

MULTI-SAMPLE SPATIAL MEDIAN TEST NOT REQUIRING DISTRIBUTIONS OF THE SAME TYPE

Ján Somorčík, František Rublík

Keywords: Location parameters, multi-sample comparison, spatial median.

Abstract: We present a test for the multivariate multi-sample location problem. It is based on spatial medians. Therefore, it is more robust than the classical tools. Moreover, it can be used also in situations when the underlying distributions of the samples are not of the same type. We prove some asymptotic properties and present results of a Monte Carlo study.

Abstrakt: Predstavujeme test rovnosti parametrov polohy viacerých mnohorozmených rozdelení, založený na priestorových mediánoch. Je robustnejší než klasické metódy. Navyše sa dá použiť aj v situáciách, keď rozdelenia pravdepodobnosti v súboroch nie sú rovnakého typu. V článku dokazujeme niekoľko asymptotických vlasností a uvádzame výsledky Monte Carlo simulácií.

1 Introduction

We consider the d-dimensional q-sample location problem. It means that we have q independent random samples from d-variate distributions with location parameters $\theta_1, \ldots, \theta_q$ and we wish to test the hypothesis

$$H_0 : \theta_1 = \ldots = \theta_q.$$

We suppose that the d-variate distributions possess densities w.r.t. Lebesgue measure. Typically, it is assumed that the densities are of the form $f(. - \theta_a)$ $(a = 1, \ldots, q)$. This means that the underlying distributions can differ only in location parameters. A lot of tests have been developed to test the above hypothesis, a good overview can be found in [6]. We have been motivated by the Lawley-Hotelling test based on the comparison of sample means (see e.g. [6]). Its test statistic is

$$T^2 := \sum_{a=1}^{q} n_a (\bar{X}^{(a)} - \bar{X})^T S^{-1} (\bar{X}^{(a)} - \bar{X}),$$

where n_a is the sample size of the a-th sample, $\bar{X}^{(a)}$ is the sample mean of the a-th sample. $\bar{X}$ and S denote the sample mean and the sample covariance matrix based on the pooled sample of all $n := n_1 + \ldots + n_q$ data points. Existence of finite second order moments of the underlying distributions is required. Then the asymptotic distribution of T^2 under H_0 is $\chi^2_{(q-1)d}$.

It is well-known that the performance of the Lawley-Hotelling test is rather poor (it has low power) when the underlying distributions are heavy-tailed (see e.g. [5]). The reason is that the sample covariance matrix and the sample means are very sensitive to outliers. A more robust estimate of location is, for example, the spatial median. Therefore, in [5] we have replaced the sample means by sample spatial medians to obtain a more robust test.

First, a few words about spatial median. The sample spatial median $\hat{\mu}$ of the data points $X_1, \ldots, X_n$ is defined as

$$\hat{\mu} := \arg \min_{M \in \mathbf{R}^d} \sum_{i=1}^{n} \|X_i - M\|, \tag{1}$$

where $\|.\|$ denotes the usual Euclidean norm. Uniqueness and existence of $\hat{\mu}$ is ensured unless the data points lie on a single line (see [3], one of the shortest contributions ever published in Annals of Statistics). There is no explicit formula to compute the spatial median. Hence, an iterative algorithm is needed. The most popular one seems to be the Weiszfeld's algorithm. It was developed already in 1937 and refined in [7] to ensure its convergence for an arbitrary starting point.

The sample spatial median $\hat{\mu}$ can be seen as an estimate of

$$\mu := \arg \min_{M \in \mathbf{R}^d} \mathbf{E}(\|X - M\| - \|X\|)$$

which is the spatial median of the underlying probability distribution. Now, we introduce a weak assumption from [2] about this distribution.

Assumption 1. *Let the density of the underlying distribution be bounded on every bounded subset of* $\mathbf{R}^d$.

Under Assumption 1 the sample spatial median is asymptomatically normal:

$$\sqrt{n}(\hat{\mu} - \mu) \to N_d(0, V) \quad \text{in distribution.} \tag{2}$$

See [5] for the definition of the asymptotic covariance matrix V and [2] for the proof.

