Computational Statistics and Data Analysis 91 (2015) 78–91

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Efficient maximum approximated likelihood inference for Tukey's *g*-and-*h* distribution

Ganggang Xu [a,*], Marc G. Genton [b]

[a] Department of Mathematical Sciences, Binghamton University, State University of New York, Binghamton, NY 13902, USA

[b] CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Article history:
Received 18 September 2014
Received in revised form 18 May 2015
Accepted 10 June 2015
Available online 18 June 2015

Keywords:
Approximated likelihood ratio test
Computationally efficient
Maximum approximated likelihood estimator
Skewness
Tukey's *g*-and-*h* distribution

ABSTRACT

Tukey's *g*-and-*h* distribution has been a powerful tool for data exploration and modeling since its introduction. However, two long standing challenges associated with this distribution family have remained unsolved until this day: how to find an optimal estimation procedure and how to make valid statistical inference on unknown parameters. To overcome these two challenges, a computationally efficient estimation procedure based on maximizing an approximated likelihood function of Tukey's *g*-and-*h* distribution is proposed and is shown to have the same estimation efficiency as the maximum likelihood estimator under mild conditions. The asymptotic distribution of the proposed estimator is derived and a series of approximated likelihood ratio test statistics are developed to conduct hypothesis tests involving two shape parameters of Tukey's *g*-and-*h* distribution. Simulation examples and an analysis of air pollution data are used to demonstrate the effectiveness of the proposed estimation and testing procedures.

1. Introduction

Datasets with skewed and/or heavy-tailed distributions are typical in many research areas. There have been numerous attempts to search for flexible and practically useful distribution families to model such data in the statistical community; see Jones (in press) for a comprehensive review. An attractive class of distributions introduced by Tukey (1977), and later named Tukey's *g*-and-*h* distribution, has been extensively studied by many researchers; see, for example, Martinez and Iglewicz (1984), Hoaglin (1985) and Morgenthaler and Tukey (2000). Let Z be a random variable from a standard normal distribution, $N(0, 1)$. A random variable, Y, is said to have Tukey's *g*-and-*h* distribution if it is obtained through the transformation

$$Y = \xi + \omega\tau_{g,h}(Z), \tag{1}$$

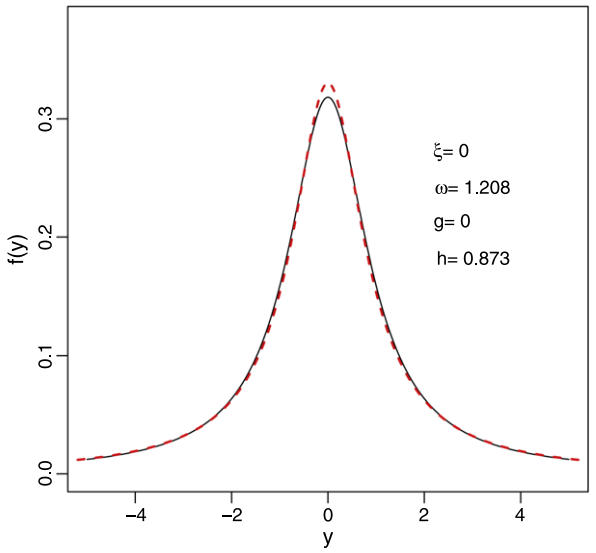
where $\xi \in \mathbb{R}$ is a location parameter, $\omega > 0$ is a scale parameter, and

$$\tau_{g,h}(z) = g^{-1}\{\exp(gz) - 1\}\exp(hz^2/2) \tag{2}$$

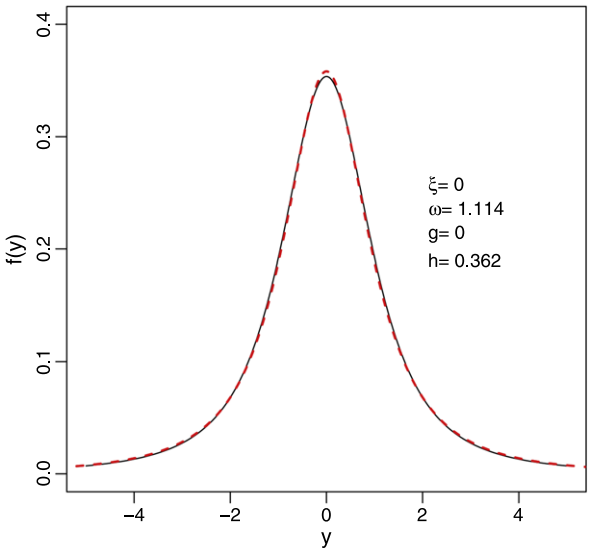
is a one-to-one monotone function of $z \in \mathbb{R}$ for $h \geq 0$, $g \in \mathbb{R}$. When $g = 0$, we use the customary definition of $\tau_{0,h}(z) = \lim_{g\to 0}\tau_{g,h}(z) = z\exp(hz^2/2)$. To simplify, from now on, values of all quantities involving the parameter g evaluated at $g = 0$ are defined as their limits attained at $g \to 0$. Aside from ξ and ω, two additional shape parameters, g and h,

* Corresponding author.
E-mail addresses: gang@math.binghamton.edu (G. Xu), marc.genton@kaust.edu.sa (M.G. Genton).

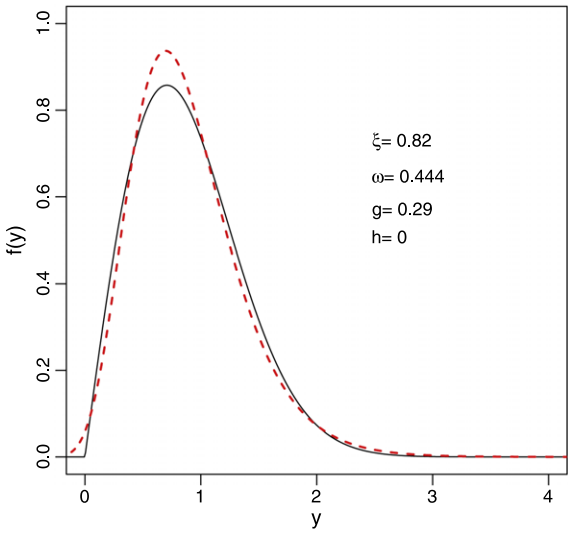
http://dx.doi.org/10.1016/j.csda.2015.06.002
0167-9473/



(a) Standard Cauchy distribution.



(b) t-distribution with $df = 2$.



(c) Weibull (2, 1) distribution.

Fig. 1. Illustrations of using Tukey's g-and-h density function (red dashed line) to approximate other density functions (black solid line): (a) Standard Cauchy distribution; (b) Student's t-distribution with degrees of freedom 2; (c) Weibull distribution with shape parameter 2 and scale parameter 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

are introduced to accommodate the potential existence of skewness and heavy-tailness in the data distribution. More precisely, $g > 0$ yields a right-skewed distribution while $g < 0$ corresponds to a left-skewed distribution. In the special case of $g = h = 0$, the resulting distribution reduces to a normal distribution with mean ξ and variance ω^2. By setting $h = 0$, one obtains a shifted log-normal distribution, whereas letting $g = 0$ gives a Pareto-like distribution. In fact, it has been shown that many commonly used distributions can be well-approximated by Tukey's g-and-h distribution (Martinez and Iglewicz, 1984; MacGillivray, 1992; Jiménez and Arunachalam, 2011). In Fig. 1, we present three examples of using Tukey's g-and-h distributions to approximate other distributions, where all parameters were estimated by the proposed estimation procedure using 10,000 random numbers generated from each distribution. As we can see, all three approximations appear to be quite good.

The flexibility of Tukey's g-and-h family makes it a powerful tool to model real data arisen from many research areas. Examples of applications include modeling short interest rate distributions (Dutta and Babbel, 2002), air pollution data (Rayner and MacGillivray, 2002a,b), extreme wind speed (Field, 2004), the value-at-risk of stock prices (Jiménez and Arunachalam, 2011), operational risk (Degen et al., 2007), and so on. Field and Genton (2006) proposed a multivariate version of Tukey's g-and-h distribution and used it to study data on Australian athletes and on wind speed. He and Raghunathan (2006, 2012) proposed to use Tukey's g-and-h distribution to perform multiple imputations for missing data. Despite its popularity, there

remain two unsolved challenges associated with Tukey's g-and-h distribution: the lack of optimal parameter estimation procedures and the lack of valid statistical inference tools. Since a small variation in g and h may result in significant changes of the shape of the distribution, an estimation procedure with high accuracy is crucial for applying Tukey's g-and-h distribution to real data. Unfortunately, the most accurate maximum likelihood estimator is not available for Tukey's g-and-h distribution because the inverse function of $\tau_{g,h}(\cdot)$ does not have a closed form unless $h = 0$, which makes the likelihood function intractable. Rayner and MacGillivray (2002a) proposed a method to numerically evaluate the log-density function of Tukey's g-and-h distribution, but their approach lacks resistance (Hoaglin, 2010) in that it can be computationally expensive when the sample size is large and it may also be numerically unstable for a reason that we will discuss later. To bypass this computational challenge, the existing literature has mainly relied on estimation procedures involving matching sample quantiles (Hoaglin, 1985; Dutta and Babbel, 2002) or sample moments (Majumder & Ali, 2008) with their population counterparts. Recently, Xu et al. (2014a) proposed a new estimation procedure named quantile least square approach to estimate the parameters. Although all these methods provide satisfactory parameter estimators in many applications, our numerical experience shows that they can be significantly less accurate than the maximum likelihood estimator. An additional problem with these methods compared to the maximum likelihood estimator is that their estimation accuracies depend on a pre-selected set of quantiles or moments, which can be subjective in practice. To the best of our knowledge, there has been no study on how to choose an "optimal" set of quantiles/moments to sharpen the estimation accuracies for these methods.

A second long-standing challenge with Tukey's g-and-h family is how to provide valid statistical inference for the parameters. Although Tukey's g-and-h distributions were first introduced as a tool to explore the data, they also have the potential to be used as an inference tool for the underlying distribution. For example, one can make statistical inference on whether the underlying distribution is symmetric by testing the hypothesis $g = 0$. While there have been numerous attempts to improve the estimation accuracy, Xu et al. (2014a,b) were the first to derive the asymptotic distribution of their estimator. However, like many other quantile-based estimators, this asymptotic distribution also depends on a subjectively selected set of quantiles and it remains unclear how will this choice affect the validity of the subsequent inference, especially when the sample size is small. Furthermore, the limiting distribution of their estimator is only valid when the true value of h, say h_0, satisfies the condition $h_0 > 0$. In the special case of $h_0 = 0$, the inference becomes irregular and the limiting distribution can be much more complicated than a normal distribution. The reason is that the restriction $h \geq 0$, which is necessary to ensure the monotonicity of $\tau_{g,h}(\cdot)$, makes $h_0 = 0$ fall on the boundary of the parameter space. Therefore, when $h_0 = 0$, the regularity conditions in Xu et al. (2014a,b) will be violated and therefore the result of Xu et al. (2014a,b) cannot be used to test hypotheses such as $h = 0$. However, because of the special interpretations of the shape parameters g and h, testing $g = 0$ or $h = 0$ may be of particular interest in many applications.

