UNIVERSITY OF TAMPERE

One sample signed-rank tests in the symmetric IC model

K. Nordhausen[1], H. Oja[1] and D. Paindaveine[2]

[1]Tampere School of Public Health, University of Tampere, Finland

[2]Département de Mathématique, I.S.R.O. and E.C.A.R.E.S., Université Libre de Bruxelles, Belgium

Introduction

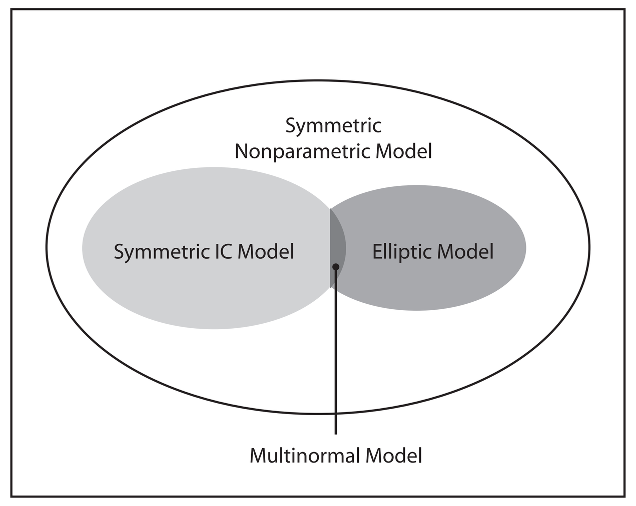
Let $X_1, ..., X_n$ be a random sample from a p-variate distribution and assume that the multivariate observations are generated by the location-scatter model

$$X_i = \Lambda Z_i + \mu, \quad i = 1, ..., n,$$

where the Z_i are *standardized variables*. The full-rank $p \times p$ matrix Λ is called the *mixing matrix*, and parameter μ is the location center. We wish to test the hypotheses

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0$$

Naturally, the test statistic one should use depends on the assumptions on the distribution of the standardized variable Z_i.
These are possible assumptions for Z_i:



- Normal Model
  $Z_i \sim N_p(0, I)$.
- Elliptic Model
  Z_i is spherical distributed around the origin with $Med(||Z_i||^2) = \chi^2_{p,.5}$.
- Symmetric IC Model
  Z_i has independent symmetric components with $Med(|Z_{ij}|^2) = \chi^2_{1,.5}$.
- Symmetric Nonparametric Model
  Z_i is symmetric around the origin ($Z_i \sim -Z_i$).

For most of the different models, the literature offers a wide variety of possible locations tests. To name only a few tests which are all asymptotically valid (after some possible adjustments) in all models:

- Hotelling's T^2, optimal in the normal model (requires however always finite 2nd moments).
- Signed-rank score tests of Hallin and Paindaveine, optimal in the elliptic model.
- Signed-rank score tests of Puri and Sen, not affine invariant and not very efficient for dependent margins.
- Spatial sign and rank tests, not affine invariant but more efficient than their marginal counterparts.
- Sign and signed-rank tests of Hettmansperger et al. based on Oja signs and ranks, affine invariant but computational expansive.

However so far nothing optimal for the IC model!

Symmetric Independent Component Model

Let μ be a p-vector, Λ be an invertible $(p \times p)$ matrix and $z = (z_1, \dots, z_p)' \mapsto g(z) = \prod_{r=1}^{p} g_r(z_r)$ be the probability density function of a p-variate symmetric random vector with independent margins.
The observations $X_1, \dots, X_n$ were generated by the assumed model:

$$X_i = \Lambda Z_i + \mu, \quad i = 1, \dots, n$$

where $Z_i = (Z_{i1}, \dots, Z_{ip})'$, $i = 1, \dots, n$ are iid with pdf g.

For a location test of the form $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ the mixing matrix Λ and the density g are nuisance parameters.

The problem of estimating Λ in this model is known as independent component analysis (ICA) and several methods are suggested in the literature. In this paper we will use the two scatter matrix method of Oja et al. (1996).

