Two-sample scale rank procedures optimal for the generalized secant hyperbolic distribution

O.Y. Kravchuk[*]

School of Physical Sciences, School of Land and Food Sciences,
University of Queensland, Australia

topic area: Nonparametrics; **presentation format:** Paper

Abstract

There are many linear rank tests for the two-sample dispersion problem presented in literature. However just a few of them, the simplest ones, are commonly used. These common tests are not efficient for many practical distributions and thus other simple tests need to be developed to serve a wider range of distributions. The generalized secant hyperbolic distribution, proposed by Vaughan in [9], includes a large family of symmetric heavy- and light-tailed distributions, from Cauchy to uniform. Kravchuk in [7] discussed location rank procedures optimal for this family. In the current paper, we introduce two-sample scale linear rank tests locally optimal for the generalized secant hyperbolic distribution. We discuss the asymptotic and exact properties of the new test statistics and illustrate the corresponding tests and rank estimators with numerical examples.

Key words: Generalized hyperbolic secant distribution; Cauchy distribution; Rank test; Scale parameter.

[*] Hartley Teakle Bld, University of Queensland, Brisbane, QLD, 4072, Australia; o.kravchuk@uq.edu.au; phone: (07) 33652171; fax: (07) 33651177

1 Introduction

There are many two-sample rank tests of the dispersion alternative discussed in the literature, e.g. [3]. The theory of these tests is well established and its full treatise may be found in [4] and [5]. In general, it is easy to derive an asymptotically efficient rank procedure for any continuous and symmetric distribution of a location-scale family, providing that its information numbers of the location and scale parameters are finite. However, a common drawback of the rank tests of the scale alternative is that the rate of convergence of their exact distributions to the asymptotic distributions is low under the null and specific alternative hypotheses, and one needs to rely upon the exact distributions even for reasonably large samples, as discussed in [3]. There are several tests that are exceptions from this rule and commonly used for testing the two-sample hypothesis of difference in scale; for example, such tests are the Siegel-Turkey and Sukhatme tests [3], and a rank test built on the second component of the Cramer-von Mises test [2]. The later is a test of scale asymptotically locally optimal for the Cauchy distribution. The asymptotic null-distribution of its rank statistic is standard normal; tables of the common percentiles of the exact null-distribution may be found in [2]. Although this test is simple, robust in the presence of outliers, and easy to interpret, its efficiency is very low for many practical distributions, being just above 65% for normal data and around 76% for logistic data, as demonstrated in [1].

As shown in [7], the Cauchy distribution is a limiting case of the generalized hyperbolic secant distribution (GSHD hereafter), which was introduced by Vaughan in [9]. Vaughan also showed in [9] that the GSHD includes the hyperbolic secant (HSD hereafter) and logistic distributions, the uniform distribution as a limiting case, and some Student's t distributions. The GSHD is governed by the tail-parameter t, and any distribution of this family may be denoted as GSHD(t). The family includes long-tailed distributions at $-\pi < t < 0$, the logistic distribution at $t = 0$, and short-tailed distributions at $t > 0$. Vaughan in [9] derived the inverse cumulative function of this distribution. Kravchuk in [7] discussed the general two-sample linear location rank procedures optimal for the GSHD and established the score-generating function of a rank statistic of the location alternative. These two results may be combined, as discussed in [4], to derive a linear scale rank statistic optimal for the GSHD.

It is known that a linear rank test of scale built on the score-generating function of the logistic distribution is not robust in the presence of outliers (see [4] for the score-generating function of the logistic distribution and [6] for a discussion of the robustness property of a rank test). On the other hand, a linear rank scale test optimal for the Cauchy distribution is robust. In the current paper, we introduce a general two-sample rank procedure of the scale alternative, discuss its distributional and applied properties, and suggest how to transform some tests in order to improve their robustness.

This paper is structured as follows. We introduce the score-generating function, φ_1, of the scale alternative in the *Score-generating functions* section. In the *Scale rank statistics* section, we propose the linear rank statistic, whose scores $a(\cdot)$ are associated with the score-generating function by $a(i) = \varphi_1\left(\frac{i}{N+1}\right)$, where N is the number of observations. Assuming that the location parameters are known, we construct the corresponding rank tests and estimators in the *Scale rank tests and estimators* section, and illustrate these procedures with numerical examples in *Numerical examples and discussion*. In *Conclusions*, we point out some immediate directions for further research.