In this paper, we aim at removing the bottle-neck of sub-optimal estimation procedures and the lack of statistical inference tools for Tukey's g-and-h distribution. By approximating the likelihood function using a much simpler tractable function, we are able to obtain a maximum approximated likelihood estimator for parameters of Tukey's g-and-h distributions, which is shown to be as efficient as the true maximum likelihood estimator under mild conditions. In addition, we derive the limiting distribution of the proposed maximum approximated likelihood estimator and develop valid approximated likelihood ratio tests for a series of hypotheses for the shape parameters g and h, regardless of the true value h_0 equals 0 or not. Our simulation studies demonstrate that the proposed approach is much more efficient than the quantile-based estimators and reaches the same efficiency as that of the maximum likelihood estimator.

The rest of the paper is organized as follows. In Section 2, an efficient estimation approach based on an approximated likelihood function for Tukey's g-and-h distribution is proposed and related computational issues are discussed. The asymptotic and finite sample properties of the proposed maximum approximated likelihood estimators are investigated in Section 3. In Section 4, simulation studies are conducted to evaluate the performance of the proposed estimation procedure and approximated likelihood ratio tests. An application of our methodology to air pollution data is presented in Section 5. The article ends with a conclusion in Section 6 and all theoretical results are collected in Appendix A.

2. Parameter estimation

2.1. Existing approaches

Denote the parameter vector of Tukey's g-and-h distribution by $\boldsymbol{\theta} = (\xi, \omega, g, h)^T$. The log-density function of the random variable Y from transformation (1) can be written as

$$\log f_{Y|\boldsymbol{\theta}}(y) = \log \phi\left\{\tau_{g,h}^{-1}\left(\frac{y-\xi}{\omega}\right)\right\} - \log\omega - \log \tau'_{g,h}\left\{\tau_{g,h}^{-1}\left(\frac{y-\xi}{\omega}\right)\right\}, \tag{3}$$

where $\phi(\cdot)$ is the standard normal density function, and $\tau_{g,h}^{-1}(\cdot)$ and $\tau'_{g,h}(\cdot)$ are the inverse function and the first derivative function of $\tau_{g,h}(\cdot)$, respectively. Suppose that we have a random sample $\{y_1, \ldots, y_n\}$ and let $\mathbf{Y} = (y_1, \ldots, y_n)^T$. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{mle,n}$ is obtained by maximizing the log-likelihood function

$$L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{Y|\boldsymbol{\theta}}(y_i). \tag{4}$$

It is well known that under mild regularity conditions, the limiting distribution of $\hat{\boldsymbol{\theta}}_{mle,n}$ has the smallest variance. One can further utilize tools such as the likelihood ratio test to make statistical inference on $\boldsymbol{\theta}$. Unfortunately, since $\tau_{g,h}^{-1}(\cdot)$ does not have a closed form, numerically evaluating $L_n(\boldsymbol{\theta})$ can be computationally expensive, especially when the sample size is large. For this reason, the existing literature has largely been focusing on searching for alternative estimators, two of such examples are given below.

For a pre-selected sequence, $0 < p_1 < p_2 < \cdots < p_K < 1$, denote by $\hat{q}_{p_1}, \ldots, \hat{q}_{p_K}$ the corresponding sample quantiles of $\{y_1, \ldots, y_n\}$ and let $z_{p_k} = \Phi^{-1}(p_k)$ for $k = 1, \ldots, K$, where $\Phi^{-1}(\cdot)$ is the inverse of the $N(0, 1)$ cumulative distribution function. The first approach aims at directly matching a sequence of sample quantiles and theoretical quantiles, which we refer to as the letter-value-based approach (Dutta and Babbel, 2002). The letter-value-based estimator $\hat{\boldsymbol{\theta}}_{lv,n} = (\hat{\xi}_{lv}, \hat{\omega}_{lv}, \hat{g}_{lv}, \hat{h}_{lv})^T$ is defined as: $\hat{\xi}_{lv} = \hat{q}_{1/2}$, $\hat{g}_{lv} = \text{median}_{k=1,\ldots,K}\{\hat{g}_k\}$, where $\hat{g}_k = -\frac{1}{z_{p_k}} \log\left(\frac{\hat{q}_{1-p_k} - \hat{q}_{1/2}}{\hat{q}_{1/2} - \hat{q}_{p_k}}\right)$ and finally $(\hat{\omega}_{lv}, \hat{h}_{lv})$ are obtained from the linear regression

$$\log\left\{\frac{\hat{g}_{lv}(\hat{q}_{p_k} - \hat{q}_{1-p_k})}{\exp(\hat{g}_{lv} z_{p_k}) - \exp(-\hat{g}_{lv} z_{p_k})}\right\} = \log\omega + h z_{p_k}^2/2, \quad k = 1, \ldots, K.$$

A second approach was proposed by Xu et al. (2014a,b), named the quantile least square estimator $\hat{\boldsymbol{\theta}}_{qls,n}$, which is defined as the minimizer of the quantile least square loss function

$$L_{qls}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \left\{\hat{q}_{p_k} - q_{p_k}(\boldsymbol{\theta})\right\}^2,$$

where $q_{p_k}(\boldsymbol{\theta}) = \xi + \omega\tau_{g,h}(z_{p_k})$ is the theoretical p_kth quantile of the random variable Y.

As one can see, both $\hat{\boldsymbol{\theta}}_{lv,n}$ and $\hat{\boldsymbol{\theta}}_{qls,n}$ rely on a suitable choice of p_k's. While this choice is largely subjective in practice, there has not yet been any research on how this choice may impact the efficiency and the limiting distribution of resulting estimators.

2.2. A second representation of $\log f_{Y|\boldsymbol{\theta}}(y)$

The main difficulty for obtaining the maximum likelihood estimator lies in the lack of closed form for $\log f_{Y|\boldsymbol{\theta}}(y)$. A natural idea is to find an explicitly computable function that can approximate $\log f_{Y|\boldsymbol{\theta}}(y)$ well. To do so, we first define $p_{\boldsymbol{\theta}}(y) = F_{Y|\boldsymbol{\theta}}(y)$ and $z_{p_{\boldsymbol{\theta}}(y)} = \Phi^{-1}\{p_{\boldsymbol{\theta}}(y)\}$, where $F_{Y|\boldsymbol{\theta}}(\cdot)$ is the cumulative distribution function of Y with a parameter vector $\boldsymbol{\theta}$. By this definition, it is straightforward to show that $z_{p_{\boldsymbol{\theta}}(y)} = \tau_{g,h}^{-1}(\frac{y-\xi}{\omega})$. Then $\log f_{Y|\boldsymbol{\theta}}(y)$ in (3) can be written as a function of $z_{p_{\boldsymbol{\theta}}(y)}$; that is,

$$\begin{aligned}\varphi_{\boldsymbol{\theta}}\{z_{p_{\boldsymbol{\theta}}(y)}\} &= \log f_{Y|\boldsymbol{\theta}}(y) = \log\phi\{z_{p_{\boldsymbol{\theta}}(y)}\} - \log\omega - \log\tau'_{g,h}\{z_{p_{\boldsymbol{\theta}}(y)}\} \\ &= -\frac{1+h}{2} z_{p_{\boldsymbol{\theta}}(y)}^2 - \log\left[\exp\{g z_{p_{\boldsymbol{\theta}}(y)}\} + g^{-1}\{\exp(g z_{p_{\boldsymbol{\theta}}(y)}) - 1\} h z_{p_{\boldsymbol{\theta}}(y)}\right] - \log\omega - \frac{1}{2}\log(2\pi). \end{aligned} \tag{5}$$

The notation $z_{p_{\boldsymbol{\theta}}(y)}$ is used here to emphasize that $z_{p_{\boldsymbol{\theta}}(y)}$ is an unknown quantity depending on $\boldsymbol{\theta}$ and y. For simplicity, from now on, we write $z_{p_{\boldsymbol{\theta}}(y)}$ as z_p whenever there is no ambiguity. Recall that p is the cumulative distribution function of Y evaluated at the observed value y. Fig. 2(a) and (b) depict an example of $\log f_{Y|\boldsymbol{\theta}}(y)$, or equivalently $\varphi_{\boldsymbol{\theta}}(z_p)$, as a function y and z_p, respectively. From Fig. 2(a)–(b), we can see that the observed data $\{y_1, \ldots, y_n\}$ are sparsely distributed in $(-\infty, \infty)$. As a function of y, $\log f_{Y|\boldsymbol{\theta}}(y)$ varies rapidly on the left-hand side of the sharp spike in the middle. On the contrary, within a bounded interval, $\varphi_{\boldsymbol{\theta}}(z_p)$ is a slowly varying function of z_p, which makes it easier to approximate using some numerical method. Furthermore, although $\varphi_{\boldsymbol{\theta}}(z_p)$ is also defined on $(-\infty, \infty)$ in theory, with a finite sample size n, one only has to focus on a bounded interval $[-b_n, b_n]$ for some $b_n > 0$. To see this, consider the case when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with $\boldsymbol{\theta}_0 = (\xi_0, \omega_0, g_0, h_0)^T$ being the true value of $\boldsymbol{\theta}$ that generated the data and choose $b_n = \Phi^{-1}\{1 - 1/(n\log n)\}$. In this case, the $p_i = F_{Y|\boldsymbol{\theta}_0}(y_i)$'s follow a uniform distribution on $[0, 1]$ and the corresponding z_{p_i}'s follow the $N(0, 1)$ distribution. Straightforward algebra yields that

$$\mathrm{P}\left(\max_{1\le i\le n} z_{p_i} > b_n\right) = 1 - \Phi(b_n)^n = 1 - \left(1 - \frac{1}{n\log n}\right)^n \to 0 \quad \text{as } n \to \infty,$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. In this sense, with a sample size $n = 10{,}000$, using $b_n = 4.25$ is already a good choice. Therefore, for $\boldsymbol{\theta}$ in a sufficiently small neighborhood of $\boldsymbol{\theta}_0$ and a finite sample size n, it is reasonable to treat $\varphi_{\boldsymbol{\theta}}(z_p)$ as a function defined on a bounded support $[-b_n, b_n]$. Our numerical experience shows that it is generally sufficient to take $b_n = 10$ in practice.

