R-ESTIMATOR OF LOCATION OF THE GENERALIZED SECANT HYPERBOLIC DISTRIBUTION

O.Y.Kravchuk

School of Physical Sciences and School of Land and Food Sciences

University of Queensland

Brisbane, Australia 3365-2171

o.kravchuk@uq.edu.au

Key Words: rank tests; rank estimators; Cauchy distribution; hyperbolic secant distribution

Mathematics Subject Classification: 62G10, 62G32

ABSTRACT

The generalized secant hyperbolic distribution (GSHD) proposed in Vaughan (2002) includes a wide range of unimodal symmetric distributions, with the Cauchy and uniform distributions being the limiting cases, and the logistic and hyperbolic secant distributions being special cases. The current paper derives an asymptotically efficient rank estimator of the location parameter of the GSHD and suggests the corresponding one and two-sample optimal rank tests. The rank estimator derived is compared to the modified MLE of location proposed in Vaughan (2002). By combining these two estimators, a computationally attractive method for constructing an exact confidence interval of the location parameter is developed. The statistical procedures introduced in the current paper are illustrated by examples.

1. INTRODUCTION

The generalized secant hyperbolic distribution (GSHD hereafter), introduced by Vaughan (2002), covers a wide range of symmetric unimodal distributions, including the hyperbolic secant distribution, the logistic distribution and, as the limiting cases, the uniform and Cauchy distributions. The location-scale family of the generalized distribution is governed by the following density

$$f(x|\mu,\sigma) = \frac{b}{2\sigma}\frac{1}{\cosh\left(\frac{x-\mu}{\sigma}\right)+a}, \tag{1}$$

where

$$\begin{aligned} a &= \cos t, \quad b = \sin t/t, \quad -\pi < t \leq 0, \\ a &= \cosh t, \quad b = \sinh t/t, \quad t > 0, \end{aligned} \tag{2}$$

and μ and σ are the location and scale parameters such that $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = (c\sigma)^2$, where

$$c = \begin{cases} \sqrt{\frac{\pi^2 - t^2}{3}}, & -\pi < t \leq 0 \\ \sqrt{\frac{\pi^2 + t^2}{3}}, & t > 0. \end{cases} \tag{3}$$

Parameter t determines the tail behaviour of the distribution as discussed in Vaughan (2002); Klein and Fischer (2003) also showed that t is the kurtosis parameter that can be used in the van Zwet's partial ordering of distributions: negative t being associated with heavier-tailed and positive t with lighter-tailed distributions, as illustrated in Figure 1.

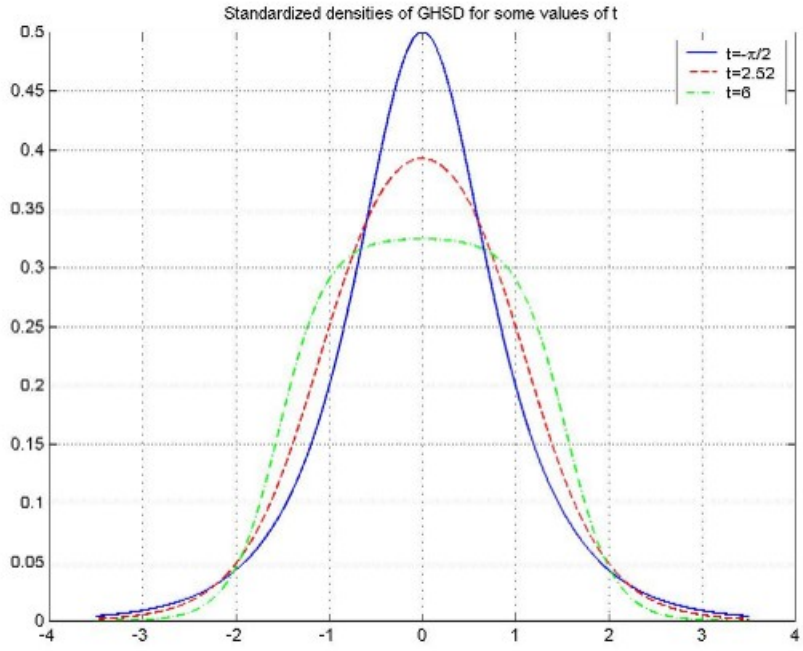

Figure 1. The standard densities of GSHD include thick-tailed, normal-tailed and thin-tailed cases.

Although the Fisher information number of the location parameter is finite for $-\pi < t < \infty$, the maximum likelihood estimator of location is not computationally feasible. Vaughan (2002) introduced the modified maximum likelihood estimator of location (MMLE hereafter) of the GSHD as a certain linear combination of the order statistics and showed that this estimator is asymptotically efficient and, for the heavy-tailed distributions, almost efficient even on small samples. Efficient estimators for the logistic and hyperbolic secant distributions can also be found among the class of rank estimators (RE hereafter). For example, the Hodges-Lehmann estimator is an asymptotically efficient RE for the logistic distribution, and the minimum distance estimator, derived by Boos (1981), can be interpreted as an efficient RE for the logistic and hyperbolic secant distributions. Jurečkova (1984) discussed many nice properties of REs: in general, REs are robust, efficient, asymptotically normally distributed and their small sample properties are easy to derive; it is not difficult to construct exact confidence intervals and assign the corresponding tests of the location alternative. The main drawback of the estimator is that the simplest computational procedure requires $O(N^2)$ steps. However, this drawback may be overcome by constructing a linearized RE. Antille (1974) calculated a linearized Hodges-Lehmann estimator in $O(N \log N)$ steps by utilizing the sample median as the consistent estimator of location. A general approach to linearized rank estimators was developed by Kraft and Van Eeden (1972).

In the current paper, we introduce an RE of location for the GSHD that is asymptotically efficient for $t \geq -\pi/2$ and highly efficient for $-\pi < t < -\pi/2$. We discuss the asymptotic and small sample properties, and construct the linearized version of the RE based on the MMLE developed in Vaughan (2002). We also suggest the corresponding locally most powerful two-sample and one-sample rank tests of the location alternative. The paper is organized in the following way: in Section 2, we discuss the efficiency of the sample mean and median of the GSHD. Section 3 and 4 discuss the two-sample and one-sample tests of location. The signed-rank estimator and the linearized rank estimator are discussed in Section 5, and Section 6

contains numerical examples of the statistical procedures introduced in this paper.