2 Score generating functions

Let us consider a simple linear rank statistic S defined as

$$S = \sum_{i=1}^{N} c_i a(R_i), \tag{1}$$

where $N < \infty$, $c(\cdot)$ are certain constants, also called coefficients, and $a(\cdot)$ are scores that may be associated with a certain continuous and piecewise monotone function φ defined on $[0, 1]$ by

$$a(i) = \varphi\left(\frac{i}{N+1}\right), \quad i = 1, 2, \ldots, N. \tag{2}$$

Let us consider a two-sample problem, when the data are observed in two random samples, of sizes m and n, independently drawn from two continuous distributions. Assume that the distributions belong to the same location-scale family of continuous density f and cumulative distribution function F, and differ either in their location or scale parameters only.

When it is possible to define the following functions for $u \in [0,1]$:

$$\varphi(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \tag{3}$$

and

$$\varphi_1^*(u) = -1 + F^{-1}(u)\varphi(u), \tag{4}$$

those functions are called the location and scale score-generating functions correspondingly. Under certain conditions on the score=generating functions, [4] and [8], a rank statistic associated to one of them may be used to define a Pitman regular, asymptotically efficient linear rank estimator of location or scale parameters correspondingly.

If we additionally require that

$$\sum_{i=1}^{N} c_i = 0, \quad \text{and} \quad \sum_{i=1}^{N} c_i^2 = 1, \tag{5}$$

then the scores of a scale rank statistic may be associated with

$$\varphi_1(u) = 1 + \varphi_1^*(u) = F^{-1}(u)\varphi(u). \tag{6}$$

If, additionally, the distribution is such that its location and scale score-generating functions satisfy

$$\int_0^1 \varphi_1(u)\varphi(u)du = \int_0^1 F^{-1}(u)\varphi(u)^2 du = 0, \tag{7}$$

then any two location and scale statistics, associated with φ and φ_1, are independent under the null hypothesis that there is a difference neither in location nor scale parameters.

For the GSHD, Vaughan derived in [9] that (up to a multiplier)

$$F^{-1}(u,t) = \begin{cases} \ln\left(\frac{\sin(tu)}{\sin(t(1-u))}\right), & -\pi < t < 0, \\ \ln\left(\frac{u}{1-u}\right), & t = 0, \\ \ln\left(\frac{\sinh(tu)}{\sinh(t(1-u))}\right), & t > 0, \end{cases} \tag{8}$$

for $u \in [0,1]$.

For the GSHD, Kravchuk derived in [7] that

$$\varphi(u,t) = \begin{cases} \frac{1}{\sin t} \sin\left(t(2u-1)\right), & -\pi < t < 0, \\ 2u-1, & t = 0, \\ \frac{1}{\sinh t} \sinh\left(t(2u-1)\right), & t > 0. \end{cases} \tag{9}$$

Substituting (8) and (9) into (6), we derive the score-generating function of a scale rank statistic for the GSHD:

$$\varphi_1(u,t) = \begin{cases} \ln\left(\frac{\sin(tu)}{\sin(t(1-u))}\right) \frac{1}{\sin t} \sin\left(t(2u-1)\right), & -\pi < t < 0, \\ (2u-1)\ln\left(\frac{u}{1-u}\right), & t = 0, \\ \ln\left(\frac{\sinh(tu)}{\sinh(t(1-u))}\right) \frac{1}{\sinh t} \sinh\left(t(2u-1)\right), & t > 0. \end{cases} \tag{10}$$

The scale score-generating function of the GSHD is plotted in Figure 1, where the standard normal distribution is approximated with the GSHD(2.54) as discussed in [7]. One can see that these scores, all but those of the Cauchy distribution, lead to non-robust tests - even a small number of outliers, being at the smallest or largest ranks, may cause an incorrect rejection of the null hypothesis.

3 Scale rank statistics

Let us introduce a general scale rank statistic (1) whose scores are associated with (10) by (2), and whose coefiicients are constrained by (5):

$$a(i,t) = \begin{cases} \cos\left(2\pi i/(N+1)\right), & t = -\pi, \\ \ln\left(\frac{\sin(ti/(N+1))}{\sin(t(1-i/(N+1)))}\right) \frac{1}{\sin t} \sin\left(t(2i/(N+1)-1)\right), & -\pi < t < 0, \\ (2i/(N+1)-1)\ln\left(\frac{2i/(N+1)}{1-i/(N+1)}\right), & t = 0, \\ \ln\left(\frac{\sinh(ti/(N+1))}{\sinh(t(1-i/(N+1)))}\right) \frac{1}{\sinh t} \sinh\left(t(2i/(N+1)-1)\right), & t > 0. \end{cases} \tag{11}$$

