{Y}\mathbf{J}\mathbf{P}) = \mathbf{B}(\mathbf{Y})$$

As, under the null hypothesis, \mathbf{Y} is a random sample from distribution symmetric around the origin, it is also true that

$$\mathbf{Z}(\mathbf{Y}) \sim \mathbf{Z}(\mathbf{Y})\mathbf{J}\mathbf{P}, \quad \forall \mathbf{J}, \mathbf{P}.$$

Clearly $\mathbf{Z} = (\mathbf{z}_1 \dots \mathbf{z}_n)$ is not a random sample any more. However, under the null hypothesis, the variables in $(\mathbf{z}_1, \dots, \mathbf{z}_n)$ are exchangeable.

One sample case cont'd

Consider the j th component of the z_i vectors. It is easy to see that

Result 2. *Under the null hypothesis, the univariate sign test statistic*

$$U_j = \sum_{i=1}^n I(z_{ji} > 0) \sim \text{Bin}(n, 0.5).$$

Thus, for all $j = 1, \dots, p$, U_j is an invariant distribution-free multivariate sign test statistic. Unfortunately, the p sign test statistics U_1, \dots, U_p are not mutually independent.

One sample case cont'd

Let next R_{ji}^+ be the rank of $|z_{ji}|$ among $|z_{j1}|, \dots, |z_{jn}|$. The univariate *Wilcoxon signed-rank test statistic*

$$W_j = \sum_{i=1}^n \text{sgn}(z_{ji}) R_{ji}^+$$

is then distribution-free as well:

Result 3. *Under the null hypothesis, the distribution of W_j is that of the one-sample Wilcoxon signed-rank test statistic.*

Two sample case

Let $\mathbf{Y} = (\mathbf{Y}_1 \mathbf{Y}_2)$ where \mathbf{Y}_1 and \mathbf{Y}_2 are independent random samples of sizes n_1 and n_2 , $n = n_1 + n_2$, from p -variate continuous distributions with cumulative density functions $F(\mathbf{y})$ and $F(\mathbf{y} - \boldsymbol{\mu})$, respectively. We wish to test the null hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$. Let $\mathbf{S}_1 = \mathbf{S}_1(\mathbf{Y})$ and $\mathbf{S}_2 = \mathbf{S}_2(\mathbf{Y})$ be two scatter matrices *calculated from the combined data set and invariant under permutations to the observations*. This is to say that, for $k = 1, 2$,

$$\mathbf{S}_k((\mathbf{A}\mathbf{Y} + \mathbf{b}\mathbf{1}')\mathbf{P}) = \mathbf{A}\mathbf{S}_k(\mathbf{Y})\mathbf{A}', \quad \forall \mathbf{A}, \mathbf{b}, \mathbf{P},$$

and $\mathbf{B}(\mathbf{Y}\mathbf{P}) = \mathbf{B}(\mathbf{Y})$. Under the null hypothesis, the combined sample $\mathbf{Y} = (\mathbf{Y}_1 \mathbf{Y}_2)$ is a random sample of size n , and

$$\mathbf{Z}(\mathbf{Y}) \sim \mathbf{Z}(\mathbf{Y})\mathbf{P}, \quad \forall \mathbf{P}.$$

Again, $\mathbf{Z} = (\mathbf{z}_1 \dots \mathbf{z}_n)$ is not a random sample but, under the null hypothesis, the variables in $(\mathbf{z}_1, \dots, \mathbf{z}_n)$ are exchangeable.

Two sample case cont'd

Affine invariant distribution-free multivariate rank tests may be constructed as follows. Let now R_{ji} be the rank of z_{ji} among z_{j1}, \dots, z_{jn} . As z_1, \dots, z_n are exchangeable,

Result 4. *Under the null hypothesis the distribution of the univariate Wilcoxon rank test statistic*

$$W_j = \sum_{i=n_1+1}^n R_{ji}$$

is that of regular two samples Wilcoxon test statistic with sample sizes n_1 and n_2 .

Testing strategies

For the testing problems several statistics are now available and the question arises how to combine them. Some possibilities are:

- (i) Using a componentwise sign test and signed rank test as described in Puri and Sen (1971) based on all p components.
- (ii) Using the same componentwise sign test and signed rank test as before but only to combine the first and last component.
- (iii) Using an exact sign test respectively a Wilcoxon signed rank test for the last component only.

However, the exact and asymptotic distributions of (i) and (ii) are still open questions. We suggest using as approximations the asymptotic distributions of Puri and Sen (1971).

Simulation set up

We performed a simulation study to evaluate the powers of the above mentioned strategies. To construct the ICS we used as S_1 the regular Covariance matrix and

$$S_2 = \frac{1}{p+2} \text{ave}[r_i^2 (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})']$$