2. SAMPLE MEAN AND MEDIAN OF THE GSHD

For the purposes of the current paper, we need the following derivative

$$\frac{\partial}{\partial\mu}\log f(x|\mu)=\frac{1}{\sigma}\frac{\sinh\left(\frac{x-\mu}{\sigma}\right)}{\cosh\left(\frac{x-\mu}{\sigma}\right)+a}, \tag{4}$$

that immediately leads to the integral expression of the information number in (5). Additionally, it can be shown that $E\left[\left(\frac{\partial}{\partial\mu}\log f(X|\mu,\sigma)\right)^2\right]=-E\left[\frac{\partial^2}{\partial\mu^2}\log f(X|\mu)\right]$ for the GSHD, where the expected value of the second partial derivative of the likelihood function has been worked out in (3.4) in Vaughan (2002). Now we can express the Fisher information number of the location parameter, I_μ, in its integral as well as closed form

$$I_\mu=\frac{1}{2\sigma^2}\begin{cases}\left(1-\frac{\sin 2t}{2t}\int\limits_0^\infty\frac{dy}{(\cosh y+\cos t)^2}\right)=\frac{1-\sin 2t/(2t)}{\sin^2 t}, & -\pi<t<0,\\ \frac{2}{3}, & t=0,\\ \left(1-\frac{\sinh 2t}{2t}\int\limits_0^\infty\frac{dy}{(\cosh y+\cosh t)^2}\right)=\frac{\sinh 2t/(2t)-1}{\sinh^2 t}, & t>0.\end{cases} \tag{5}$$

Neither the sample mean nor the sample median reaches 100% efficiency as shown in Figure 2. The sample median and sample mean are equally efficient, 81.06%, at $t=-\pi/2$. This corresponds to the hyperbolic secant distribution (HSD hereafter) and makes this distribution somewhat special within the GSHD family: one can use $t=-\pi/2$ as the turning point in choosing the sample mean against the sample median. The sample mean is more than 81% efficient for $-\pi/2\leq t\leq 5.7$ and its maximal efficiency, 98.44%, corresponds to $t=2.521$ (Vaughan (2002) gave a different value of $t=\pi\sqrt{21/31}\approx 2.586$ when comparing the GSH with a specific Student's t distribution). The sample median is more than 81% efficient for $-\pi<t\leq-\pi/2$ and its maximal efficiency, 87.20%, is achieved at $t=-2.587$.

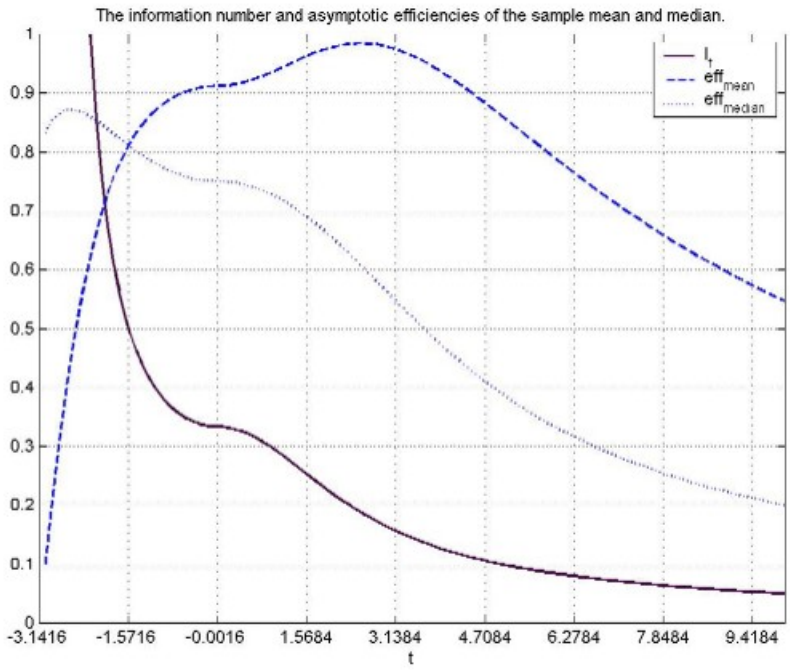

Figure 2. The information number and asymptotic efficiencies of the sample mean and sample median for the GHSD.

Since the GSHD includes smooth, continuous, unimodal, symmetric distributions, one may speculate about the limiting behaviour of the family by referring to the efficiencies of its sample mean and median. From one side, as $t \to \infty$, the efficiencies of the sample mean and the sample median tend to zero and a uniform-like behaviour may be expected. From the other side, as $t \to -\pi$, the efficiency of the sample mean tends to zero while the sample median is still a good estimator, therefore, we would expect a Cauchy-like behaviour at this limit. These speculations are justified by the following derivations (the argument related to the Cauchy distribution has been suggested by the referee):

$$\lim_{t\to\infty} \frac{1}{2}\frac{c}{t}\frac{\sinh t}{\cosh t}\frac{1}{1+\frac{\cosh cx}{\cosh t}} \approx \frac{1}{2\sqrt{3}} \lim_{t\to\infty} \frac{1}{1+\frac{\cosh xt/\sqrt{3}}{\cosh t}} = \text{Uniform}\left(-\sqrt{3},\sqrt{3}\right),$$

where c is given by (3).