2.3. Maximum approximated likelihood estimator

In this section we develop a numerical algorithm for computing the log-density value, $\log f_{Y|\boldsymbol{\theta}}(y) = \varphi_{\boldsymbol{\theta}}(z_p)$, at an observation y for a known parameter vector $\boldsymbol{\theta}$, which cannot be computed directly because the value of z_p is unknown due

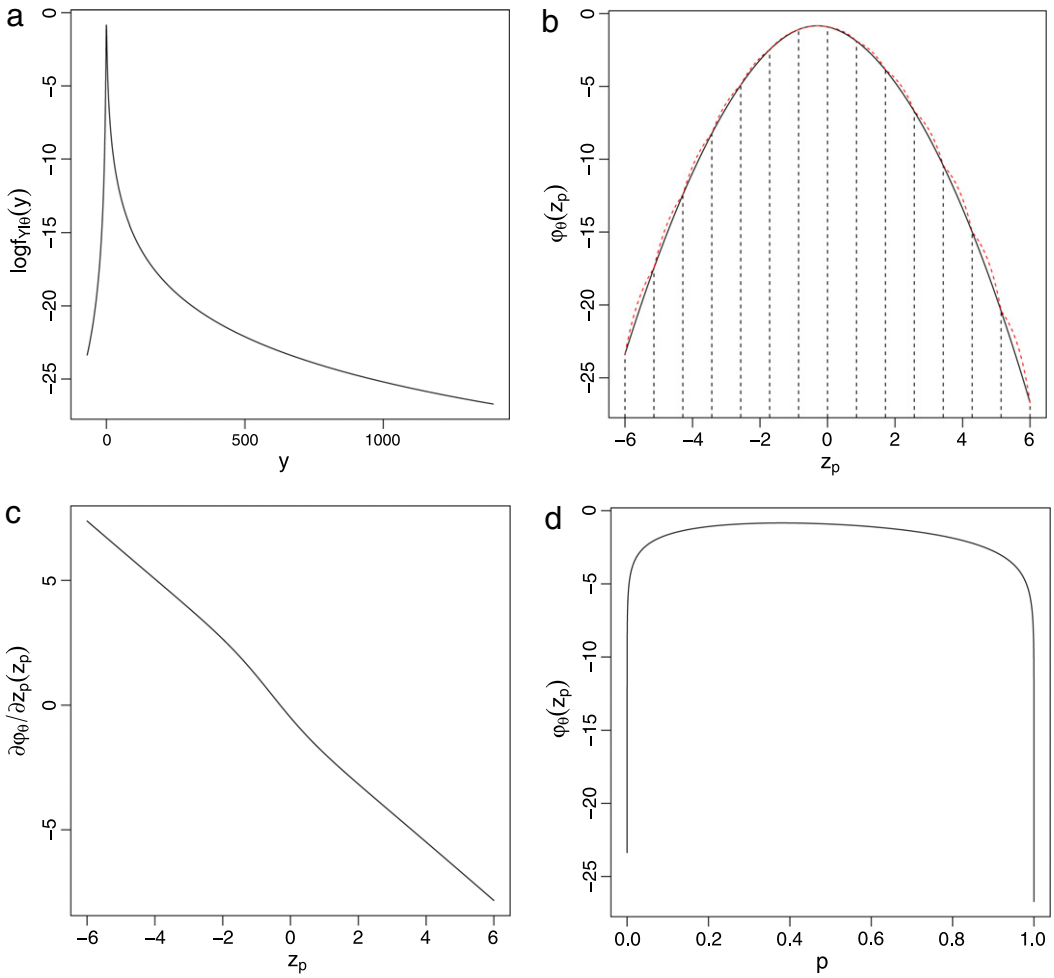


Fig. 2. The log-density function of Tukey's g-and-h distribution with $(\xi, \omega, g, h) = (0, 1, 0.5, 0.2)$. (a) $\log f_{Y|\boldsymbol{\theta}}(y)$ as a function of y; (b) $\varphi_{\boldsymbol{\theta}}(z_p)$ (solid line) vs $\tilde{\varphi}_{\boldsymbol{\theta}}(z_p)$ (dashed line, $K_n = 15$, $b_n = 6$) as functions of z_p; (c) $\frac{\partial \varphi_{\boldsymbol{\theta}}}{\partial^\dagger z_p}(z_p)$ as a function of z_p; (d) $\varphi_{\boldsymbol{\theta}}(z_p)$ as a function of p.

to the fact that $\tau_{g,h}^{-1}(\cdot)$ does not have a closed form. To overcome this, we propose the following approach to approximate the value of z_p. With a sample size n and a pre-given $b_n > 0$, we first introduce K_n equally spaced knots over the interval $[-b_n, b_n]$, denoted as $-b_n = \mathsf{Z}_1 < \mathsf{Z}_2 < \cdots < \mathsf{Z}_{K_n} = b_n$, and then compute the corresponding knots in the transformed scale as $\mathsf{Y}_{k,\boldsymbol{\theta}} = \xi + \omega\tau_{g,h}(\mathsf{Z}_k),\ k = 1, \ldots, K_n$. For a given $y \in [\mathsf{Y}_{1,\boldsymbol{\theta}}, \mathsf{Y}_{K_n,\boldsymbol{\theta}}]$, we find the knot Z_k such that $\mathsf{Y}_{k,\boldsymbol{\theta}} \le y < \mathsf{Y}_{k+1,\boldsymbol{\theta}}$. The monotonicity of $\tau_{g,h}(\cdot)$ ensures that the z_p associated with y must satisfy $\mathsf{Z}_k \le z_p < \mathsf{Z}_{k+1}$. Instead of computing z_p by numerically solving the equation $y = \xi + \omega\tau_{g,h}(z_p)$, we use the following linear approximation

$$\tilde{z}_{p,k} = \mathsf{Z}_k + \frac{y - \mathsf{Y}_{k,\boldsymbol{\theta}}}{\mathsf{Y}_{k+1,\boldsymbol{\theta}} - \mathsf{Y}_{k,\boldsymbol{\theta}}}(\mathsf{Z}_{k+1} - \mathsf{Z}_k) \quad \text{if } \mathsf{Y}_{k,\boldsymbol{\theta}} \le y < \mathsf{Y}_{k+1,\boldsymbol{\theta}}. \tag{6}$$

Because $|\tilde{z}_{p,k} - z_p| \le 2b_n/K_n$ by design, we can expect that if K_n is sufficiently large such that $b_n/K_n \to 0$, $\tilde{z}_{p,k}$ should approximate z_p well. Then, for any observed value y such that $y = \xi + \omega\tau_{g,h}(z_p)$, we can define an approximation function for $\varphi_{\boldsymbol{\theta}}(z_p)$ as

$$\tilde{\psi}_{\boldsymbol{\theta}}(y) = \tilde{\varphi}_{\boldsymbol{\theta}}(z_p) = \sum_{k=1}^{K_n-1} \varphi_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) I_{[\mathsf{Y}_{k,\boldsymbol{\theta}}, \mathsf{Y}_{k+1,\boldsymbol{\theta}}]}(y), \tag{7}$$

where $I_A(y) = 1$ if $y \in A$ and 0 otherwise. In Fig. 2(b), we can see that $\tilde{\varphi}_{\boldsymbol{\theta}}(z_p)$ is a piecewise convex function yielding a good approximation, even though only $K_n = 15$ knots were used in this example. To see why this is the case, it is straightforward to show that, for any given $\boldsymbol{\theta}$,

$$\sup_{y \in [\mathsf{Y}_{1,\boldsymbol{\theta}}, \mathsf{Y}_{K_n,\boldsymbol{\theta}}]} \left|\varphi_{\boldsymbol{\theta}}(z_p) - \tilde{\varphi}_{\boldsymbol{\theta}}(z_p)\right| \le \frac{2b_n}{K_n} \sup_{z_p \in [-b_n, b_n]} \left|\frac{\partial \varphi_{\boldsymbol{\theta}}}{\partial^\dagger z_p}(z_p)\right|,$$

where the notation $\partial^\dagger$ is used here to distinguish it from the usual partial derivative. More specifically, by using $\partial^\dagger$, we treat z_p as an argument of the function $\varphi_{\boldsymbol{\theta}}(z_p)$ that is independent of $\boldsymbol{\theta}$, even though z_p is actually a function of $\boldsymbol{\theta}$. For example,

$$\frac{\partial \varphi_{\boldsymbol{\theta}}}{\partial^\dagger z_p}(z_p) = -(1+h)z_p - g - \frac{g^{-1}\{\exp(gz_p)-1\}+z_p}{\exp(gz_p)+g^{-1}\{\exp(gz_p)-1\}hz_p}h.$$

Fig. 2(c) illustrates that for the same example used in Fig. 2(b), $\frac{\partial \varphi_{\boldsymbol{\theta}}}{\partial^\dagger z_p}(z_p)$ is essentially bounded within the interval $[-b_n, b_n]$ and therefore $\sup_{y\in[\mathsf{Y}_{1,\boldsymbol{\theta}},\mathsf{Y}_{K_n,\boldsymbol{\theta}}]}\left|\varphi_{\boldsymbol{\theta}}(z_p)-\tilde{\varphi}_{\boldsymbol{\theta}}(z_p)\right| \to 0$ as long as $b_n/K_n \to 0$ when $n \to \infty$.

Given a random sample $\{y_1, \ldots, y_n\}$, we can repeat the above process to calculate $p_i = F_{Y|\boldsymbol{\theta}}(y_i)$ and $z_{p_i} = \Phi^{-1}(p_i)$ for each data point. Then, the log-likelihood function (4) can be approximated by the function

$$\tilde{L}_n(\boldsymbol{\theta}) = \begin{cases} \sum_{i=1}^{n} \tilde{\psi}_{\boldsymbol{\theta}}(y_i), & \text{if } \mathsf{Y}_{1,\boldsymbol{\theta}} \le y_{\min} < y_{\max} \le \mathsf{Y}_{K_n,\boldsymbol{\theta}}, \\ -\infty, & \text{otherwise,} \end{cases} \tag{8}$$

where $\tilde{\psi}_{\boldsymbol{\theta}}(y)$ is defined in (7) and $y_{\min}, y_{\max}$ are the smallest and largest observed values. The maximum approximated likelihood estimator, $\hat{\boldsymbol{\theta}}_{male,n}$, is defined as the maximizer of (8).

2.4. Some computational issues

Rayner and MacGillivray (2002a) proposed a similar "change of support" approach to numerically approximate $\log f_{Y|\boldsymbol{\theta}}(y_i)$ for each y_i, $i = 1, \ldots, n$. The idea is essentially to numerically solve $y = \xi + \omega\tau_{g,h}(z_p)$ for p and then plug the solution $\hat{p}$ back into $\log f_{Y|\boldsymbol{\theta}}(y) = \varphi_{\boldsymbol{\theta}}(z_p)$. However, this approach can be computationally expensive when n is large. In addition, as illustrated in Fig. 2(d), $\varphi_{\boldsymbol{\theta}}(z_p)$ as a function of p is rather steep on both ends of the interval $[0, 1]$, which is very different from Fig. 2(b). Consequently, for p close to 0 or 1, a small error in $\hat{p}$ can lead to non-negligible errors when evaluating $\varphi_{\boldsymbol{\theta}}(z_p)$ using $\varphi_{\boldsymbol{\theta}}(z_{\hat{p}})$. In this sense, Rayner and MacGillivray's (2002a) method can also be numerically unstable.