(1) defines the sample spatial median as the point from which the sum of distances to the data points is minimal. The robustness of spatial median against outliers is not obvious from this definition. To see it, compute the gradient of the function $\sum_{i=1}^{n} \|X_i - M\|$ with respect to M which must be a zero vector for $M := \hat{\mu}$. It follows that the sample spatial median is such a point that the unit-length vectors pointing from $\hat{\mu}$ to the data points are balanced, i.e. their sum is a zero vector. Now, the robustness of the spatial median can be seen from Figure 1: irrespective of how far X_4 has moved to the "north-east" the spatial median of $X_1, \ldots, X_4$ does not change, whereas the sample mean "follows" X_4.

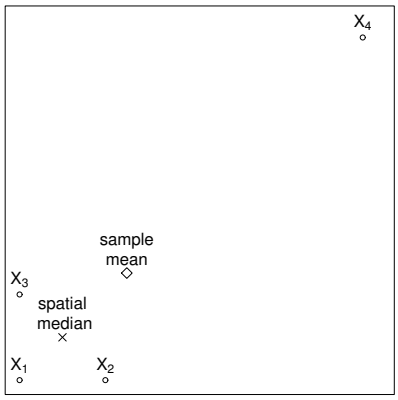

Figure 1: Robustness of spatial median.

2 Test statistics based on spatial medians

In [5] we introduced two test statistics whose form is inspired by the Lawley-Hotelling T^2:

$$M_1 := \sum_{a=1}^{q} n_a(\hat{\mu}_a - \bar{\mu})^T \hat{V}^{-1}(\hat{\mu}_a - \bar{\mu}),$$

$$M_2 := \sum_{a=1}^{q} n_a(\hat{\mu}_a - \hat{\mu})^T \hat{V}^{-1}(\hat{\mu}_a - \hat{\mu}),$$

where the $\hat{\mu}_a$'s are the sample spatial medians, $\bar{\mu} := (1/n)\sum_{a=1}^{q} n_a\hat{\mu}_a$, $\hat{\mu}$ is the sample spatial median of the pooled sample and $\hat{V}$ (see [5]) is an estimate of the asymptotic covariance matrix V of the sample spatial medians. We have shown that the asymptotic distribution of M_1 and M_2 under H_0 is $\chi^2_{(q-1)d}$. See [5] for more asymptotic properties and comparison with other test statistics.

However, there is an important deficiency concerning M_1 and M_2: they require that the underlying densities are of the form $f(.-\theta_a)$, i.e. of the same type, differing at most in location parameters. But it is natural to test H_0 also in more general situations. Think, for example, about the situation that we have random samples from d-variate spherically symmetric distributions with (possibly) different centers of symmetry $\theta_1, \ldots, \theta_q$ but also with different scatter matrices describing the variability of the spherical distributions.

This motivated us to adjust M_1 or M_2 to get rid of the assumption that the distributions of the samples must be of the same type. Our solution is as follows:

$$M_3 := \sum_{a=1}^{q} n_a (\hat{\mu}_a - \tilde{\mu})^T \hat{V}_a^{-1} (\hat{\mu}_a - \tilde{\mu}),$$

where $\hat{V}_a$ $(a = 1, \ldots, q)$ are the estimates of the asymptotic covariance matrices V_a of the sample spatial medians $\hat{\mu}_a$ and

$$\tilde{\mu} := \hat{S}^{-1} \sum_{a=1}^{q} n_a \hat{V}_a^{-1} \hat{\mu}_a$$

where $\hat{S} := \sum_{a=1}^{q} n_a \hat{V}_a^{-1}$. Hence, $\tilde{\mu}$ is a weighted average of the sample spatial medians $\hat{\mu}_a$. The impact of a particular $\hat{\mu}_a$ increases with increasing sample size n_a and decreases with "increasing" $\hat{V}_a$ because $\hat{V}_a$ measures the variability of the estimate $\hat{\mu}_a$. Our idea is not new, it was used e.g. in [4] in a different multi-sample testing problem.