Let us introduce a variable $Y = X/c^2$, where the distribution function of X is given by (1) and c is given by (3), then $\mathrm{Var}(Y) = (\sigma/c)^2$ and $\lim_{t\to-\pi} \mathrm{Var}(Y) = \infty$. For the simplicity of arguments, let $\sigma = 1$ and $\mu = 0$ in (1). The cumulative distribution function of Y is thus (see (1.5) in Vaughan (2002))

$$F(y|t) = 1 + \frac{1}{t}\cot^{-1}\left(\frac{\exp(c^2y) + \cos(t)}{\sin(t)}\right) = 1 + \frac{\pi}{2t} + \frac{1}{\pi}\tan^{-1}\left(\frac{\exp(c^2y) + \cos(t)}{\sin(t)}\right),$$

and

$$\lim_{t\to-\pi} F(y|t) = 0.5 + \frac{1}{\pi}\tan^{-1}\left(\frac{2\pi y}{3}\right) = \mathrm{Cauchy}\left(0, \frac{3}{2\pi}\right).$$

We can see that the sample mean and median, two convenient estimators of location, are not always good choices for the GSHD. As an immediate option, one would think of deriving the maximum likelihood location estimator. As Chuiv and Sinha (1998) pointed out, for the logistic distribution the maximum likelihood estimators of μ and σ are extremely difficult to calculate and their small sample properties are unknown at the moment. Vaughan (2002) demonstrated that there are computational difficulties in calculating the MLE for the GSHD and proposed a modified maximum likelihood location estimator (MMLE) based on the order statistics of a random sample of size N. Vaughan (2002) stated that for the thick-tailed distributions (in our understanding, $t < -\pi/2$) the estimator is almost fully efficient even when the sample size is small. Computing the MMLE is not expensive: it requires estimating several converging integrals and rapidly converging infinite sums, and is especially attractive for the logistic distribution whose moments of the order statistics can be computed iteratively (Shah, 1970). Conditional on σ known, the MMLE is unbiased, efficient and asymptotically normally distributed, $\mathrm{Normal}(\mu, (NI_\mu(t))^{-1})$. The small sample properties of the estimator are not immediately available. Using Student's t distribution for small samples, proposed by Vaughan (2002), still implies the asymptotic normality of the estimator for $-\pi < t \leq \pi$, and is not helpful in constructing a confidence interval for $t > \pi$. This problem is eliminated in an exact distribution-free rank estimator.

Rank estimators are exact distribution free and it is easy to construct an exact confidence interval for a small sample. The Hodges-Lehmann estimator, median $\left(\frac{x^{(i)}+x^{(j)}}{2}\right)\Big|_{i\leq j}$, is a widely used RE whose exact interval estimator is constructed on the basis of the Wilcoxon signed rank statistic. Inspired by the simplicity of the Hodges-Lehmann estimator, in the current paper, we derive an efficient RE of the GSHD, construct an exact confidence interval and suggest the corresponding one and two-sample tests of location. The following section introduces the two-sample rank test of location.

3. TWO-SAMPLE LOCATION INFERENCE

Let us consider a two-sample location problem. There are two samples of size m and n, $m+n=N$, collected in order to draw inference about the difference in location of two continuous distributions, both of which belong to the same location-scale family and may differ in their location parameter, μ_1, μ_2, only. Let Δ_μ denote the difference between the location parameters, $\Delta_\mu = \mu_1 - \mu_2$. The null hypothesis of no difference in location is to be tested with the help of the following two-sample simple linear rank statistic:

$$S_2 = \sum_{i=1}^{N} c_i a_2(R_i), \tag{7}$$

where the following constants are assigned to the pooled sample of the observations

$$c_i = \begin{cases} \frac{1}{m}\sqrt{\frac{mn}{m+n}}, & 1 \leq i \leq m, \\ -\frac{1}{n}\sqrt{\frac{mn}{m+n}}, & m < i \leq N, \end{cases} \tag{8}$$

so that $\sum\limits_{i=1}^{N} c_i = 0$ and $\sum\limits_{i=1}^{N} c_i^2 = 1$.

As shown by Hajek et al. (1999), Chapter 6, a rank test built on (7) is asymptotically optimal for a distribution F of a symmetric, continuous and differentiable density f when the scores $a_2(\cdot)$ satisfy

$$a_2(i) = N \int\limits_{(i-1)/N}^{i/N} \varphi(u)\mathrm{d}u, \quad 1 \leq i \leq N, \tag{9}$$

where φ is the following square integrable function

$$\varphi(u,f) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \quad 0 < u < 1,$$

which can be expressed as a finite sum of monotonic functions.

The score-generating function, φ, of the GSHD is derived by substituting the inverse cumulative distribution function, given in Vaughan (2002) into (4):

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

Notice that the score-generating function (10) is monotonic for $t \geq -\pi/2$, and may be expressed as the sum of two monotonic functions for $-\pi < t < -\pi/2$. Nondecreasing score-function allows one to construct a Pitman regular rank estimator of the difference in location. Therefore, in the following derivations we will consider the case $t \geq -\pi/2$ first and then show how to modify the score-generating function (10) to obtain a Pitman regular statistic for $-\pi < t < -\pi/2$.

The optimal, in the sense of (9), score function is

$$a_2(i,t) = \frac{1}{2}\begin{cases} \frac{1}{\sin t}\frac{\sin(t/N)}{t/N}\sin\left(t(\frac{2i-1}{N}-1)\right), & -\pi/2 \leq t < 0, \\ \frac{2i-1}{N}-1, & t = 0, \\ \frac{1}{\sinh t}\frac{\sinh(t/N)}{t/N}\sinh\left(t(\frac{2i-1}{N}-1)\right), & t > 0. \end{cases} \tag{11}$$

With the constants (8) and the scores (11), the expected value of the statistic (7) is zero, $\mathrm{E}(S_2) = 0$. The choice (11) leads to $\bar{a}_2 = N^{-1}\sum_{i=1}^{N} a_2(i) = 0$, and $\mathrm{var}(S_2) = (N-1)^{-1}\sum_{i=1}^{N} a_2^2(i)$ is expressed as follows

$$\mathrm{var}(S_2) = \frac{1}{2}\begin{cases} \frac{N}{N-1}\frac{1}{\sin^2 t}\left(1-\frac{\sin 2t}{2t}\frac{2t/N}{\sin 2t/N}\right)\left(\frac{\sin t/N}{t/N}\right)^2, & -\pi/2 \leq t < 0, \\ \frac{2}{3}\frac{N+1}{N}, & t = 0, \\ \frac{N}{N-1}\frac{1}{\sinh^2 t}\left(\frac{\sinh 2t}{2t}\frac{2t/N}{\sinh 2t/N}-1\right)\left(\frac{\sinh t/N}{t/N}\right)^2, & t > 0. \end{cases} \tag{12}$$

As N increases, the variance of the statistic (7) quickly converges to its asymptotic variance (5), I_μ, with $\sigma^2 = 1$. The rate of convergence is illustrated in Table I for several

values of t.