The computational complexity of $\tilde{L}_n(\boldsymbol{\theta})$ is equivalent to assigning n data points to $K_n - 1$ bins in a histogram, which can be efficiently implemented using some bucket sort algorithm (Corwin and Logar, 2004). The average computational complexity for such an algorithm can be easily achieved as $O(n + K_n)$, which is fast even for very large n and K_n. In this paper, we use the function `.bincode` from the R software (R Development Core Team, 2014) to implement this algorithm. Based on Theorems 1–2, to guarantee estimation efficiency, we need to ensure that K_n is large enough so that Condition 5 in Appendix A is met. Our numerical experience indicates that using $K_n = \max(1000, n)$ is sufficient for most applications. With such a choice of K_n, our simulation studies in Section 4 demonstrate that the proposed estimation method can be several hundred times faster than the numerical maximum likelihood estimation approach proposed in Rayner and MacGillivray (2002a).

Another important issue in maximizing $\tilde{L}(\boldsymbol{\theta})$ defined in (8) is the choice of initial values for $\boldsymbol{\theta}$. In the supplementary material (see Appendix B), we give the Gradient and the Hessian matrix of $\tilde{L}(\boldsymbol{\theta})$, which can be utilized for any standard optimization routine. We use the "L-BFGS-B" method (Byrd et al., 1995) provided in the R function `optim` to maximize $\tilde{L}(\boldsymbol{\theta})$ since it is a constrained optimization problem due to the restriction $h \ge 0$. To find $\hat{\boldsymbol{\theta}}_{male,n}$, the initial value $\boldsymbol{\theta}^0$ must be chosen such that Condition 1 in Appendix A is met, that is, $\mathsf{Y}_{1,\boldsymbol{\theta}^0} \le y_{\min} < y_{\max} \le \mathsf{Y}_{K_n,\boldsymbol{\theta}^0}$. Based on our discussion in Section 2.1, we need essentially to find a $\boldsymbol{\theta}^0$ that is reasonably close to the true value $\boldsymbol{\theta}_0$. Fortunately, such a requirement can be easily fulfilled by taking $\boldsymbol{\theta}^0$ as the letter-value-based estimator $\hat{\boldsymbol{\theta}}_{lv,n}$ or the quantile least square estimator $\hat{\boldsymbol{\theta}}_{qls,n}$. In this paper, we choose $\hat{\boldsymbol{\theta}}_{lv,n}$ as the initial value of our algorithm.

Finally, we would like to point out that, unlike the quantile-based estimators $\hat{\boldsymbol{\theta}}_{lv,n}$ and $\hat{\boldsymbol{\theta}}_{qls,n}$, whose performances depend on a subjective choice of quantiles, the estimation accuracy of the maximum approximated likelihood estimator $\hat{\boldsymbol{\theta}}_{male,n}$ does not depend on the positions of the chosen knots as long as the number of knots K_n is sufficiently large. This is reflected in the fact that the results of Theorems 1–2 do not depend on K_n at all. In addition, we have proposed to use equally spaced knots merely for the simplicity of our technical investigations. Generally speaking, any knot sequence $-b_n = \mathsf{Z}_1 < \mathsf{Z}_2 < \cdots < \mathsf{Z}_{K_n} = b_n$ satisfying the condition that $\inf_{1\le i\le K_n-1}|\mathsf{Z}_{i+1} - \mathsf{Z}_i| \to 0$ sufficiently fast as $n \to \infty$ will do the job.

3. Statistical inference

It is often desirable to provide an uncertainty measure for a point estimator in statistical research, which remains a challenge when fitting Tukey's g-and-h distribution to data. For the popular letter-value-based approach, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{lv,n}$ remains unknown. Although the quantile least square approach proposed by Xu et al. (2014a,b) results in an asymptotic normal distribution for $\hat{\boldsymbol{\theta}}_{qls,n}$, there remain two problems to be addressed. First, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{qls,n}$ depends on a preselected set of p_k's and it remains unknown how this choice affects the subsequent statistical inference. Furthermore, their theory relies on the implied assumption that $h > 0$ and does not hold if the true value of h is 0. However, for Tukey's g-and-h distribution, $h = 0$ is a special case of particular interest. On the one hand, by testing $h = 0$,

one can tell whether or not the data have heavy tails. On the other hand, if we have significant evidence to believe that the true value of h is 0, then letting $h = 0$ would make the function $\tau_{g,h}(\cdot)$ invertible and thus makes the maximum likelihood estimator achievable.

3.1. Asymptotic properties of $\hat{\boldsymbol{\theta}}_{male,n}$

Denote by $\Omega = \mathbb{R} \times (0, \infty) \times \mathbb{R} \times [0, \infty)$ the parameter space of $\boldsymbol{\theta} = (\xi, \omega, g, h)^T$ and define $U_n(\boldsymbol{\theta}) = \partial L_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, $\tilde{U}_n(\boldsymbol{\theta}) = \partial\tilde{L}_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $I_n(\boldsymbol{\theta}) = -\partial^2 L_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$. Let $I(\boldsymbol{\theta}_0)$ be the expected value of $I_n(\boldsymbol{\theta})$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and define a random vector $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)^T \sim N_4(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$, where $N_4(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$ is a 4-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $I^{-1}(\boldsymbol{\theta}_0)$. The following theorem gives the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{male,n}$ under the scenario $h_0 = 0$ or $h_0 > 0$. The proof is given in Appendix A.

Theorem 1 (*Asymptotic Distribution*). *Under Conditions* 1–5 *in Appendix A, as* $n \to \infty$ *and* $K_n \to \infty$, *we have:* (a) *If the true value* $h_0 > 0$, *then* $\sqrt{n}(\hat{\boldsymbol{\theta}}_{male,n} - \boldsymbol{\theta}_0) \to \mathbf{Z}$ *in distribution;* (b) *If the true value* $h_0 = 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{male,n} - \boldsymbol{\theta}_0) \to \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix} I(Z_4 > 0) + \begin{pmatrix} Z_1 - I^{14}/I^{44}Z_4 \\ Z_2 - I^{24}/I^{44}Z_4 \\ Z_3 - I^{34}/I^{44}Z_4 \\ 0 \end{pmatrix} I(Z_4 < 0) \quad \textit{in distribution},$$

where I^{ij} *is the* ij*th element of* $I^{-1}(\boldsymbol{\theta}_0)$, $\boldsymbol{\theta} = (\xi, \omega, g, h)^T$ *and* $\boldsymbol{\theta}_0 = (\xi_0, \omega_0, g_0, h_0)^T$.

Theorem 1 essentially states that when $h_0 > 0$, if $K_n \to \infty$ fast enough as $n \to \infty$, $\hat{\boldsymbol{\theta}}_{male,n}$ has the same asymptotic distribution as $\hat{\boldsymbol{\theta}}_{mle,n}$ and reaches the Cramér–Rao efficiency lower bound. This is confirmed by our simulation example 1 in Section 4. When $h_0 = 0$, asymptotically, $\hat{\boldsymbol{\theta}}_{male,n}$ has a mixture normal distribution and there are 50% of times that h will be estimated as exactly 0. In the supplementary material (see Appendix B), we give a detailed description of how to compute $U_n(\boldsymbol{\theta})$ and $I_n(\boldsymbol{\theta})$ numerically and thus approximate the information matrix $I(\boldsymbol{\theta}_0)$ by $I_n(\hat{\boldsymbol{\theta}}_{male,n})$.

Based on Theorem 1, we can see that it is critical to determine whether $h_0 = 0$ or not because the limiting distributions of $\hat{\boldsymbol{\theta}}_{male,n}$ under these two scenarios are radically different. So the next question becomes: can we perform a hypothesis test for H_0: $h = 0$? Fortunately, the answer is yes. Let $\Omega_0 \subseteq \Omega$ be a subset of the parameter space of $\boldsymbol{\theta}$. We define the approximated likelihood ratio test statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta} \in \Omega_0$ as

$$D_n = -2\{\sup_{\boldsymbol{\theta}\in\Omega_0} \tilde{L}_n(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}\in\Omega} \tilde{L}_n(\boldsymbol{\theta})\}. \tag{9}$$

In this paper, we consider three types of Ω_0's: $\Omega_0^{(1)} = \mathbb{R} \times (0, \infty) \times \{0\} \times [0, \infty)$, $\Omega_0^{(2)} = \mathbb{R} \times (0, \infty) \times \mathbb{R} \times \{0\}$ and $\Omega_0^{(3)} = \mathbb{R} \times (0, \infty) \times \{0\} \times \{0\}$, which correspond to testing $H_0^{(1)} : g = 0$, $H_0^{(2)} : h = 0$, $H_0^{(3)} : h = g = 0$, respectively. The following theorem gives the limiting distributions of D_n under these three null hypotheses. The proof is given in Appendix A.

Theorem 2 (*Approximated Likelihood Ratio Tests, ALRT*). *Under Conditions* 1–5 *in Appendix A, as* $n \to \infty$ *and* $K_n \to \infty$,

1. *Under* $H_0^{(1)} : g = 0$, *one has* $D_n \to D \sim \chi_1^2$ *in distribution;*
2. *Under* $H_0^{(2)} : h = 0$, *one has* $D_n \to D \sim 0.5\chi_0^2 + 0.5\chi_1^2$ *in distribution;*
3. *Under* $H_0^{(3)} : g = h = 0$, *one has* $D_n \to D \sim 0.5\chi_1^2 + 0.5\chi_2^2$ *in distribution;*

where χ_{df}^2 *is the chi-square distribution with df degrees of freedom and* $0.5\chi_{df_1}^2 + 0.5\chi_{df_2}^2$ *stands for a mixture distribution of* $\chi_{df_1}^2$ *and* $\chi_{df_2}^2$ *with* 50% *of each component. In particular,* χ_0^2 *represents the distribution with a point mass at* 0, *i.e.* $P(X = 0) = 1$ *if* $X \sim \chi_0^2$.

The critical values of the mixed χ^2 distribution can be found using the function `qchibarsq` in the R package "emdbook" (Bolker, 2008). From Theorem 2, it is interesting that D_n for testing $H_0^{(1)} : g = 0$ has the same limiting distribution, whether or not the value of h_0 is 0. Under the null hypotheses $H_0^{(2)} : h = 0$ or $H_0^{(3)} : g = h = 0$, the first component of the asymptotic distributions of D_n comes from the cases when $\hat{h}_{male,n} = 0$, which should happen, asymptotically 50% of times by Theorem 1.