A New Test in the Symmetric IC Model

The test we are proposing has the following idea:

1. Recover the underlying components:
   Easy when Λ is know, $Z_i(\Lambda) = \Lambda^{-1} X_i$. But if unknown any estimate $\hat{\Lambda}$ that is $\sqrt{n}$-consistent and not affected under individual sign changes of observations can be used instead.
2. Compute for each component the marginal signs and ranks:
   These will be denoted by $S_i(\Lambda) = (S_{i1}(\Lambda), \dots, S_{ip}(\Lambda))'$ and $R_i(\Lambda) = (R_{i1}(\Lambda), \dots, R_{ip}(\Lambda))'$, where $R_{ir}(\Lambda)$ denotes the marginal rank of $|Z_{ir}(\Lambda)|$ among $|Z_{1r}(\Lambda)|, \dots, |Z_{nr}(\Lambda)|$.
3. Choose an appropriate score function for each component:
   The score vector is defined as a vector of $(2 + \delta)$-integrable score functions ($\delta > 0$), $K(u) = (K_1(u_1), \dots, K_p(u_p))'$.
4. Combine the marginal scores to form a test statistic.

The proposed test statistic is given by:

$$Q_K(\Lambda) = \big(T_K(\Lambda)\big)' \Gamma_K^{-1} T_K(\Lambda),$$

where

- $T_K(\Lambda) = n^{-1/2} \sum_{i=1}^n T_{K;i}(\Lambda) = n^{-1/2} \sum_{i=1}^n S_i(\Lambda) \odot K\left(\frac{R_i(\Lambda)}{n+1}\right)$
- $\Gamma_K = diag(E[(K_1(U))^2], \dots, E[(K_p(U))^2])$ is under H_0 the asymptotic covariance matrix of $T_K(\Lambda)$, when U is uniformly distributed over $(0, 1)$.
- under H_0 $Q_K(\Lambda)$ is asymptotically chi-square distributed with p degrees of freedom.
- If an estimate $\hat{\Lambda}$ is used, the statistics will be denoted correspondingly as $\hat{Q}_K, \hat{T}_K, \dots$

Important particular cases are:

- Sign test:

$$\hat{Q}_S = \hat{T}_S' \hat{T}_S = n^{-1} \sum_{i,j=1}^{n} \hat{S}_i' \hat{S}_j = n^{-1} \sum_{i,j=1}^{n} \sum_{r=1}^{p} \hat{S}_{ir} \hat{S}_{jr}.$$

- Wilcoxon type test:

$$\hat{Q}_W = 3 \hat{T}_W' \hat{T}_W = \frac{3}{n(n+1)^2} \sum_{i,j=1}^{n} \sum_{r=1}^{p} \hat{S}_{ir} \hat{S}_{jr} \hat{R}_{ir} \hat{R}_{jr}.$$

- Van der Waerden type test:

$$\hat{Q}_{vdW} = \hat{T}_{vdW}' \hat{T}_{vdW} = n^{-1} \sum_{i,j=1}^{n} \sum_{r=1}^{p} \hat{S}_{ir} \hat{S}_{jr} \Phi_+^{-1}\Big(\frac{\hat{R}_{ir}}{n+1}\Big) \Phi_+^{-1}\Big(\frac{\hat{R}_{jr}}{n+1}\Big),$$

  where $\Phi_+^{-1}(u) = \Phi^{-1}((u+1)/2)$.

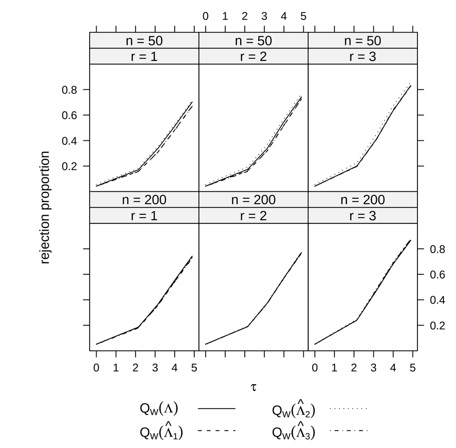
The test has the following properties:

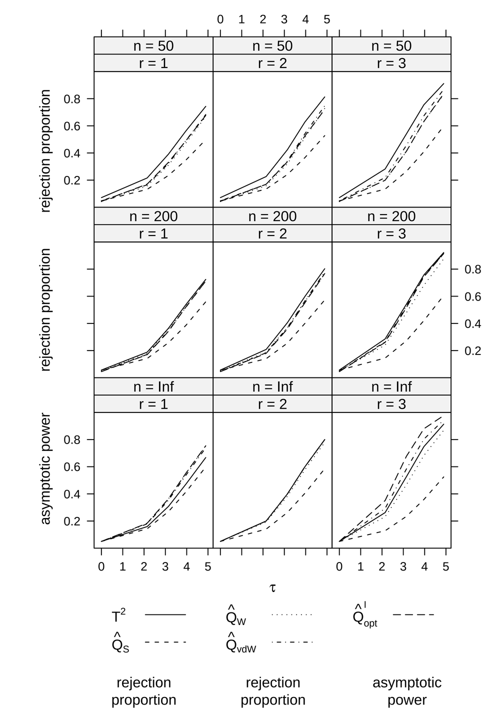
- It is affine invariant, given $\hat{\Lambda}$ is affine equivariant.
- In the case of $H_0 : \mu = \mu_0$ against $H_1(\tau) : \mu = \mu_0 + n^{-1/2}\tau$, the efficiencies of our tests compared to Hotelling's T^2 are weighted univariate efficiencies of the univariate scores compared to Student's t-test, where the weights depend on the shift τ through the "standardized" shift $\Lambda^{-1}\tau$.