Table I: Exact and asymptotic variance of the statistic (7)

N	$t=-\pi/2$	$t=0$	$t=2.52$	$t=6$
6	0.5864	0.3889	0.2113	0.1523
20	0.5252	0.3500	0.1966	0.0852
60	0.5084	0.3389	0.1909	0.0845
∞	0.5000	0.3333	0.1879	0.0833

The statistic (7) may also be expressed in terms of the anti-ranks, D, of the pooled observations as shown below.

$$S_2 = \frac{1}{N}\sum_{i=1}^{N} b\left(\frac{i}{N}\right) Z_i, \quad Z_i = \sum_{j=R_i}^{N} c_{D_j}, \quad u \in [0,1], \tag{12}$$

where $c(\cdot)$ are given by (8). This dual form of the statistic (7) is useful in computing its exact distribution. When the score-generating function is differentiable, the statistic (7) may be thought of as a functional on the random process Z constructed on the ordered observations. The following choice of b for GSHD

$$b(u) = 2\begin{cases} \frac{t}{\sin t}\cos\left(t(2u-1)\right), & -\pi/2 \le t < 0, \\ 1, & t = 0, \\ \frac{t}{\sinh t}\cosh\left(t(2u-1)\right), & t > 0, \end{cases}$$

satisfies (9) as discussed in Kravchuk (2005).

Under the null hypothesis of no difference in location, statistic S_2, in both its forms, (7) and (12), is asymptotically normal, Normal$(0, I_\mu(t))$, with I_μ given by (5).

For $t = 0$, the statistic (7) is a linear transformation of the Wilcoxon two-sample statistic (or the first component of the Anderson-Darling statistic as shown in Pettitt (1976)). For $t = -\pi/2$, (7) is the first component of the Cramer-von Mises two-sample statistic and has been studied in detail in Kravchuk (2005). For $t = 2.52$, the test is a competitor to the Student's t and van der Waerden tests with the ARE of 98.4% on the normal data. For

any t, the sampling distribution of the statistic (7) may be constructed by the enumeration method for small samples ($N < 15$) and the normal approximation may be used for moderate and large samples ($N > 30$).

When the score-generating function is nondecreasing, the statistic (7) is a regular estimating function of the shift parameter. A rank-estimator of the difference between the location parameters, $\hat{\Delta}_\mu$, may be derived in the manner discussed in Hajek et al. (1999), Chapter 9, as

$$\hat{\Delta}_\mu = \arg\min_{\Delta} |S_2(\Delta)| \,. \tag{13}$$

The confidence interval, $\left(\hat{\Delta}_{\mu,L}, \hat{\Delta}_{\mu,R}\right)$ at a given significance level α satisfies $P(S_2(\hat{\Delta}_{\mu,L}) < S_2 < S(\hat{\Delta}_{\mu,R})) \approx \alpha$ under the null-distribution. We will not discuss the computational procedure of the point estimate in the current paper, but will give some numerical examples of the estimator in Section 6.

Let us now briefly discuss the robustness properties of the statistic (7); the following discussion is basic and the reader is referred to Hettmansperger and McKean (1998) for more detail. For the simplicity, let $m = n$. The influence function of the statistic is bounded and the breakdown point of the estimator (13) reaches its bound, 0.25, for $t \geq -\pi/2$. The estimator (13) is thus robust and resistant. Let us consider the simplest interpretation of the rejection and acceptance breakdown points of the test: rejection breakdown is the smallest proportion of outliers in one of the samples that guarantees that the test will reject the null hypothesis when it is true; similarly, acceptance breakdown is the proportion of outliers that leads to accepting the null hypothesis when the alternative is true. We simplify the definition even further by considering the worst-case scenario for a one-side test. The closed form of the acceptance breakdown of the test, taking the asymptotic variance (5), is derived to be

$$\epsilon_a = \begin{cases} 0.25 - \frac{1}{2t} \arcsin\left(\frac{\sqrt{t(2t-\sin(2t))}}{2\sin(t/2)} \frac{Z_\alpha}{\sqrt{N}}\right), & -\pi/2 \leq t < 0, \\ 0.25 - \frac{1}{2\sqrt{3}} \frac{Z_\alpha}{\sqrt{N}}, & t = 0, \\ 0.25 + \frac{1}{2t} \text{arcsinh}\left(\frac{\sqrt{t(\sinh(2t)-2t)}}{2\sinh(t/2)} \frac{Z_\alpha}{\sqrt{N}}\right), & t > 0, \end{cases}$$

where Z_α is the standard normal $(1-\alpha)$ percentile. The worst-case rejection breakdown is $0.5-\epsilon_a$. Notice that the rank test of $t=2.52$, which is almost efficient on normal data, is robust in contrast to the van der Waerden and Student's t tests.

The Pitman asymptotic efficiency (AE hereafter) of the test built on (7) is expressed in terms of the score-function (10) and information number (5):

$$AE = \left(\int_0^1 \varphi(u,t^*)\varphi(u,t_d) \right)^2 \left(I_\mu(t^*)I_\mu(t)\right)^{-1}, \tag{14}$$

where t^* is the value of t that governs the test statistic (7).

The test is robust against misspecification of the data distribution and stays efficient within a wide range of t. This property allows us to suggest using a few key-values of the t parameter in hypothesis testing that maintain a reasonably high efficiency. In Figure 3, the efficiency of the test is plotted for the three values of t^*: $t^*=-\pi/2$, optimal for the HSD, $t^*=0$, optimal for the logistic distribution, and $t^*=2.521$, nearly optimal for the normal distribution. In practice, we would recommend to use the test of $t^*=-\pi/2$ for heavy-tailed distributions, $t^*=2.521$ for light-tailed and $t^*=0$ whenever one cannot decide upon the tail behaviour. It is interesting to notice the good performance of the Wilcoxon two-sample test, $t^*=0$, in such a wide range of the distributions: it is more than 80% efficient in $-2.5<t<4.5$.