3.2. Finite sample performance of ALRT tests

Although by Theorem 1, when the sample size $n \to \infty$, the occurrence rate of h being estimated as 0, denoted by $P_{0,n}$, should approach 0.5, our empirical studies show that $P_{0,n}$ is almost always bigger than 0.5 when n is small. In our simulation studies, we noticed that, for a fixed n, $P_{0,n}$ is closely related to another quantity, $C_{0,n} = P(S_n < 0)$, where S_n is the fourth entry in the vector $n^{-1/2}U_n(\boldsymbol{\theta}_0)$. As illustrated in Fig. 3(b), when the sample size n grows, the difference between $P_{0,n}$ and $C_{0,n}$ becomes smaller and smaller. An intuitive explanation is that if $S_n < 0$, the approximated likelihood function $\tilde{L}(\boldsymbol{\theta})$ cannot be improved by moving the parameter h away from 0 toward ∞, given that the other parameters are fixed at (ξ_0, ω_0, g_0), and

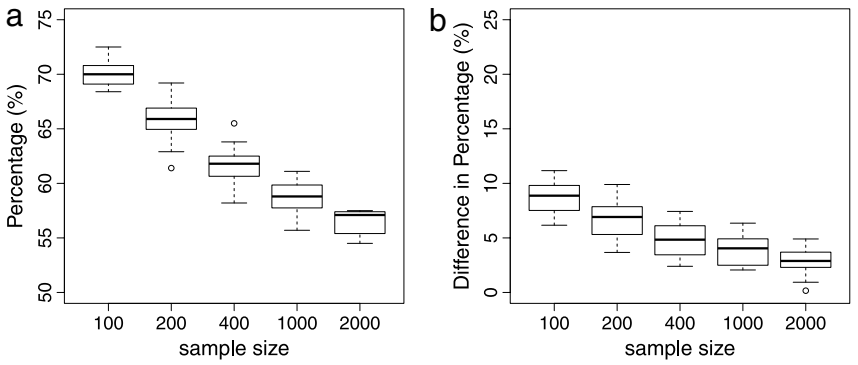


Fig. 3. (a) Percentage of times, $P_{0,n}$, when $\hat{h}_{male,n} = 0$ with $h_0 = 0$; (b) Difference between $P_{0,n}$ and $C_{0,n}$: $P_{0,n} - C_{0,n}$ vs sample size n.

therefore $h = 0$ should be the resulting solution. Although the explicit relationship between $P_{0,n}$ and $C_{0,n}$ remains unclear, studying the behavior of $C_{0,n}$ will provide some insights on why $P_{0,n} \to 0.5$ at such a slow rate.

In the special case of $g_0 = h_0 = 0$, some tedious algebra yields that $S_n = -n^{-1/2} \sum_{i=1}^{n} (z_i^4/3 - z_i^2)$ with z_i's being independent samples from the $N(0, 1)$ distribution. By the Central Limit Theorem, S_n converges in distribution to a normal distribution with mean 0 and thus $C_{0,n} = P(S_n < 0) \to 0.5$ as $n \to \infty$. However, even for a large n, the finite sample distribution of S_n can still be severely skewed to the left due to the existence of the term z_i^4 in the summation. As a result, even for a large n, we can still have $C_{0,n} > 0.5$. Our simulation studies show that this left-skewness in S_n becomes even more severe when $|g_0|$ deviates from 0. Because of the association between $P_{0,n}$ and $C_{0,n}$, this partially explains why $P_{0,n}$ is always greater than 0.5.

Next, we use a small simulation study to illustrate this phenomenon. The data were generated using model (1) with $\xi_0 = \omega_0 = 3$, $h_0 = 0$ and $g_0 = -0.5, -0.4, \ldots, 0.4, 0.5$. In Fig. 3, for each n, we plot the values of $P_{0,n}$ and $C_{0,n}$ obtained using different values of g_0 in a boxplot. We can see that $P_{0,n}$ is always greater than 0.5 regardless of the values of g_0 and n and that it slowly regresses to 0.5 as n increases. In addition, as n increases, the difference $P_{0,n} - C_{0,n}$ consistently decreases for various values of g_0, indicating strong association between $P_{0,n}$ and $C_{0,n}$.

Such a phenomenon indicates that, for a finite n, the critical values based on the asymptotic distributions in Theorem 2 are often too large for the null hypotheses $H_0^{(2)}$ and $H_0^{(3)}$. For example, to test $H_0^{(2)}$, if the estimated value for h is $\hat{h}_{male,n} = 0$ under the alternative hypothesis, the resulting approximate likelihood ratio test statistic $D_n = 0$. And if $\hat{h}_{male,n} = 0$ occurs more than 50% of times, then the finite sample distribution of D_n consists of more than 50% of 0's when $H_0^{(2)}$ is true, which means we will fail to reject $H_0^{(2)}$ more often than $(1-\alpha)100\%$ of times using the $(1-\alpha)$th quantile of the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution. Similar arguments apply to $H_0^{(3)}$ too. This explains why the size of the approximated likelihood ratio tests for $H_0^{(2)}$ and $H_0^{(3)}$ are often smaller than the nominal level in the simulation example 2. However, even with such conservativeness, the approximated likelihood ratio tests for $H_0^{(2)}$ and $H_0^{(3)}$ still display significant local powers.

4. Simulation studies

In this section, we use some numerical examples to show the effectiveness of our method. The data were generated using model (1) with different values of $\boldsymbol{\theta}_0 = (\xi_0, \omega_0, g_0, h_0)^T$ and five sample sizes $n = 100, 200, 400, 1000, 2000$. All summary statistics were based on 1000 repetitions and for each repetition, we used $b_n = 10$ and $K_n = \max(1000, n)$.

4.1. Evaluation of the estimation accuracy

In the simulation example 1, we mainly focused on comparing estimation accuracies for different methods described in Section 2. Three alternative approaches were considered: the letter-value-based method (LV), the quantile least square method (QLS) and the numerical maximum likelihood estimation (NMLE, Rayner and MacGillivray, 2002a). For the LV method, six quantiles were used: $\boldsymbol{p}_{lv} = (0.005, 0.01, 0.025, 0.05, 0.10, 0.25)^T$. For the QLS method, 10 quantiles were used with $p_i = \frac{i-1/3}{m+1/3}$, $i = 1, \ldots, m$ and $m = 10$. For the NMLE method, the function `uniroot` from the R package was used to numerically solve $y = \xi + \omega\tau_{g,h}(z_p)$ for p.

We set $\boldsymbol{\theta}_0 = (\xi_0, \omega_0, g_0, h_0)^T = (3, 3, 0.5, 0.2)^T$ and the empirical standard errors based on 1000 simulation runs are summarized in Fig. 4. The asymptotic standard error derived from Theorem 1, which is also the Cramér–Rao efficiency lower bound of the maximum likelihood estimator (MLE), is reported as well. We can see that the maximum approximated likelihood estimator (MALE) is uniformly more accurate than both LV and QLS methods for all sample sizes. In terms of estimation efficiencies, the NMLE and MALE methods are almost identical and reach the efficiency lower bounds, which confirms the theoretical findings in Theorem 1. In addition, the LV method did a better job at estimating the parameters g and h than the

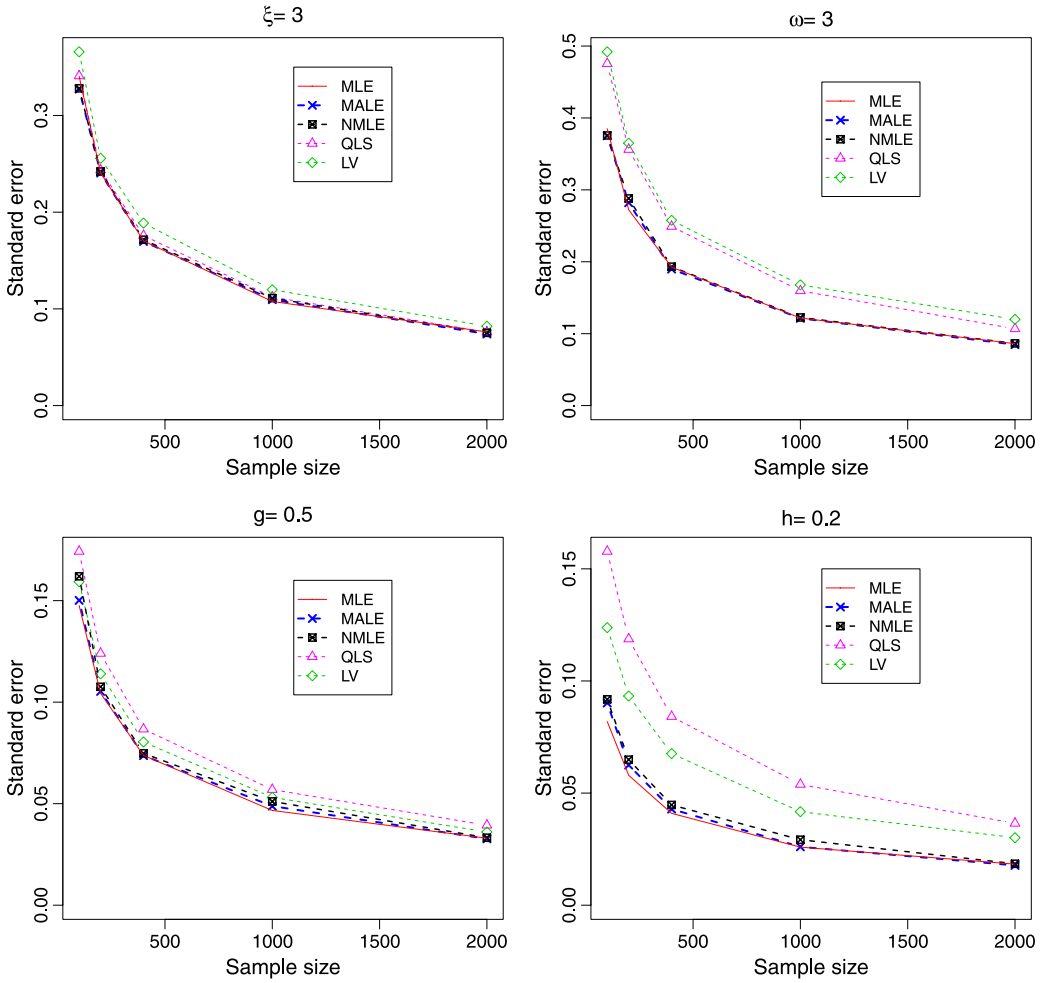


Fig. 4. The estimation efficiencies of four methods for Tukey's g-and-h distribution; MLE: asymptotic standard error of the maximum likelihood estimator; MALE: maximum approximated likelihood estimator; NMLE: numerical maximum likelihood estimator; QLS: quantile least square estimator; LV: letter-value-based estimator.

Table 1
Comparisons of computational costs for MALE and NMLE.

Sample size n	Average CPU time per run		Average of CPU time ratio per run
	MALE	NMLE	NMLE/MALE
100	0.13	10.86	91.68
200	0.14	22.60	173.80
400	0.17	47.15	301.86
1000	0.29	113.68	445.58
2000	0.53	229.01	464.58

QLS method did, while it seems to be the reverse for estimating ξ and ω. Additional simulation results using different values of $\boldsymbol{\theta}_0$ can be found in the supplementary material (see Appendix B), where the findings are all consistent with Fig. 4.