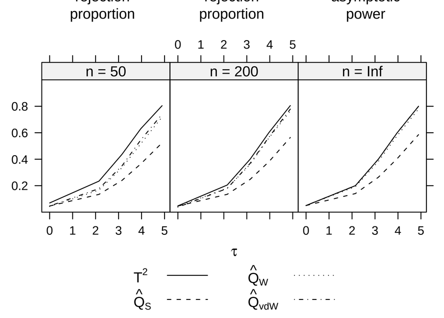
Some selected univariate efficiencies of signed-rank test compared to Student's t-test:

		underlying density						
		t_3	t_6	t_{12}	$\mathcal{N}$	e_2	e_3	e_5
	S	1.621	0.879	0.733	0.637	0.411	0.370	0.347
	W	1.900	1.164	1.033	0.955	0.873	0.881	0.907
score	vdW	1.639	1.093	1.020	1.000	1.129	1.286	1.533
	t_{12}	1.816	1.151	1.040	0.981	0.973	1.024	1.102
	t_6	1.926	1.167	1.026	0.936	0.820	0.800	0.779
	t_3	2.000	1.124	0.944	0.820	0.569	0.479	0.385

Simulation







We compared in a simulation study the efficiency of our tests in the finite sample case and compared if the estimator for Λ used has any impact. For that we set up a trivariate model $X = I_3 Z + n^{-1/2} \tau_l e_r$, where the components of Z are independent and have a t_9, $N(0, 1)$ and $PE(0, 1, 2)$ distribution (number of repetitions: 5000). The shift τ_l occurs in the direction of e_r, where e_r is the canonical basis vector in direction r. The magnitude of τ_l is such that Hotelling's T^2 would have power 0.2, 0.4, 0.6 and 0.8 in a trivariate normal model.
To compare the impact of the choice of $\hat{\Lambda}$ on the test we compare for the Wilcoxon version of the test using three different estimates which are based on the two scatter matrix approach of Oja et al. (2006) with the original Λ. The top figure on the left shows that the efficiencies are not affected by an estimate for Λ.
Next we evaluate then in this setting different choices of score functions. We compare here Hotelling's T^2 with our test using sign, Wilcoxon, van der Waerden and optimal scores. We noted that with $n = 50$ scores that assume underlying non-heavy-tailed marginal distributions converge not fast enough to the asymptotic χ^2_p distribution and therefore the critical value is here obtained by simulation from a multivariate normal distribution under H_0. Note that for $n = 50$ also Hotelling's T^2 was biased. In general shows the 2nd Figure on the left the expected behavior of our tests also in finite sample sizes.

In ICA the case with $k > 1$ Gaussian components is considered the worst case since then an estimate $\hat{\Lambda}$ can recover only the $p - k$ non-Gaussian components whereas the Gaussian components have still a random rotation and are only scaled and uncorrelated. However our test is also still asymptotically valid in this case since given such a rotation the p components still converge in distribution to Z because (possibly rotated) uncorrelated Gaussian variables with a common scale are independent. The 3rd Figure on the left shows simulation results when setting in the previous setting all components to $N(0, 1)$.

Summary

The nonparametric signed-rank score test we suggest here is affine invariant, robust and optimal in the symmetric IC model and does not depend on the choice of an estimate for the mixing matrix Λ. Furthermore does this show that Hotelling's T^2 is in the Pitman efficiency sense non-admissible in the symmetric IC model.

Key References

K. Nordhausen, H. Oja and D. Paindaveine (2008). Signed-rank tests for location in the symmetric independent component model. Submitted.

H. Oja, S. Sirkiä, and J. Eriksson (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35, 175-189.

M. L. Puri, and P. K. Sen (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley & Sons, New York.