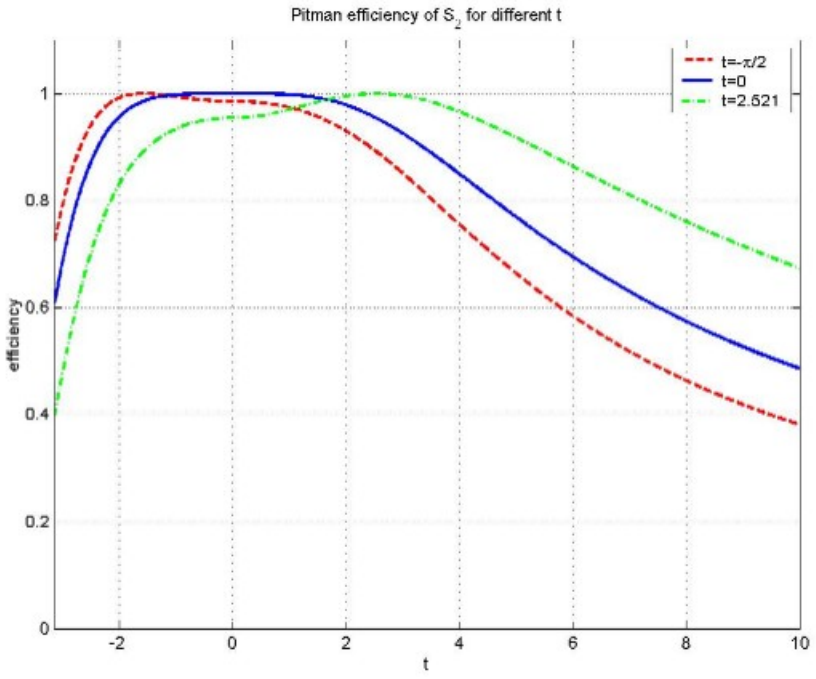

Figure 3. The Pitman asymptotic efficiencies of the test built on (7) for $t = -\pi/2$, $t = 0$ and $t = 2.521$ for various distributions of the GSHD family, $-3 < t < 10$.

Now let us consider the case $-\pi < t < -\pi/2$. To make the estimator (13) regular, we change the score-generating function to

$$\varphi^*(u) = \frac{1}{\sin t} \begin{cases} 1, & 0 < u \le \frac{\pi+2t}{4t}, \\ \sin\left(t(2u-1)\right), & \frac{\pi+2t}{4t} < u < 1 - \frac{\pi+2t}{4t}, \\ -1, & 1 - \frac{\pi+2t}{4t} \le u < 1. \end{cases} \tag{16}$$

Certainly, we lost efficiency and the estimator is not fully efficient anymore although still performs well: for $-2.8 < t < -\pi/2$ the efficiency is greater than 95% and declines to 86% for $t \to -\pi$. This loss could be tolerated in practice while the estimator is now Pitman regular, resistant and robust.

In the following section, we derive the one-sample test of location.

4. ONE-SAMPLE LOCATION INFERENCE

Let us consider the following simple linear signed rank statistic for the one-sample location problem:

$$S_1 = \sum_{i=1}^{N} a_1(R_i^+)\text{sign}\,(X_i)\,, \tag{17}$$

where $\text{sign}\,(x) = 1$ or -1 according to whether $x \geq 0$ or $x < 0$, and $R_i^+ = r_i(|X_i|)$. The sample is drawn from a continuous distribution whose density function is twice differentiable. Following Hajek et al. (1999), Chapter 4, the score-generating function of an optimal statistic (17) for the GSHD is derived as

$$\varphi_1(u,t) = \varphi\left(\frac{u+1}{2}, t\right) = \begin{cases} \frac{1}{\sin t}\sin\,(ut), & -\pi < t < 0, \\ u, & t = 0, \\ \frac{1}{\sinh t}\sinh\,(ut), & t > 0, \end{cases} \tag{18}$$

that leads to the following scores under (9):

$$a_1(i,t) = \begin{cases} \frac{1}{\sin t}\frac{\sin\,(t/(2N))}{t/(2N)}\sin\left(\frac{(2i-1)t}{2N}\right), & -\pi < t < 0, \\ \frac{2i-1}{2N}, & t = 0, \\ \frac{1}{\sinh t}\frac{\sinh\,(t/(2N))}{t/(2N)}\sinh\left(\frac{(2i-1)t}{2N}\right), & t > 0. \end{cases} \tag{19}$$

The expected value of the statistic (17) is zero, $E[S_1] = 0$, and the variance is

$$\text{var}(S_1) = \sum_{i=1}^{N} a_1^2(i) = \begin{cases} \frac{N/2}{\sin^2 t}\left(\frac{\sin t/2N}{t/2N}\right)^2\left(1 - \frac{t/N}{\sin t/N}\frac{\sin 2t}{2t}\right), & -\pi < t < 0, \\ \frac{1}{N}\left[\frac{(N^2-1)}{3} + \frac{1}{4}\right], & t = 0, \\ \frac{N/2}{\sinh^2 t}\left(\frac{\sinh t/2N}{t/2N}\right)^2\left(\frac{t/N}{\sinh t/N}\frac{\sinh 2t}{2t} - 1\right), & t > 0. \end{cases} \tag{20}$$

The statistic (17) is asymptotically normally distributed Normal$(0, NI_\mu)$ with I_μ given by (5).

For $t = 0$, the statistic (17) is a linear transformation of the Wilcoxon one-sample statistic, and for $t = -\pi/2$, it is the first component of the Cramer-von Mises one-sample statistic; the Wilcoxon statistic is extensively tabulated in the statistical literature, and tables of critical values of (17) for $t = -\pi/2$ may be found in Knott and Durbin (1972). The one-sample statistic (17) exhibits the same optimality as does the two-sample statistic (7). In

particular, the statistic (17) is a good alternative to the van der Waerden and Student's t tests at $t = 2.52$. For $t < -\pi/2$, the statistic (17) is not monotonic in shift and the nominal and actual sizes of the test may not match for small and moderate samples. For small samples, we recommend using the test optimal for the HSD for $-2.6 \leq t \leq -\pi/2$ and the median test for $-\pi < t < -2.6$. Alternatively, the score-function (18) may be modified by utilizing (16).

In practice, the following version of the signed rank statistic (17) may also be used:

$$S_1 = 0.5 \sum_{i=1}^{N} a_1(R_i^+)(\text{sign}\,(X_i) + 1),$$

whose exact null distribution for small samples ($N \leq 10$) may be calculated directly with StatXact.

In the following section we construct the corresponding point and interval estimators of the location parameter based on the statistic (17).