Strictly speaking, the test statistics M_1, M_2 and M_3 test the equality of the spatial medians $\mu_1, \ldots, \mu_q$ of the underlying distributions. If these distributions are of the same type the equality of the μ_a's implies also the equality of the location parameters θ_a's, no matter how the term "location parameter" is defined. But if the underlying distributions are not of the same type (it is the case when M_1 and M_2 can not be used but M_3 can) the above implication is not necessarily true. However, in many practical situations each of the underlying distributions possesses certain kind of symmetry and the location parameters θ_a's are defined as the centers of these symmetries. Typically, also the spatial medians μ_a's of the distributions coincide with the centers of the symmetries. Therefore, the equality of the location parameters θ_a's is the consequence of the equality of the spatial medians μ_a's.

To establish the asymptotic distribution of M_3 we will need an assumption about the asymptotic "proportions" of the samples:

Assumption 2. *Let* $\exists p_a := \lim(n_a/n) > 0$ *for* $a = 1, \ldots, q$.

The following theorem provides a tool for testing the hypothesis H_0. Its proof and also the proofs of the following theorems can be found in the Appendix.

Theorem 1. *Let the underlying densities of the samples satisfy Assumption 1. Then under Assumption 2 the asymptotic distribution of* M_3 *under* H_0 *is* $\chi^2_{(q-1)d}$.

In [5] it was shown that in case of underlying distributions of the same type the test statistics M_1 and M_2 are asymptotically equal under H_0, i.e. $M_1 = M_2 + o_P(1)$. Now, a natural question arises about the relationship between M_3 and M_1 (or M_2). The following theorem gives the answer.

Theorem 2. *Let the distributions of the samples be of the same type (i.e. their densities are of the form $f(.-\theta_a)$). Let f satisfy Assumption 1. Then under Assumption 2: $M_3 = M_1 + o_P(1)$.*

Now, we are going to study the asymptotic performance of M_3 when H_0 is not true. Consider the sequence of Pitman alternatives, i.e. the spatial medians do not share the same value μ but the spatial median of the distribution of the a-th sample is

$$\mu + \frac{h_a}{\sqrt{n}},$$

where the h_a's are some constant vectors satisfying

$$\sum_{a=1}^{q} p_a V_a^{-1} h_a = 0, \tag{3}$$

which means that asymptotically the "shifts" of the distributions are balanced.

Theorem 3. *Under Pitman alternatives and Assumptions 1 and 2 the asymptotic distribution of M_3 is noncentral chi-squared $\chi^2_{(q-1)d}(\delta)$ where the noncentrality parameter is $\delta := \sum_{a=1}^{q} p_a h_a^T V_a^{-1} h_a$.*

Condition (3) is just technical and enables us to compare M_3 with other tests by the ratio of the noncentrality parameters. Note that in case of underlying distributions of the same type the noncentrality parameters of M_1, M_2 and M_3 are the same (cf. [5]).

3 Monte Carlo study

We have performed a simulation study to illustrate the finite sample performance of the test statistics M_1, M_2 and especially the performance of M_3. Also four other multivariate multi-sample test statistics were included in the study: Lawley-Hotelling T^2, L_N based on component-wise ranks (see [6]) and W_{ϕ_1}, W_{ϕ_2} based on spatial signs (see [5]). Only M_3 and W_{ϕ_1} do not require underlying distributions of the same type, however, W_{ϕ_1} needs spherical symmetry.

We have been generating $q = 3$ samples of $n_1 = n_2 = n_3$ data points from 3-variate distributions (i.e. $d = 3$). The first sample has been generated from $N_3(\theta_1, I_3)$, the second and third from spherically symmetric Cauchy distributions (see [6]) with centers of symmetry θ_2 and θ_3. If not stated otherwise, the location parameters θ_1, θ_2, θ_3 were set to $(0, 0, 0)^T$. We have simulated three different cases, each of them 5000-times. The 5% critical value of χ^2_6 was used to reject H_0. The results are in Table 1.