Under the null hypothesis of no difference in the scale (and location) parameters, the asymptotic null-distribution of such a statistic is normal, $Normal(0, \int\limits_0^1 \varphi_1^2(u)du)$. However, the rate

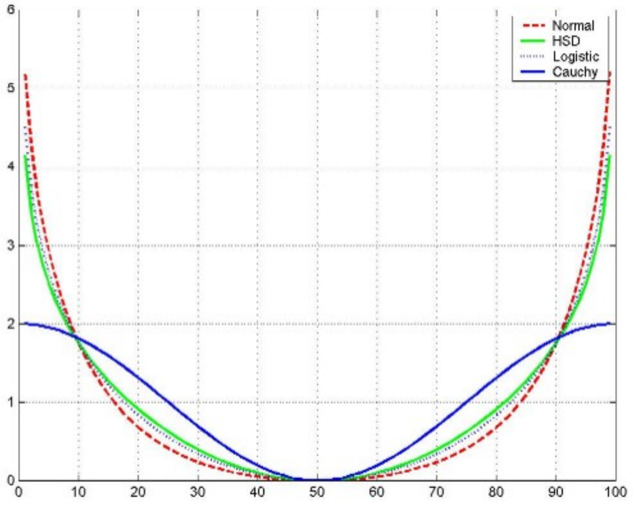

Figure 1: The scale score-generating functions for some diffuse-tailed members of the GSHD.

of convergence of the exact distribution is not fast and in practice, for some t, the approximate normal distribution with exact variance, $Normal(0,(N-1)^{-1}\sum_{i=1}^{N}a_i^2)$, is required even for reasonably large samples (up to 20 observations). The only exception is the statistic at $t=-\pi$, whose exact distribution converges quickly to its asymptotic normal distribution, which is acceptable for more than 7 observations per sample [2].

Under a specific local alternative hypothesis of a difference δ in the scale parameters of a distribution of a continuous density f, the asymptotic distribution of the linear rank statistic (1) built on (11) is normal, $Normal(\delta\int_0^1 \varphi_1(u)\varphi_1(u,f)du, \int_0^1 \varphi_1^2(u)du)$. Again, the rate of convergence is not fast and it is recommended to use the normal approximation with exact variance.

4 Scale rank tests and estimators

Locally optimal rank tests of the scale alternative and efficient estimators of the ratio of scale parameters may be established basing on the rank statistic (1) of the scores (11). A detailed discussion of how to construct a test procedure may be found in [3]; a good treatise of scale rank estimators is provided in [5]. The efficiency of several rank procedures among members of the GSHD family is presented in Figure 2. We can see that the rank procedures that correspond to $t = -\pi$ may be recommended for heavy-tailed distributions, and those that correspond to $t = 0$ - for normal-tailed distributions.

Although their optimality and efficiency properties make the rank test and estimation procedures attractive, there are several limitations that have to be taken into account when one considers using these procedures in practice. The main drawbacks are that the tests are not robust in the presence of outliers for all t but $t = -\pi$, and that the estimators are not regular in the presence of shift again for all t but $t = -\pi$. To overcome these limitations, it is suggested to trim the data whenever there is a possibility of outliers, and center the samples around zero by using the corresponding one-sample location estimators discussed in [7].