Although the MALE and NMLE methods yield roughly the same estimation accuracy, the proposed MALE method is more computationally efficient. To show this, we record their CPU times (in seconds) from the above simulation study. The same initial values and optimization routine were used for both methods. All simulation runs were carried out in the software R on a cluster of 120 commodity Linux machines using a total of 300 CPU cores, with each core running at approximately 2 GFLOPS, or 2 billion floating point operations per second. The results based on the 1000 simulation runs are summarized in Table 1, where the first two columns are average CPU times (in seconds) for each simulation run for both methods and the third column is the average of the ratios of CPU times for these two methods applied to the same dataset. From Table 1 we can see that, when the sample size grows to $n = 2000$, our method is on average over 400 times faster than the NMLE method. We also would like to point out that the MALE method converges much faster than the NMLE method for almost all simulation runs, which also contributes to the smaller computation time of the proposed method.

Table 2
Empirical powers of the approximated likelihood ratio tests in the simulation example 2.

d	n	$H_0 : g = 0$			$H_0 : h = 0$			$H_0 : g = h = 0$		
		Nominal levels			Nominal levels			Nominal levels		
		10.0	5.0	1.0	10.0	5.0	1.0	10.0	5.0	1.0
d_1	100	12.3	6.3	1.8	4.0	2.0	0.4	7.7	4.5	1.1
	200	11.4	5.7	1.4	4.5	2.2	0.1	9.6	4.7	0.8
	400	10.4	5.0	1.6	5.8	2.4	0.5	8.7	4.0	0.4
	1000	9.8	3.8	0.5	5.6	2.7	0.3	8.9	3.9	0.5
	2000	8.8	4.1	1.3	7.1	3.4	0.8	8.9	4.6	0.7
d_2	100	30.2	19.9	7.0	27.9	20.6	8.7	56.4	46.8	30.0
	200	27.9	19.4	5.8	33.9	25.4	10.4	58.9	47.7	29.2
	400	28.0	17.6	6.3	40.4	29.0	11.6	60.5	49.3	30.3
	1000	28.5	19.0	7.2	47.8	33.1	15.6	64.6	53.2	32.5
	2000	28.1	18.4	7.9	53.7	38.8	17.8	65.0	53.6	30.7
d_3	100	67.4	55.9	30.8	58.0	48.7	30.3	93.0	88.6	78.4
	200	68.6	57.3	32.6	68.2	58.7	41.7	95.3	93.3	84.0
	400	67.7	57.9	35.9	79.5	69.7	49.8	97.2	95.2	88.4
	1000	70.9	59.8	35.2	83.2	74.8	58.9	98.5	96.6	91.7
	2000	68.9	58.0	34.9	89.5	83.3	63.9	98.8	97.5	92.1

4.2. Evaluation of approximated likelihood ratio tests

In the simulation example 2, we study the empirical size and the local power of the proposed approximated likelihood ratio tests. The null and alternative hypotheses are

1. $H_0^{(1)} : g = 0$ vs $H_1 : g = d/\sqrt{n}$ with $(d_1, d_2, d_3) = (0, 1.5, 3)$. The data were generated using model (1) with $(\xi_0, \omega_0, g_0, h_0) = (3, 3, d/\sqrt{n}, 0.2)$;
2. $H_0^{(2)} : h = 0$ vs $H_1 : h = d/\sqrt{n}$ with $(d_1, d_2, d_3) = (0, 0.5, 1.0)$. The data were generated using model (1) with $(\xi_0, \omega_0, g_0, h_0) = (3, 3, 0.5, d/\sqrt{n})$;
3. $H_0^{(3)} : g = h = 0$ vs $H_1 : g = 3d/\sqrt{n}, h = d/\sqrt{n}$ with $(d_1, d_2, d_3) = (0, 0.5, 1.0)$. The data were generated using model (1) with $(\xi_0, \omega_0, g_0, h_0) = (3, 3, 3d/\sqrt{n}, d/\sqrt{n})$.

To study the ability of the proposed tests in detecting small departures from the null hypotheses, we deliberately set up alternative hypotheses as a function of $\sqrt{n}$, which are commonly referred to as local alternatives in the literature. Studying the power of hypothesis tests under local alternatives is more accurate in characterizing asymptotic behavior of the test statistic and has been widely used, for example, see Zhang (2001) and Xu and Wang (2010). The empirical powers based on 1000 simulation runs are summarized in Table 2. For $H_0^{(1)}$ and $H_0^{(3)}$, we see that under the null hypotheses, the empirical sizes are quite close to the nominal sizes. However, under the null hypothesis $H_0^{(2)}$, the empirical sizes are much smaller than the nominal level, even for relatively large n. In other words, in such a situation, the approximated likelihood ratio test tends to be conservative, which confirms our discussion in Section 3.2. However, all tests possess significant local powers in detecting very small deviations from the null hypotheses and these local powers increase when the departure from the null hypothesis grows from $d_2/\sqrt{n}$ to $d_3/\sqrt{n}$.

5. Air pollution data

In this section, we apply the proposed method to two years of air pollution data obtained from Ledolter and Hogg (2009, Chapter 1, Section 1.4), which can be downloaded at the website http://www.biz.uiowa.edu/faculty/jledolter/AppliedStatistics/. In each of 1976 and 1977, 64 weekday afternoon lead concentrations, in micrograms/m^3, were collected near the San Diego Freeway in Los Angeles. The same datasets were also used by Rayner and MacGillivray (2002a,b). In Fig. 5, the histogram and the density plots show possible skewness and some outliers in both datasets. Following Rayner and MacGillivray (2002b), we use Tukey's g-and-h distribution to fit both datasets.

For parameter estimation, the four methods described in Section 4.1 were used. All estimates are summarized in Table 3. As we can see, the estimates $\hat{\boldsymbol{\theta}}_{male,n}$ and $\hat{\boldsymbol{\theta}}_{nmle,n}$ are identical. This is not surprising because both of these methods should yield estimators that are very close to the true maximum likelihood estimator. Although $\hat{\boldsymbol{\theta}}_{lv,n}$ and $\hat{\boldsymbol{\theta}}_{qls,n}$ yield consistent estimates of ξ, ω and g, compared to $\hat{\boldsymbol{\theta}}_{male,n}$ and $\hat{\boldsymbol{\theta}}_{nmle,n}$, they produce very different estimates for h. For example, for 1977, $\hat{\boldsymbol{\theta}}_{qls,n}$ gives $\hat{h}_{qls,n} = 0.49$, which is very questionable because with this value the fitted distribution basically has an infinite variance.

We also ran a series of hypothesis tests using the approximated likelihood ratio test statistics and the results are summarized in Table 3, where $C_{0.95}$ stands for the critical value derived from the asymptotic distribution given in Theorem 2 for each test at the 0.05 significance level. Based on these tests, there is significant evidence that h_0 is not 0 but it seems reasonable to set $g = 0$ for both datasets. This indicates that for both datasets, the data do not have a normal distribution but a heavy-tailed symmetric distribution.

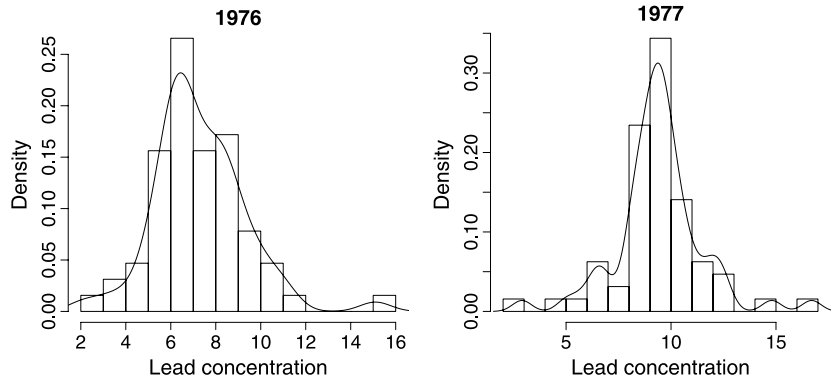


Fig. 5. Histograms of lead concentrations in 1976 and 1977.

Table 3
Estimates and hypothesis tests of lead concentration data.

	Methods	Year 1976	Year 1977
Estimation $(\hat{\xi}, \hat{\omega}, \hat{g}, \hat{h})$	$\hat{\boldsymbol{\theta}}_{male,n}$	(7.11, 1.60, 0.19, 0.12)	(9.38, 1.22, 0.04, 0.34)
	$\hat{\boldsymbol{\theta}}_{nmle,n}$	(7.11, 1.60, 0.19, 0.12)	(9.38, 1.22, 0.04, 0.34)
	$\hat{\boldsymbol{\theta}}_{lv,n}$	(6.95, 1.62, 0.25, 0.07)	(9.40, 1.33, 0.04, 0.21)
	$\hat{\boldsymbol{\theta}}_{qls,n}$	(7.12, 1.62, 0.24, 0.01)	(9.41, 1.08, 0.01, 0.49)
Hypothesis tests $(D_n, C_{0.95})$	H_0: $g = h = 0$	(7.26, 5.14), Reject;	(14.34, 5.14), Reject;
	H_0: $h = 0$	(4.46, 2.71), Reject;	(14.05, 2.71), Reject;
	H_0: $g = 0$	(1.54, 3.84), Fail to reject.	(0.03, 3.84), Fail to reject.

6. Conclusion

In this paper, we proposed a computationally efficient estimation method for Tukey's g-and-h distribution. The proposed maximum approximated likelihood estimator has the same limiting distribution as the true maximum likelihood estimator and the resulting approximated likelihood ratio tests can be used to test interesting hypotheses involving the two shape parameters g and h. The performances of the proposed method have been demonstrated through extensive simulation studies and an application to a air pollution data.

There are several ways to extend the current results. Tukey's g-and-h distribution can be used as the residual process of a regression model to conduct robust model selection and model averaging (Xu et al., 2014a,b). Another interesting application is to conduct model selection for an autoregressive model with an infinite variance as in Xu et al. (2012), where Tukey's g-and-h distribution can be used to replace the stable distribution as the innovation process. It would also be interesting to conduct some theoretical investigations on the local powers of the ALRT tests in Section 3 and how to construct valid confidence intervals for g and h, without knowing the true value $h_0 = 0$ or not.