5. RANK ESTIMATOR OF THE LOCATION PARAMETER

Under the simple null hypothesis of location, $\mu = 0$, the signed rank statistic (17) is symmetric about 0. We define the following rank estimator of location, T:

$$T = 0.5\left(\sup\left\{\theta : S_1(X_i - \theta) > 0\right\} + \inf\left\{\theta : S_1(X_i - \theta) < 0\right\}\right), \tag{21}$$

and establish the corresponding lower and upper boundaries, L and U, of a confidence interval:

$$L = \sup\left\{\theta : P\left(S_1 \leq \sum_{i=1}^{N} a_1(R_i^+)\text{sign}\left((X_i - \theta)\right)\right) \leq \alpha/2\right\}$$
$$U = \inf\left\{\theta : P\left(S_1 \geq \sum_{i=1}^{N} a_1(R_i^+)\text{sign}\left((X_i - \theta)\right)\right) \leq \alpha/2\right\}$$

The estimator (21) is well-defined if its score-generating function, φ_1, is monotonic. This condition is satisfied for $t \geq -\pi/2$; we discuss this case first and then show how to construct a nearly efficient estimator for $-\pi < t < -\pi/2$.

For positive t, as t increases, the coefficients (19) quickly diminishes everywhere but at the few last positions, so that $\lim_{t\to\infty} \theta^s = x^{(1)}$, and $\lim_{t\to\infty} \theta^i = x^{(N)}$. Therefore, the estimator (21) for a large positive t is

$$T_{t\to\infty^+} = \frac{x^{(1)} + x^{(N)}}{2},$$

which is the MLE of location of the uniform distribution.

The test statistic (17) and the rank estimator (21) are reasonably robust against misspecification in the exact t-parameter. In practice, a threshold efficiency, γ, may be assigned in order to partition the domain of t, U_t, into $k < \infty$ disjoint intervals, $U_t = \{U_{t,i}, 1 \le i \le k\}$, and estimate the location parameter by analysing one of the k statistics defined for $t_i^*, 1 \le i \le k$:

$$U_t = \{U_{t,1}, U_{t,2}, \ldots, U_{t,k} | \mathrm{AE}(S_1(t_i^*)) \ge \gamma, t_i^* \in U_{t,i}\}.$$

For example, assigning the threshold of efficiency at 98% for (17), we partition the t-domain into the five intervals shown in Table II. Another threshold may be chosen and another strategy may be developed by referring to the efficiency of (17).

Table II: The partition of the t-domain of the GSHD for $AE \ge 98\%$ for rank tests and estimators.

t	$[-2.2, -1.1)$	$[-1.1, 2]$	$(2, 4.5]$	$(4.5, 7]$	$(7, \infty)$
t^*	$-\pi/2$	0	π	2π	2.5π

The φ_1 function is not monotonic for $t < -\pi/2$. Moreover, for $t < -\pi/2$, under the scores (19), an exact confidence interval corresponding to (21) may not exist especially for small samples, although the problem lessens as the sample size increases. A standard remedy is to adjust the score-generating function by referring to the function (16) (to bound the score-function for $t \to -\pi$, we multiply (16) by $\sin(t)$):

$$\varphi_1^*(u) = \begin{cases} \sin(ut), & 0 < u \le -\frac{1}{2}\frac{\pi}{t}, \\ 1, & -\frac{1}{2}\frac{\pi}{t} < u < 1, \end{cases} \tag{22}$$

In comparing the adjusted rank estimator to the sample median, we support the well-known argument that there is room for improvement in the estimation of the location parameters of heavy-tailed distributions. For example, Chan(1970) showed that certain quantile-based estimators substantially outperform the median for the Cauchy distribution. The estimator built on (22) is 86% efficient for Cauchy data in comparison to the 81% efficiency of the sample median.

Vaughan (2002) proposed to symmetrically disregard order statistics up to a certain term for heavy-tailed distribution, $t < -\pi/2$. Investigating the MMLE for Cauchy distribution, Vaughan (1992) built an estimator of the location parameter by utilizing approximately the middle third of the order statistics. The asymptotic efficiency of the MMLE for Cauchy distribution proposed in Vaughan (1992) is approximately 83%. This is higher than the asymptotic efficiency of the sample median and slightly lower than the efficiency of the rank estimator introduced in this paper.

The influence function for the estimator (21) is bounded and the estimator is thus robust. The asymptotic breakdown of the estimate, ϵ is given by (see Hettmansperger and McKean (1998))

$$\int_0^{1-\epsilon} \varphi_1(u)du = \frac{1}{2}\int_0^1 \varphi_1(u)du,$$

and can be solved in the closed form

$$\epsilon = \begin{cases} 1 + \frac{1}{t}\text{acos}((\cos(t)+1)/2), & -\pi/2 \le t < 0, \\ 1 - 1/\sqrt{(2)}, & t = 0, \\ 1 - \frac{1}{t}\text{acosh}((\cosh(t)+1)/2), & t > 0. \end{cases}$$

Notice that for $t > 10$ the breakdown point is almost zero.

One can see that the Hodges-Lehmann estimator, $t = 0$, and the HSD estimator, $t = -\pi/2$, cover wide intervals of the t-domain for $AE > 90\%$. The Hodges-Lehmann estimator, broadly available in the statistical software, is an efficient estimator for $-1.1 < t < 2$. There are several practically appealing computational algorithms (see for example Monahan

(1984)) that can be used for calculating the Hodges-Lehmann estimator. To simplify the computational procedure for the HSD estimator, and for any $t > -\pi/2$, $t \neq 0$, we suggest using the technique of a linearized estimate, $\hat{\mu}$, discussed for a general case in Kraft and Van Eeden (1972):

$$\hat{\mu} = \hat{\mu}_1 + \frac{\hat{\sigma}}{N I_{\mu,t^*}} S_1(t^*, \hat{\mu}_1), \tag{23}$$

where $\hat{\mu}_1$ is a consistent estimator of the location parameter such that

$$\hat{\mu}_1\left(\frac{X-\mu}{a}\right) = \frac{\hat{\mu}_1(X)-\mu}{a}, \quad \text{for all} \quad \mu \quad \text{and all} \quad a > 0,$$

the event $(\hat{\mu}_1 - \mu)\sqrt{N}$ is measurable, and $\hat{\sigma}$ is a consistent estimate of the scale parameter σ. Following Theorem 3.2 in Kraft and Van Eeden (1972), the linearized estimator $\hat{\mu}$ is asymptotically normally distributed with the moments