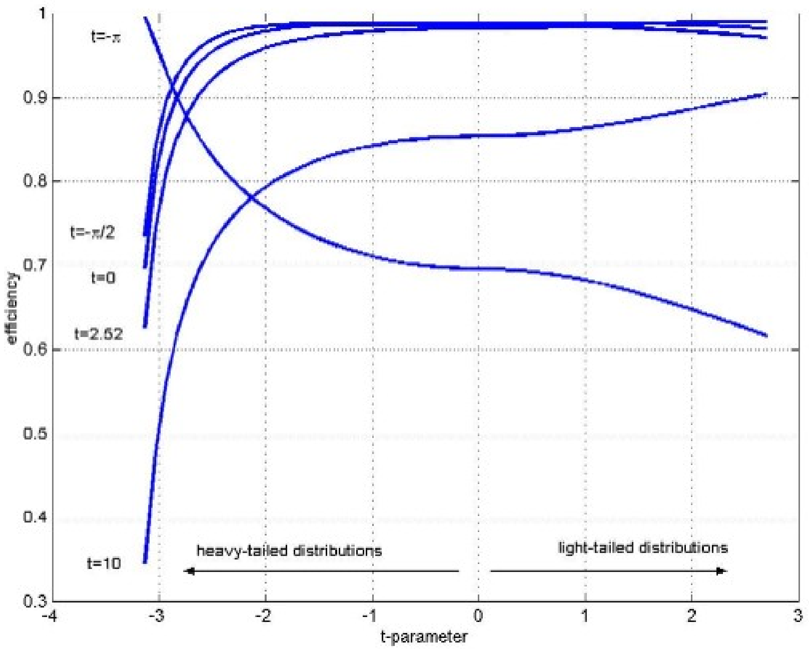

Figure 2: The scale score-generating functions for some diffuse-tailed members of the GSHD.

The test of $\varphi_1(u) = \cos(2\pi u)$ is robust and should be used for heavy-tailed distributions which are similar to the Cauchy distribution in their tails. The test of $\varphi_1(u) = \ln(u/(1-u))(2u-1)$ is efficient for normal-tailed distributions but exhibits the same over-sensitivity to outliers as do the F-test and Klotz test [3].

5 Numerical examples and discussion

Let us illustrate the test and ratio estimator of the scale parameters with several numerical examples of distributions with heavy (*Cauchy*), normal (*Logistic* and *Normal*) and light (*Uniform*) tails. The samples presented in Table 1 have been generated with Minitab 14. The Cauchy and Logistic samples were generated from the *Cauchy*(0,1) and *Cauchy*(0,3),and *Logistic*(0,1) and *Logistic(0,3)* distributions, where the parameters in brackets denote the location and scale parameters correspondingly. The Normal samples were generated from the *Normal*(0,1) and *Normal*(0,3) distributions, where the parameters correspond to the expected value and standard deviation correspondingly. The Uniform samples were generated from the *Uniform*(0,1) and *Uniform*(-1,2) distributions, where the parameters correspond to the lower and upper limits of the uniform interval.

It is known that it is hard to detect any difference in scale parameters when their ratio is smaller than 4 and the samples are not large. In our example, we can see that all the tests perform reasonably well on such a small ratio as we used to generate the samples in Table 1. As could be expected, the GSHD test performs well on the Cauchy data but fails to detect the difference for the Uniform example.

For such small samples, the interval estimates are not of much use although each of them contains the real value of the ratio of the two scale parameters. These intervals were constructed by using the normal approximation with exact variance in place of the exact null-distribution. The examples in Table 1 illustrate the general feature of the scale ratio estimation that large samples are required, in practice, to attain a reasonable accuracy of the estimation.

These examples do not contain outliers. However, it is easy to see that the Klotz test would report an unusually high significance if the data were contaminated in such a way that the narrowest sample contains the highest and the lowest values of the pooled sample. The main drawback of these scale procedures is that they are not robust in the presence of outliers. An immediate step that may be suggested to overcome this problem is to bound the score-generating functions in such a way that the resulting procedures still remain reasonably efficient for a chosen range of distributions. In particular, at $t = -\pi/2$, the score generating function may be

$$\varphi_1(u) = \begin{cases} \Big[2\left(\frac{\pi/2(u-1)-1}{\pi/2(u-1)+1} + \frac{1}{3}\left(\frac{\pi/2(u-1)-1}{\pi/2(u-1)+1}\right)^3 + \frac{1}{5}\left(\frac{\pi/2(u-1)-1}{\pi/2(u-1)+1}\right)^5\right) + \\ \frac{(\pi/2)^2}{3}(1-u)^2 + \frac{7(\pi/2)^4}{90}(1-u)^4 + \frac{62(\pi/2)^6}{2835}(1-u)^6\Big] \cos(\pi(1-u)), & u \in [0, 0.5) \\ \\ \Big[2\left(\frac{u\pi/2-1}{u\pi/2+1} + \frac{1}{3}\left(\frac{u\pi/2-1}{u\pi/2+1}\right)^3 + \frac{1}{5}\left(\frac{u\pi/2-1}{u\pi/2+1}\right)^5\right) + \\ \frac{(\pi/2)^2}{3}u^2 + \frac{7(\pi/2)^4}{90}u^4 + \frac{62(\pi/2)^6}{2835}u^6\Big] \cos(\pi u), & u \in [0.5, 1] \end{cases}$$

This score-generating function allows one to introduce a robust scale procedure whose relative efficiency varies from 87% to 96% for the heavy-tailed members of the GSHD ($-\pi < t < 0$). A similar approach has been discussed in [7] for improving the robustness of the location procedure for the Cauchy distribution.