The simulated probabilities of Type I error of M_1, M_2 and M_3 are slightly higher than the nominal level 5%. In case of M_3 it is just because of too

	M_1	M_2	M_3	T^2	L_N	W_{ϕ_1}	W_{ϕ_2}
H_0 true	.060	.064	.063	.029	.049	.049	0.057
$\theta_1 = (0.3,\, 0.3,\, 0)^T$	.497	.500	.580	.038	.421	.567	.369
$\theta_2 = (0.3,\, 0.3,\, 0)^T$	.480	.489	.485	.042	.315	.456	.283

Table 1: Simulated probabilities of Type I error and powers.

small sample sizes. For M_1 and M_2 it is because of underlying distributions of different type (compare with simulated probabilities of Type I error of M_1 and M_2 proposed in [5] in case of $n_1 = n_2 = n_3 = 100$ and underlying distributions of the same type).

If the first sample is shifted (i.e. the one with lower variability) the simulated powers of M_1 and M_2 lag behind the simulated powers of M_3 and W_{ϕ_1}. However, in case of shift in the second sample (here the variability is higher) the simulated powers of M_1 and M_2 are similar to that of M_3. The presence of the heavy-tailed Cauchy distribution makes the performance of the Lawley-Hotelling T^2 really poor. Also note that the powers of L_N and W_{ϕ_2} are significantly smaller than the powers of M_3 and W_{ϕ_1}.

4 Conclusions

The test statistic M_3 based on spatial medians turns out to be a quite robust tool for testing the multivariate multi-sample location problem when the underlying distributions are not of the same type. Some non-parametric multivariate test statistics are computationally intensive and, therefore, not easy to use for larger data sets. Thanks to Weiszfeld's spatial median algorithm M_3 can be obtained quite quickly. Also its χ^2-approximation seems to work already for relatively small sample sizes.

As the simulations suggest, the violation of the assumption of underlying distributions of the same type does not necessarily mean a poor performance of tests based on that assumption. Nevertheless, M_3 ensures that the favourable properties of the spatial median test statistics M_1, M_2 remain valid also in more general situations.

Appendix

Proof of Theorem 1:

M_3 can be written in the matrix form:

$$M_3 = \begin{pmatrix} \sqrt{n_1}(\hat{\mu}_1 - \tilde{\mu}) \\ \vdots \\ \sqrt{n_q}(\hat{\mu}_q - \tilde{\mu}) \end{pmatrix}^T \underbrace{\begin{pmatrix} \hat{V}_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{V}_q^{-1} \end{pmatrix}}_{=:\hat{W}^{-1}} \underbrace{\begin{pmatrix} \sqrt{n_1}(\hat{\mu}_1 - \tilde{\mu}) \\ \vdots \\ \sqrt{n_q}(\hat{\mu}_q - \tilde{\mu}) \end{pmatrix}}_{=:Z}.$$

In [1] it was shown that $\hat{V}_a = V_a + o_P(1)$. It means that $\hat{W}^{-1} = W^{-1} + o_P(1)$, where

$$W := \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_q \end{pmatrix}.$$

Further, Z can be rewritten into the form

$$Z = \underbrace{\left[I_{qd} - \begin{pmatrix} \sqrt{\frac{n_1}{n}} I_d \\ \vdots \\ \sqrt{\frac{n_q}{n}} I_d \end{pmatrix} \left(\sqrt{\frac{n_1}{n}} n\hat{S}^{-1}, \ldots, \sqrt{\frac{n_q}{n}} n\hat{S}^{-1} \right) \hat{W}^{-1} \right]}_{=:\hat{B}} \cdot X,$$

where $X := (\sqrt{n_1}(\hat{\mu}_1 - \mu)^T, \ldots, \sqrt{n_q}(\hat{\mu}_q - \mu)^T)^T$, μ is the common value of $\mu_1, \ldots, \mu_q$. Assumption 2 and the fact that $\hat{V}_a = V_a + o_P(1)$ imply that $\hat{B} = B + o_P(1)$ where

$$B := I_{qd} - (\sqrt{p} \otimes I_d)(\sqrt{p}^T \otimes R^{-1})W^{-1},$$

$\sqrt{p} := (\sqrt{p_1}, \ldots, \sqrt{p_q})^T$, $R := \sum_{a=1}^{q} p_a V_a^{-1}$ and $\otimes$ denotes the Kronecker product. Summarizing the above asymptotic results about $\hat{W}$ and $\hat{B}$ we obtain that