$$\mathrm{E}(\hat{\mu}) = \mu, \quad \mathrm{var}(\hat{\mu}) = \frac{1}{N} \frac{I_{\mu,t^*}}{\left(\int\limits_0^1 \varphi(u,t)\varphi(u,t^*)du\right)^2}, \tag{24}$$

where I_μ is given by (5) and φ by (10). The efficiency of $\hat{\mu}$ corresponds to the efficiency of the one-sample signed rank procedures. For $\hat{\mu}_1$ and $\hat{\sigma}$, we suggest using the MML location and scale estimators of the GSHD proposed by Vaughan (2002). We refer the reader to the corresponding derivations of $\hat{\mu_m}$ and $\hat{\sigma}_m$ in Vaughan (2002), page 228. For moderate sample sizes, $N > 20$, the calculations in Vaughan (2002) may be simplified by using the corresponding values of the inverse cumulative function, $F^{-1}(i/(N+1)$ in place of the exact moment of the ith order statistic (this simplification has been suggested by the referee).

The linearized estimator does not always coincide to the Hodges-Lehmann estimator but should be in the vicinity of the latter. For a non-contaminated small data set, a few additional steps may be required to calculate the estimator as t gets closer to the end of the test intervals for t^*. The secant method may be used to calculate the point estimate in a manner similar to that proposed by Antille (1974).

Next section illustrates the procedures introduced in this paper by numerical examples.

6. DISCUSSION

The discussion in this section is based on the artificial samples of 20 elements each from the logistic, hyperbolic secant, Cauchy, normal and uniform distributions (see Table III). For all the examples, the location parameter $\mu = 1.5$ and the scale parameter $\sigma = 1$. The entire samples are used for the one-sample procedures; for the two-sample procedures, the samples are halved: the first ten numbers being used as the first sample, and the last ten numbers as the second sample.

Table III: Location estimates and one-sample hypothesis testing, $H_0 : \mu = 1.5$, $H_A : \mu \neq 1.5$.

	Logistic(1.5,1)	HSD(1.5,1)	Cauchy(1.5,1)	Normal(1.5,1)	Uniform(1.5,1)
1	-1.55	-1.21	3.57	2.17	1.19
2	1.01	3.11	-4.07	2.04	1.17
3	4.19	1.51	1.66	1.48	1.87
4	1.91	-0.36	1.19	0.97	1.99
5	2.03	3.48	0.80	0.16	1.59
6	0.23	5.22	2.57	1.93	1.08
7	3.05	1.52	-1.97	0.39	1.01
8	1.16	1.19	-3.71	2.91	1.58
9	-0.36	3.18	1.32	1.86	1.64
10	0.79	2.30	3.64	1.94	1.80
11	-2.14	-0.27	0.51	0.49	1.95
12	0.59	2.25	1.14	1.89	1.37
13	1.44	5.35	1.37	2.84	2.00
14	-1.40	1.01	4.66	1.41	1.59
15	0.30	2.35	-2.09	2.89	1.39
16	4.03	3.09	6.79	1.82	1.90
17	1.78	1.62	-7.22	1.40	1.73
18	-0.72	2.56	1.47	2.60	1.34
19	1.69	1.12	1.27	1.37	1.45
20	1.34	5.74	1.49	0.68	1.48

The results of the one-sample estimation and hypothesis testing are presented in Table IV. For the Cauchy example, the truncated score-generating function (22) was used with the 86% efficiency. All the other examples were analysed with the rank procedures indicated in Table II with no less than 98% efficiency; the same t^* was also used for the corresponding MMLE of the GSHD (see Vaughan (2002)). One can see that even for these small samples, the MMLEs, $\hat{\mu}_1$, are in a good agreement to the rank estimates, T, which are reached in a few (< 4) steps from the linearized estimates, $\hat{\mu}$. As would be expected, the RE of the logistic sample coincides to its Hodges-Lehmann estimator (1.008), the RE of the HSD sample is not

far from both the sample mean (2.238) and the Hodges-Lehman estimate (2.248), the RE of the Cauchy sample is close to the sample median (1.295), the RE of the normal sample is close to the sample mean (1.662) and the RE of the uniform sample is close to the average of its extremes (1.500). The exact rank confidence intervals at the 95% confidence level are shown in Table IV together with the achieved levels. We compare these estimates to the common signed-rank, median and Student's t confidence intervals indicated in Table IV. One can see that all these intervals are similar in the case of the logistic, normal and hyperbolic secant samples. However, for the Cauchy sample, as one would expect, the RE significantly outperforms the signed-rank and Student's t intervals and does slightly better than the median estimate. As could be anticipated, the median estimate of the uniform sample is the worst among these confidence intervals. The same conclusions may be derived when one compares the p-values of the test built on (17) and the corresponding tests of location for the null hypothesis $H_0 : \mu = 1.5$ against the two-side alternative. The values of the test-statistic (17) together with the p-values are given in Table IV, the p-values being calculated under the normal distribution with the exact variance (20).

For the two-sample problem, in Table V, we present the 95% confidence intervals and the point estimates that correspond to the procedures discussed in Section 3 and to the common two-sample Studen's t and Wilcoxon procedures. For all the distributions under consideration but the Cauchy distribution, we again promote using the Wilcoxon statistic by noticing that the optimal intervals of the difference in location are similar to the Wilcoxon intervals. For the Cauchy example, the optimal procedure outperforms the Wilcoxon and Student's t associated procedures, as one would expect.

In the examples discussed here, we do not include the case of contaminated data. However, the robustness properties, discussed in Sections 3 and 4, should be taken into consideration when selecting a statistical procedure.