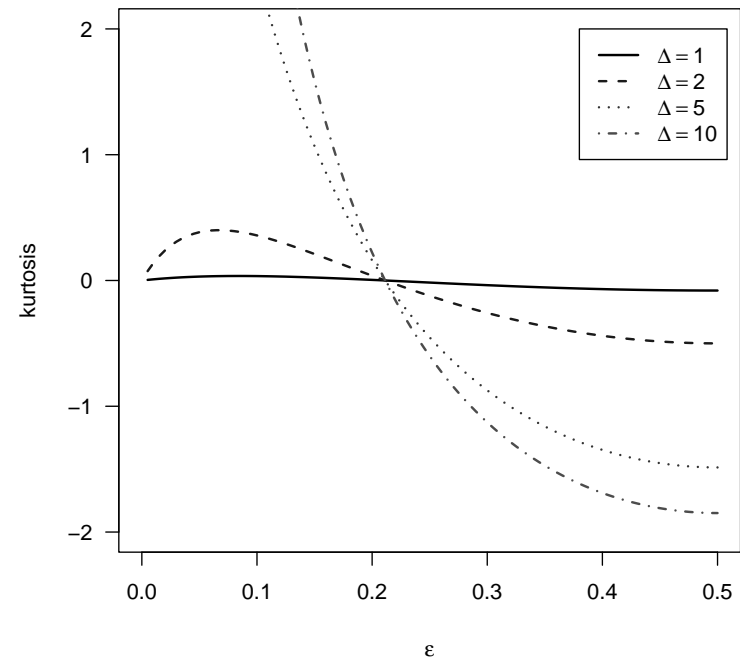
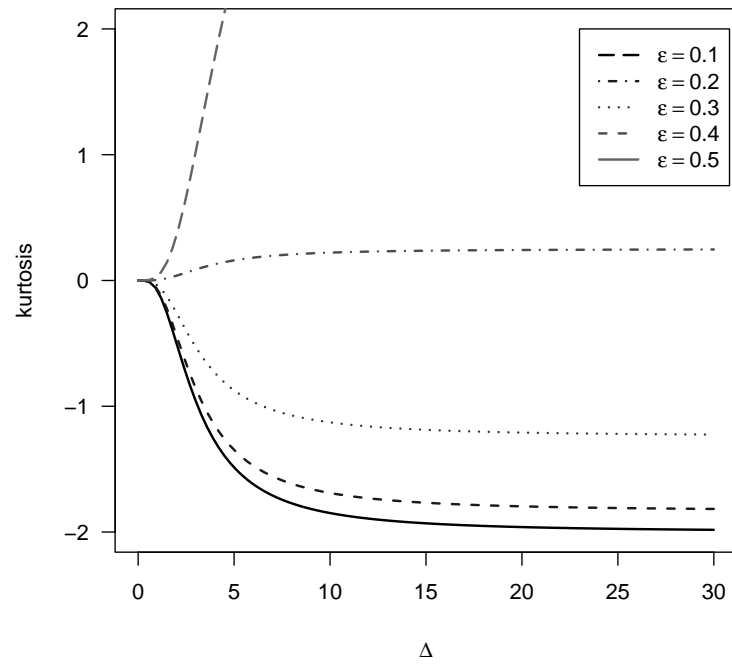
which is a one-step M-estimate of scatter, respectively a scatter matrix based on fourth moments.

The simulation data come from multivariate normal, t_{10} and t_3 distributions, where the shift Δ away from H_0 occurred only in the first coordinate and was chosen in such a way that under normality Hotelling's T^2 has power 0.5.

Simulation - relation to Kurtosis

Assuming our random sample \mathbf{Y} comes from a mixture of two multivariate normal distributions which differ only in location, that is: \mathbf{y}_i has a $N_p(\mathbf{0}, \mathbf{I})$ -distribution with probability $1 - \epsilon$ and a $N_p(\Delta \mathbf{e}_p, \mathbf{I})$ -distribution with probability ϵ ($\epsilon \leq 0.5$). Then $\mathbf{S}_1(\mathbf{Y}) \rightarrow_P \mathbf{I}$ and $\mathbf{S}_2(\mathbf{Y}) \rightarrow_P \mathbf{D}$ where \mathbf{D} is a diagonal matrix with $D_{11} = \dots = D_{p-1,p-1} = 1$. The last diagonal element is $1 + b_2/(p + 2)$ where b_2 is the *classical univariate kurtosis* measure for the last component.

Simulation - relation to Kurtosis cont'd



Simulation - One sample case

Dist.	p	n	T^2	sign tests			signed rank tests		
				$U[1 : p]$	$U[1, p]$	$U[p]$	$W[1 : p]$	$W[1, p]$	$W[p]$
normal	2	50	499	340	340	208	472	472	323
		200	502	333	333	220	479	479	327
	5	50	500	281	180	122	441	257	203
		200	508	319	197	137	472	283	213
	10	50	507	204	124	89	385	168	159
		200	503	288	140	104	458	195	152
t_3	2	50	261	286	286	180	334	334	257
		200	221	281	281	193	334	334	249
	5	50	244	237	169	117	299	200	182
		200	215	267	177	129	315	205	168
	10	50	270	173	130	95	267	155	149
		200	213	246	131	105	313	153	135

Simulation - Two sample case

Dist.	p	n_1	n_2	T^2	sign tests			signed rank tests		
					$U[1 : p]$	$U[1, p]$	$U[p]$	$W[1 : p]$	$W[1, p]$	$W[p]$
normal	2	50	50	504	321	321	177	482	482	321
		200	50	494	326	326	205	477	477	332
		200	200	494	329	329	203	477	477	317
	5	50	50	504	307	201	117	475	278	210
		200	50	491	309	191	137	464	267	199
		200	200	507	316	203	130	482	292	211
	10	50	50	499	259	136	86	449	192	159
		200	50	484	282	144	99	443	198	154
		200	200	501	212	145	92	310	199	153
t_3	2	50	50	233	277	277	160	334	334	244
		200	50	219	278	278	179	330	330	239
		200	200	214	285	285	179	336	336	246
	5	50	50	233	249	177	102	320	221	175
		200	50	213	254	181	122	321	211	164
		200	200	194	268	174	110	318	210	152
	10	50	50	230	204	123	72	296	145	132
		200	50	209	241	136	99	306	152	126
		200	200	197	183	135	84	211	149	119

Conclusions

- ICS is easier to apply than the transformation retransformation technique (Chakraborty and Chaudhuri, 1996)
- Further research necessary, for example to study the effect of different choices for the two scatter matrices