$$M_3 = X^T(B+o_P(1))^T(W^{-1}+o_P(1))(B+o_P(1))X = X^T B^T W^{-1} B X + o_P(1),$$

where the second equality follows from the fact $X = O_P(1)$ (ensured by (2)). It is easy to verify that

$$(\sqrt{p}^T \otimes I_d)W^{-1}(\sqrt{p} \otimes R^{-1}) = I_d, \tag{4}$$

which implies that

$$M_3 = X^T \underbrace{[W^{-1} - W^{-1}(\sqrt{p} \otimes R^{-1})(\sqrt{p}^T \otimes I_d)W^{-1}]}_{=:A} X + o_P(1).$$

From (2) we have that asymptotically $X \sim N_{qd}(0, W)$. Hence, to show that the asymptotic distribution of the quadratic form $X^T A X$ is $\chi^2_{(q-1)d}$ it is sufficient to prove that $WAWAW = WAW$ and $trace(AW) = (q-1)d$. These two equalities can be verified by a straightforward computation making use of (4).

Proof of Theorem 2:

Let V denotes the common value of $V_1, \ldots, V_q$. Then $W = I_q \otimes V$, $R = V^{-1}$ and one easily obtains that $A = (I_q - \sqrt{p}\sqrt{p}^T) \otimes V^{-1}$. But for this form of

the matrix A it was shown (see the proofs in [5]) that $M_1 = X^T AX + o_P(1)$ and the proof is complete.

Proof of Theorem 3:

As in the proof of Theorem 1 one obtains that $M_3 = X^T AX + o_P(1)$. But here the asymptotic distribution of X is $N_{qd}(h^*, W)$, where

$$h^* := (\sqrt{p_1}h_1^T, \ldots, \sqrt{p_q}h_q^T)^T.$$

It was already shown that $WAWAW = WAW$ and $trace(AW) = (q-1)d$. To complete the proof we have to verify that $WAh^* \in \mathcal{M}(WAW)$ and $h^{*T}Ah^* = h^{*T}AWAh^*$, where $\mathcal{M}(.)$ denotes the linear subspace spanned by the columns of the matrix. The first condition is satisfied because W is regular and therefore we can write $WAh^* = WAWW^{-1}h^* \in \mathcal{M}(WAW)$. The validity of the second condition follows from the equality $AWA = A$ which can be verified by (4). The noncentrality parameter is given by $\delta = h^{*T}AWAWAh^*$. From (4) it is easy to see that $AWAWA = A$. Further, we apply (3) and obtain that $\delta = \sum_{a=1}^{q} p_a h_a^T V_a^{-1} h_a$.

References

[1] Bai Z.D., Chen X.R., Miao B.Q., Rao C.R. (1990). *Asymptotic theory of least distances estimate in multivariate linear models.* Statistics **21**, 503 – 519.

[2] Chaudhuri P. (1992). *Multivariate location estimation using extension of R-estimates through U-statistics type approach.* Ann. Statist. **20**, 897 – 916.

[3] Milasevic P., Ducharme G.R. (1987). *Uniqueness of the spatial median.* Ann. Statist. **15**, 1332 – 1333.

[4] Rublík F. (2001). *Tests of some hypotheses on characteristic roots of covariance matrices not requiring normality assumptions.* Kybernetika **37**, 61 – 78.

[5] Somorčík J. (2006). *Tests using spatial median.* Austrian Journal of Statistics **35**, 331 – 338.

[6] Um Y., Randles R.H. (1998). *Nonparametric tests for the multivariate multi-sample location problem.* Statistica Sinica **8**, 801 – 812.

[7] Vardi Y., Zhang C.H. (2000). *The multivariate L_1-median and associated data depth.* Proceedings of the National Academy of Science USA **97**, 1423 – 1426.

Acknowledgement: The work has been supported by the VEGA grant No. 1/0077/09 of the Science Grant Agency of the Slovak Republic.

Address: J. Somorčík, FMFI UK, KAMŠ, Mlynská dolina, 842 48 Bratislava; F. Rublík, ÚM SAV, Dúbravská cesta 9, 841 04 Bratislava

E-mail: somorcik@fmph.uniba.sk, umerrubl@savba.sk