Table IV: Location estimates and one-sample hypothesis testing for data in Table III;
$H_0 : \mu = 1.5$, $H_A : \mu \neq 1.5$

	Logistic(1.5,1)	HSD(1.5,1)	Cauchy(1.5,1)	Normal(1.5,1)	Uniform(1.5,1)
t^*	0	$-\pi/2$	$-\pi$	π	2.5π
$\hat{\mu}_1$	0.969	2.165	1.295	1.649	1.567
$\hat{\mu}$	1.013	2.159	1.295	1.661	1.489
T	1.006	2.157	1.281	1.657	1.535
$\theta^s, S_1(\theta^s)$	1.007, -0.050	2.160, -0.04	1.281, -0.075	1.655, -0.303	1.535, -0.189
$\theta^i, S_1(\theta^i)$	1.005, 0.050	2.155, 0.150	1.280, 0.014	1.650, 0.007	1.534, 0.067
95%-CI, RE	(0.15, 1.74)	(1.30, 3.04)	(0.75, 1.55)	(1.20, 2.10)	(1.42, 1.68)
Achieved level, %	95.0	95.1	94.3	95.3	94.2
95%-CI, signed-rank	(0.15, 1.74)	(1.36, 3.15)	(-0.64, 2.35)	(1.23, 2.06)	(1.41, 1.72)
95%-CI, median	(0.25, 1.76)	(1.27, 3.11)	(0.58, 1.62)	(1.38, 2.02)	(1.38, 1.78)
95%-CI, Student's t	(0.17, 1.77)	(1.37, 3.10)	(-0.79, 2.23)	(1.28, 2.05)	(1.41, 1.70)
S_1, p-value	-3.651, 0.16	5.176, 0.11	-1.749, 0.08	1.798, 0.32	0.867, 0.44
signed-rank, p-value	0.16	0.10	0.23	0.37	0.47
sign, p-value	0.26	0.11	0.11	0.82	0.82
Student's t, p-value	0.18	0.09	0.29	0.39	0.43

Table V: Two-sample confidence interval and point estimates of the difference in location for the samples given in Table III.

Sample	RE($S_2(t^*)$), (7)	Wilcoxon two-sample	two-sample Student t
Logistic	(-1.0, 2.41), -0.45	(-1.0, 2.41), -0.45	(-1.1, 2.18), -0.55
HSD	(-2.25, 1.45), -0.13	(-2.33, 1.46), -0.12	(-2.26, 1.28), -0.49
Cauchy	(-3.82, 2.30), -0.055	(-4.22, 2.27), -0.23	(-3.80, 2.50), -0.65
Normal	(-1.05, 0.69), -0.38	(-0.98, 0.64), 0.03	(-0.95, 0.64), -0.15
Uniform	(-0.45, 0.22), -0.18	(-0.40, 0.22), -0.15	(-0.42, 0.16), -0.13

7. CONCLUSION

In this paper we have introduced and investigated the efficient rank estimator of location of the generalized secant hyperbolic distribution. The GSHD, proposed by Vaughan (2002), is interesting for a diverse range of applications. The distribution is governed by the tail

parameter t and describes a wide family of unimodal symmetrical distributions from heavy-tailed, $-\pi < t < 0$, to light-tailed, $t > 0$, which includes the Cauchy and the uniform distributions as the limiting cases. The rank estimator of location derived in the current paper is efficient, robust, asymptotically normally distributed and its exact distribution for small sample is easily available. It is easy to compute the estimator by applying the linearization technique and using the MMLE, suggested by Vaughan, as an initial consistent estimator. The rank estimator is robust against misspecification of the data distribution and remains efficient around its optimal distribution; this allows one to assign the finite set of the rank procedures for the GSHD family for any chosen efficiency. In practice, we suggest using the procedures based on $t = -\pi/2$ for heavy-tailed, $t = 2.52$ for normal and $t = 0$ for logistic-like distributions. These statistics demonstrate good efficiency and robustness in a wide interval of the t parameter; $t = 2.52$ provides a sound competitor to the Student's t procedure in the case of contaminated normal data.

The immediate continuation of this work would be to construct the rank estimator of the scale parameter and develop the corresponding location-scale rank test for the GSHD.

ACKNOWLEDGEMENT

The author would like to express her gratitude to Phil Pollett and Kaye Basford for their valuable comments. The author although thanks the anonymous referee for many constructive suggestions on this research.

BIBLIOGRAPHY

Antille, A. (1974). A linearized version of the Hodges-Lehmann estimator. *Annals of Statistics*, **2**, 1308–1313.

Boos, D.D. (1981). Minimum distance estimators for location and goodness of fit. *J. Amer. Statist. Assoc.*, **76**, 663–670.

Chan, L.K. (1970). Linear estimation of the location and scale parameters of the Cauchy distribution based on sample quantiles. *J. Amer. Statist. Assoc.*, **65**, 851–859.

Chuiv, N.N., and Sinha, B.K. (1998) Order statistics: application. *Handbook of Statistics*, **17**, pp. 364–370. (Eds. N.Balakrishan and C.R.Rao) Elsevier Science, Amsterdam.

Durbin, J., and Knott, M. (1972). Components of Cramer-von Mises statistics. Part I. *J. Roy. Statist. Soc. Ser. B*, **34**, 290–307.

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

Hettmansperger, T., and McKean, J.W. (1998). *Robust nonparametric statistical methods.* J. Wiley&Sons, NY.

Jurečkova, J. (1984). Nonparametric metods. M-, L- and R-estimators. *Handbook of Statistics*, **4**, pp. 463–487. (Eds. P.R. Krishnaiah and P.K.Sen) North-Holland, Amsterdam.

Klein, I., and Fischer, M. (2003). Kurtosis ordering of the generalized secant hyperbolic distribution - a technical note. *Diskussionpapier, Fridrich-Alexander Universitat*, **54**.

Kraft, C.H., and van Eeden, C. (1972). Linearized rank estimates and signed-rank estimates for the general linear hypothesis. *Ann. Math. Statist.*, **43**, 42–57.

Kravchuk, O.Y. (2005). Rank tests of location optimal for hyperbolic secant distribution. *Comm. Statist. Theory Methods* (to appear).

Monahan, J.F. (1984). Algorithm 616: fast computation of the Hodges-Lehmann location estimator. *ACM Transactions on Mathematical Software*, **10**, 265–270.

Pettitt, A.N. (1976). A two-sample Anderson-Darling statistic. *Biometrika*, **63**, 161–168.

Shah, B.K. (1970). Note on moments of a logistic order statistic. *Ann. Math. Statist.*, **41**, 2150–2152.

Vaughan, D.C. (1992). On the Tiku-Suresh method of estimation. *Comm. Statist. Theory Methods*, **21**, 451–469.

Vaughan, D.C. (2002). The generalized secant hyperbolic distribution and its properties. *Comm. Statist. Theory Methods*, **31**, 219–238.