Appendix A. Technical proofs

We first define some functions $C_1(z,\boldsymbol{\theta}) = \frac{\partial \varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger z_p}\Big|_{z_p=z}$, $C_2(z,\boldsymbol{\theta}) = \frac{\partial^2 \varphi_{\boldsymbol{\theta}}(z_p)}{\partial \boldsymbol{\theta}\partial^\dagger z_p}\Big|_{z_p=z}$, $C_3(z,\boldsymbol{\theta}) = \frac{\tau''_{g,h}(z)}{\tau'_{g,h}(z)}\frac{\partial \varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger z_p}\frac{\partial z_p}{\partial \boldsymbol{\theta}}\Big|_{z_p=z}$, $C_4(z,\boldsymbol{\theta}) = \frac{1}{K_n}\frac{\partial \log \tau'_{g,h}(z)}{\partial \boldsymbol{\theta}}\frac{\partial \varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger z_p}\Big|_{z_p=z}$, $C_5(z,\boldsymbol{\theta}) = \frac{\partial \varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger z_p}\frac{\partial^2 z_p}{\partial \boldsymbol{\theta}\partial^\dagger z_p}\Big|_{z_p=z}$. We list some technical conditions needed for proofs of Theorems 1–2, where Conditions 2–3 are taken from Self and Liang (1987).

Condition 1: There exists an open neighborhood of $\boldsymbol{\theta}_0$, denoted by $\Theta_{0,n} \subseteq \mathbb{R}^4$, such that $Y_{1,\boldsymbol{\theta}} \le y_{\min} < y_{\max} \le Y_{K_n,\boldsymbol{\theta}}$ with probability 1 for all $\boldsymbol{\theta} \in \Theta_{0,n}$;

Condition 2: The first three derivatives of $L_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist with probability 1 on $\Theta_{0,n} \cap \Omega$;

Condition 3: The third derivative of $n^{-1}L_n(\boldsymbol{\theta})$ is bounded by some function $\psi_n(\mathbf{Y})$ with $|E\{\psi_n(\mathbf{Y})\}| < \infty$;

Condition 4: $I(\boldsymbol{\theta}) = E\{I_n(\boldsymbol{\theta})\}$ is positive definite on $\Theta_{0,n}$ and $I(\boldsymbol{\theta}_0) = \text{var}\left\{n^{-1/2}U_n(\boldsymbol{\theta}_0)\right\}$;

Condition 5: $\frac{2\sqrt{n}b_n}{K_n}\sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega, z\in[-b_n,b_n]}\left|C_j(z,\boldsymbol{\theta})\right| \to 0$ for $j = 1,\ldots,5$ as $n \to \infty$ and $K_n \to \infty$.

Lemma A.1. *Under Conditions* 1–5, *as* $K_n \to \infty$ *and* $n \to \infty$, *we have that*

$$\frac{1}{\sqrt{n}} \sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega} |\tilde{L}_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta})| = o_p(1), \tag{A.1}$$

$$\frac{1}{\sqrt{n}} \sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega} |\tilde{U}_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta})| = o_p(1). \tag{A.2}$$

Proof. For a fixed $\boldsymbol{\theta} \in \Theta_{0,n} \cap \Omega$, let $z_p = \tau_{g,h}^{-1}(\frac{y-\xi}{\omega})$ and suppose $\mathsf{Z}_k \le z_p < \mathsf{Z}_{k+1}$ for some $1 \le k \le K_n$. Then we can define the corresponding $\tilde{z}_{p,k}$ as in (6) such that $|z_p - \tilde{z}_{p,k}| \le 2b_n/K_n$, which implies that

$$\left|\varphi_{\boldsymbol{\theta}}(z_p) - \tilde{\varphi}_{\boldsymbol{\theta}}(z_p)\right| \le \frac{2b_n}{K_n} \sup_{z_p\in[-b_n,b_n]} \left|\frac{\partial\varphi_{\boldsymbol{\theta}}}{\partial^\dagger z_p}(z_p)\right|.$$

Then, it is straightforward to show that, under Conditions 1 and 5, we have

$$\begin{aligned}\frac{1}{\sqrt{n}} \sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega} |\tilde{L}_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta})| &\le \sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |\varphi_{\boldsymbol{\theta}}(z_{p_i}) - \tilde{\varphi}_{\boldsymbol{\theta}}(z_{p_i})| \\ &\le \frac{2\sqrt{n}b_n}{K_n} \sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega, z_p\in[-b_n,b_n]} |C_1(z_p,\boldsymbol{\theta})| = o_p(1).\end{aligned}$$

By definition, we can rewrite (6) as

$$\tilde{z}_{p,k} = \mathsf{Z}_k + \frac{y - \mathsf{Y}_{k,\boldsymbol{\theta}}}{\mathsf{Y}_{k+1,\boldsymbol{\theta}} - \mathsf{Y}_{k,\boldsymbol{\theta}}}(\mathsf{Z}_{k+1} - \mathsf{Z}_k) = \mathsf{Z}_k + \frac{\tau_{g,h}(z_p) - \tau_{g,h}(\mathsf{Z}_k)}{\tau_{g,h}(\mathsf{Z}_{k+1}) - \tau_{g,h}(\mathsf{Z}_k)}(\mathsf{Z}_{k+1} - \mathsf{Z}_k).$$

Applying the chain rule of differentiation, we have

$$u_{\boldsymbol{\theta}}(z_p) = \frac{\partial\varphi_{\boldsymbol{\theta}}(z_p)}{\partial\boldsymbol{\theta}} = \frac{\partial\varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger\boldsymbol{\theta}} + \frac{\partial\varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger z_p}\frac{\partial z_p}{\partial\boldsymbol{\theta}} = v_{\boldsymbol{\theta}}(z_p) + w_{\boldsymbol{\theta}}(z_p)x_{\boldsymbol{\theta}}(z_p),$$

where we denote $v_{\boldsymbol{\theta}}(z_p) = \frac{\partial\varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger\boldsymbol{\theta}}$, $w_{\boldsymbol{\theta}}(z_p) = \frac{\partial\varphi_{\boldsymbol{\theta}}(z_p)}{\partial^\dagger z_p}$ and $x_{\boldsymbol{\theta}}(z_p) = \frac{\partial z_p}{\partial\boldsymbol{\theta}}$. Similarly, we have

$$\tilde{u}_{\boldsymbol{\theta}}(z_p) = \frac{\partial\varphi_{\boldsymbol{\theta}}(\tilde{z}_{p,k})}{\partial\boldsymbol{\theta}} = \frac{\partial\varphi_{\boldsymbol{\theta}}(\tilde{z}_{p,k})}{\partial^\dagger\boldsymbol{\theta}} + \frac{\partial\varphi_{\boldsymbol{\theta}}(\tilde{z}_{p,k})}{\partial^\dagger\tilde{z}_{p,k}}\frac{\partial\tilde{z}_{p,k}}{\partial\boldsymbol{\theta}} = v_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) + w_{\boldsymbol{\theta}}(\tilde{z}_{p,k})\tilde{x}_{\boldsymbol{\theta}}(z_p),$$

where we denote $\tilde{x}_{\boldsymbol{\theta}}(z_p) = \frac{\partial\tilde{z}_{p,k}}{\partial\boldsymbol{\theta}}$. Straightforward algebra yields

$$\begin{aligned}u_{\boldsymbol{\theta}}(z_p) - \tilde{u}_{\boldsymbol{\theta}}(z_p) &= u_{\boldsymbol{\theta}}(z_p) - u_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) + w_{\boldsymbol{\theta}}(\tilde{z}_{p,k})\left\{x_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) - \tilde{x}_{\boldsymbol{\theta}}(z_p)\right\} \\ &= \frac{\partial^2\varphi_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}\partial^\dagger z_p}(z^*_{p,1})(z_p - \tilde{z}_{p,k}) + w_{\boldsymbol{\theta}}(\tilde{z}_{p,k})\left\{x_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) - \tilde{x}_{\boldsymbol{\theta}}(z_p)\right\},\end{aligned} \tag{A.3}$$

where $z^*_{p,1} \in [\mathsf{Z}_k, \mathsf{Z}_{k+1})$. Furthermore, we can show that

$$\begin{aligned}\tilde{x}_{\boldsymbol{\theta}}(z_p) - x_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) &= \tilde{x}_{\boldsymbol{\theta}}(z_p) - x_{\boldsymbol{\theta}}(z_p) + x_{\boldsymbol{\theta}}(z_p) - x_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) = \frac{\partial\tilde{z}_{p,k}}{\partial\boldsymbol{\theta}} - x_{\boldsymbol{\theta}}(z_p) + x_{\boldsymbol{\theta}}(z_p) - x_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) \\ &= \frac{\partial\tilde{z}_{p,k}}{\partial^\dagger z_p}\frac{\partial z_p}{\partial\boldsymbol{\theta}} + \frac{\partial\tilde{z}_{p,k}}{\partial^\dagger\boldsymbol{\theta}} - x_{\boldsymbol{\theta}}(z_p) + x_{\boldsymbol{\theta}}(z_p) - x_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) \\ &= \left\{\frac{\tau'_{g,h}(z_p)}{\tau_{g,h}(\mathsf{Z}_{k+1}) - \tau_{g,h}(\mathsf{Z}_k)}(\mathsf{Z}_{k+1} - \mathsf{Z}_k) - 1\right\}\frac{\partial z_p}{\partial\boldsymbol{\theta}} + \frac{\partial\tilde{z}_{p,k}}{\partial^\dagger\boldsymbol{\theta}} + x_{\boldsymbol{\theta}}(z_p) - x_{\boldsymbol{\theta}}(\tilde{z}_{p,k}) \\ &= \frac{\tau''_{g,h}(z^*_{p,3})(z_p - z^*_{p,2})}{\tau'_{g,h}(z^*_{p,2})}\frac{\partial z_p}{\partial\boldsymbol{\theta}} + \frac{\frac{\partial\tau'_{g,h}}{\partial\boldsymbol{\theta}}(z^*_{p,4})(z^*_{p,5} - z^*_{p,6})}{\tau'_{g,h}(z^*_{p,2})K_n} + \frac{\partial^2 z_p}{\partial\boldsymbol{\theta}\partial^\dagger z_p}(z^*_{p,7})(z_p - \tilde{z}_{p,k}),\end{aligned} \tag{A.4}$$

where $\mathsf{Z}_k \le z^*_{p,i} \le \mathsf{Z}_{k+1}$ with $2 \le i \le 7$. Then, combining (A.3) and (A.4), under Condition 5 and the fact that $\mathsf{Z}_{k+1} - \mathsf{Z}_k = 2b_n/K_n$, we have

$$\sup_{\boldsymbol{\theta}\in\Theta_0\cap\Omega, z_p\in[-b_n,b_n]} \sqrt{n}|u_{\boldsymbol{\theta}}(z_p) - \tilde{u}_{\boldsymbol{\theta}}(z_p)| = o_p(1),$$

which further leads to, under Conditions 1 and 5, that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sup_{\boldsymbol{\theta}\in\Theta_0\cap\Omega}|\tilde{U}_n(\boldsymbol{\theta})-U_n(\boldsymbol{\theta})| &\le \sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|u_{\boldsymbol{\theta}}(z_{p_i})-\tilde{u}_{\boldsymbol{\theta}}(z_{p_i})|\\ &\le \sup_{\boldsymbol{\theta}\in\Theta_{0,n}\cap\Omega,z_p\in[-b_n,b_n]}\sqrt{n}|u_{\boldsymbol{\theta}}(z_p)-\tilde{u}_{\boldsymbol{\theta}}(z_p)|=o_p(1). \quad \square\end{aligned}$$

Before proceeding to proofs of Theorems 1–2, we first define quantities $\zeta_n = n^{-1}I^{-1}(\boldsymbol{\theta}_0)U_n(\boldsymbol{\theta}_0)$ and $H_n(\boldsymbol{\theta}) = -\{\zeta_n-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}^T I(\boldsymbol{\theta}_0)\{\zeta_n-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}+n^{-2}U_n^T(\boldsymbol{\theta}_0)^T I^{-1}(\boldsymbol{\theta}_0)U_n^T(\boldsymbol{\theta}_0)$.