6 Conclusions

In this paper, we have introduced several two-sample scale rank procedures efficient for the generalized hyperbolic secant distribution. These procedures allow one to work on the dispersion problem for a wide range of symmetric unimodal distributions, from heavy-tailed to light-tailed. The general scale rank procedures may be further improved by bounding their score-generating function in such a way that an increase in their robustness compensates for a slight decrease in their efficiency.

Table 1: The (statistic/p-value) of the Ansari-Bradley (I), Klotz (II), and the GSHD rank test with (11) at $t = -\pi$ (III), and its rank ratio estimators.

Sample	i	*Cauchy*$(0, 1/3)$	*Logistic*(0,1/3)	*Normal*$(0, 1/3)$	*Uniform*$((0, 1)/(-1, 3))$
1	1	-4.68	0.69	1.95	0.54
1	2	-0.16	0.48	1.51	0.05
1	3	0.18	-2.35	-0.51	0.92
1	4	1.22	-0.73	0.26	0.32
1	5	-1.27	-1.20	0.32	0.08
1	6	0.11	0.30	0.64	0.76
1	7	20.82	-2.68	-0.41	0.44
1	8	9.93	-0.81	1.92	0.43
1	9	0.48	-0.99	0.13	0.47
1	10	2.49	1.28	-2	0.96
2	11	4.17	1.15	-3.94	0.03
2	12	-1.00	1.89	-7.73	0.18
2	13	-1.87	1.84	2.06	0.83
2	14	2.15	2.24	0.65	0.10
2	15	-3.58	-7.84	0.53	0.12
2	16	4.94	2.01	0.66	0.22
2	17	-8.57	-0.15	3.81	-0.87
2	18	8.42	1.08	-3.44	1.64
2	19	-21.69	7.47	0.46	0.24
2	20	-1.64	1.47	0.10	1.92
I		4.47/0.06	3.57/0.11	3.13/0.14	4.02/0.09
II		$-0.55/0.26$	$-1.20/0.09$	$-1.48/0.05$	$-1.38/0.06$
III		$-1.58/0.06$	$-1.09/0.14$	$-1.06/0.17$	$-1.18/0.13$
Ratio point estimator		2.5	0.8	1.7	1.4
Confidence interval		$(0.40, 7.50)(94\%)$	$(0.25, 5.50)(94\%)$	$(0.11, 5.80)(94\%)$	$(0.20, 6.80)(96\%)$

The scores of the Ansari-Bradley's statistic are defined as $a(i) = ((N+1)/2 - |i - (N+1)/2|)$; the scores of the Klotz's statistic are defined as $a(i) = \left(\Phi^{-1}(i/(N+1))\right)^2$, where Φ^{-1} is the standard normal inverse cumulative function.

Acknowledgements

The author gratefully acknowledges the travel assistance provided by the University of Queensland.

References

[1] D.D. Boos. Minimum distance estimators for location and goodness of fit. *Journal of the American Statistical Association*, 76(375):663–670, 1981.

[2] J. Durbin, M. Knott, and C.C. Taylor. Components of Cramer-von Mises statistics. Part II. *Journal of the Royal Statistical Society. Series B.*, 37(2):216–237, 1975.

[3] J.D. Gibbons and S. Chakraborti. *Nonparametric statistical inference*. Statistics: Textbook and Monographs. Marcel Dekker, Inc., N.Y., 2003.

[4] J. Hájek, Z. Šidák, and P. K. Sen. *Theory of rank tests.* Academic Press, San Diego, California, 1999.

[5] T. Hettmansperger and J.W. McKean. *Robust nonparametric statistical methods.* J. Wiley & Sons, NY, 1998.

[6] T.P. Hettmansperger, J.W. Mckean, and S.J. Sheather. Robust nonparametric methods. *Journal of the American Statistical Association*, 95(452):1308–1312, 2000.

[7] O.Y. Kravchuk. R-estimator of location of the generalized secant hyperbolic distribution. *Communications in Statistics - Simulation and Computation*, 2006. to appear.

[8] J.S. Maritz. *Distribution-free statistical methods.* Monographs on Applied Probability and Statistics. Chapman and Hall, London, 1981.

[9] D.C. Vaughan. The generalized secant hyperbolic distribution and its properties. *Communications in Statistics - Theory and Methods*, 31(2):219–238, 2002.