Proof of Theorem 1. Under Conditions 2–4, by Lemma 1 in Self and Liang (1987), for any $\boldsymbol{\theta}$ such that $\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=O(1)$, we have

$$\frac{2}{n}\{L_n(\boldsymbol{\theta})-L_n(\boldsymbol{\theta}_0)\}=H_n(\boldsymbol{\theta})+R_n(\boldsymbol{\theta}), \tag{A.5}$$

where $R_n(\boldsymbol{\theta})=O_p(1)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^3$.

Following the same arguments in the proof of Theorem 1 in Self and Liang (1987) and using Eq. (A.1) in Lemma A.1, we can show that there exists a sequence $\hat{\boldsymbol{\theta}}_{male,n}\in\Omega$ such that $|\hat{\boldsymbol{\theta}}_{male,n}-\boldsymbol{\theta}_0|=O_p(1/\sqrt{n})$. Denote by $\hat{\boldsymbol{\theta}}_{H_n}$ the maximizer of the quadratic function $H_n(\boldsymbol{\theta})$. Because of the convexity of Ω, it is straightforward to show that $|\hat{\boldsymbol{\theta}}_{H_n}-\boldsymbol{\theta}_0|=O_p(1/\sqrt{n})$ and hence $|\hat{\boldsymbol{\theta}}_{male,n}-\hat{\boldsymbol{\theta}}_{H_n}|=O_p(1/\sqrt{n})$. Therefore, we must have that

$$\begin{aligned}\frac{2}{n}\{\tilde{L}_n(\hat{\boldsymbol{\theta}}_{male,n})-\tilde{L}_n(\hat{\boldsymbol{\theta}}_{H_n})\} &= \frac{2}{n}\{\tilde{L}_n(\hat{\boldsymbol{\theta}}_{male,n})-\tilde{L}_n(\hat{\boldsymbol{\theta}}_{H_n})\}-\frac{2}{n}\{L_n(\hat{\boldsymbol{\theta}}_{male,n})-L_n(\hat{\boldsymbol{\theta}}_{H_n})\}+\frac{2}{n}\{L_n(\hat{\boldsymbol{\theta}}_{male})-L_n(\hat{\boldsymbol{\theta}}_H)\}\\ &= \frac{2}{n}\{\tilde{L}_n(\hat{\boldsymbol{\theta}}_{male,n})-L_n(\hat{\boldsymbol{\theta}}_{male,n})\}-\frac{2}{n}\{\tilde{L}_n(\hat{\boldsymbol{\theta}}_{H_n})-L_n(\hat{\boldsymbol{\theta}}_{H_n})\}+\frac{2}{n}\{L_n(\hat{\boldsymbol{\theta}}_{male,n})-L_n(\hat{\boldsymbol{\theta}}_{H_n})\}\\ &= \frac{2}{n}\{\tilde{U}_n(\hat{\boldsymbol{\theta}}_1^*)-U_n(\hat{\boldsymbol{\theta}}_1^*)\}(\hat{\boldsymbol{\theta}}_{male,n}-\hat{\boldsymbol{\theta}}_{H_n})+H_n(\hat{\boldsymbol{\theta}}_{male,n})\\ &\quad -H_n(\hat{\boldsymbol{\theta}}_{H_n})+R_n(\hat{\boldsymbol{\theta}}_{male,n})-R_n(\hat{\boldsymbol{\theta}}_{H_n})\\ &= H_n(\hat{\boldsymbol{\theta}}_{male,n})-H_n(\hat{\boldsymbol{\theta}}_{H_n})+o_p(1/n),\end{aligned} \tag{A.6}$$

where $\hat{\boldsymbol{\theta}}_1^*\in(\hat{\boldsymbol{\theta}}_{male,n},\hat{\boldsymbol{\theta}}_{H_n})$. The last equation follows by combining (A.2) and (A.5) in Lemma A.1.

Comparing the left-hand side and right-hand side of (A.6), by definitions of $\hat{\boldsymbol{\theta}}_{male,n}$ and $\hat{\boldsymbol{\theta}}_{H_n}$, we must have $\tilde{L}_n(\hat{\boldsymbol{\theta}}_{male,n})-\tilde{L}_n(\hat{\boldsymbol{\theta}}_{H_n})\ge 0$ and $H_n(\hat{\boldsymbol{\theta}}_{male,n})-H_n(\hat{\boldsymbol{\theta}}_{H_n})\le 0$, which implies that $|H_n(\hat{\boldsymbol{\theta}}_{male,n})-H_n(\hat{\boldsymbol{\theta}}_{H_n})|=o_p(1/n)$. Furthermore, since $H_n(\boldsymbol{\theta})$ is a quadratic function, we conclude that $|\hat{\boldsymbol{\theta}}_{male,n}-\hat{\boldsymbol{\theta}}_{H_n}|=o_p(1/\sqrt{n})$.

The rest of the proof follows from Theorem 2 in Self and Liang (1987) and its application to cases 1 and 2 in that paper. $\square$

Proof of Theorem 2. Denote $\hat{\boldsymbol{\theta}}^0_{male,n}=\arg\sup_{\boldsymbol{\theta}\in\Omega_0}\tilde{L}_n(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_{male,n}=\arg\sup_{\boldsymbol{\theta}\in\Omega}\tilde{L}_n(\boldsymbol{\theta})$. Similarly, define the true maximum likelihood estimator as $\hat{\boldsymbol{\theta}}^0_{mle,n}=\arg\sup_{\boldsymbol{\theta}\in\Omega_0}L_n(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_{mle,n}=\arg\sup_{\boldsymbol{\theta}\in\Omega}L_n(\boldsymbol{\theta})$. Then it is easy to show that

$$\begin{aligned}D_n &= -2\{\tilde{L}_n(\hat{\boldsymbol{\theta}}^0_{male,n})-\tilde{L}_n(\hat{\boldsymbol{\theta}}_{male,n})\}\\ &= \underbrace{-2\{\tilde{L}_n(\hat{\boldsymbol{\theta}}^0_{male,n})-\tilde{L}_n(\hat{\boldsymbol{\theta}}_{male,n})\}+2\{L_n(\hat{\boldsymbol{\theta}}^0_{male,n})-L_n(\hat{\boldsymbol{\theta}}_{male,n})\}}_{\mathrm{I}}\\ &\quad \underbrace{-2\{L_n(\hat{\boldsymbol{\theta}}^0_{male,n})-L_n(\hat{\boldsymbol{\theta}}_{male,n})\}+2\{L_n(\hat{\boldsymbol{\theta}}^0_{mle,n})-L_n(\hat{\boldsymbol{\theta}}_{mle,n})\}}_{\mathrm{II}}\\ &\quad \underbrace{-2\{L_n(\hat{\boldsymbol{\theta}}^0_{mle,n})-L_n(\hat{\boldsymbol{\theta}}_{mle,n})\}}_{D_n^*}.\end{aligned} \tag{A.7}$$

Under the null hypothesis $H_0:\boldsymbol{\theta}\in\Omega_0$, we have shown in the proof of Theorem 1 that $|\hat{\boldsymbol{\theta}}^0_{male,n}-\boldsymbol{\theta}_0|=O_p(1/\sqrt{n})$ and $|\hat{\boldsymbol{\theta}}_{male,n}-\boldsymbol{\theta}_0|=O_p(1/\sqrt{n})$. Hence, $|\hat{\boldsymbol{\theta}}^0_{male,n}-\hat{\boldsymbol{\theta}}_{male,n}|=O_p(1/\sqrt{n})$. Using (A.2) in Lemma A.1, we can show that

$$\begin{aligned}\mathrm{I} &= -2\{\tilde{L}_n(\hat{\boldsymbol{\theta}}^0_{male,n})-L_n(\hat{\boldsymbol{\theta}}^0_{male,n})\}+2\{\tilde{L}_n(\hat{\boldsymbol{\theta}}_{male,n})-L_n(\hat{\boldsymbol{\theta}}_{male,n})\}\\ &= -2\{\tilde{U}_n(\hat{\boldsymbol{\theta}}_1^*)-U_n(\hat{\boldsymbol{\theta}}_1^*)\}(\hat{\boldsymbol{\theta}}^0_{male,n}-\hat{\boldsymbol{\theta}}_{male,n})=o_p(1),\end{aligned}$$

where $\hat{\boldsymbol{\theta}}_1^*\in(\hat{\boldsymbol{\theta}}^0_{male,n},\hat{\boldsymbol{\theta}}_{male,n})$.

Lemma 1 in Self and Liang (1987) showed that, under the null hypothesis, $H_0 : \boldsymbol{\theta} \in \Omega_0$, one has $|\hat{\boldsymbol{\theta}}^0_{mle,n} - \hat{\boldsymbol{\theta}}^0_{H_n}| = o_p(1/\sqrt{n})$ and $|\hat{\boldsymbol{\theta}}_{mle,n} - \hat{\boldsymbol{\theta}}_{H_n}| = o_p(1/\sqrt{n})$. Following the proof of Theorem 1, we can also show that $|\hat{\boldsymbol{\theta}}^0_{male,n} - \hat{\boldsymbol{\theta}}^0_{H_n}| = o_p(1/\sqrt{n})$ and $|\hat{\boldsymbol{\theta}}_{male,n} - \hat{\boldsymbol{\theta}}_{H_n}| = o_p(1/\sqrt{n})$. Therefore $|\hat{\boldsymbol{\theta}}^0_{male,n} - \hat{\boldsymbol{\theta}}^0_{mle,n}| = o_p(1/\sqrt{n})$ and $|\hat{\boldsymbol{\theta}}_{male,n} - \hat{\boldsymbol{\theta}}_{mle,n}| = o_p(1/\sqrt{n})$. Then, it is easy to show that

$$\begin{aligned} \text{II} &= -2\{L_n(\hat{\boldsymbol{\theta}}^0_{male,n}) - L_n(\hat{\boldsymbol{\theta}}^0_{mle,n})\} + 2\{L_n(\hat{\boldsymbol{\theta}}_{male,n}) - L_n(\hat{\boldsymbol{\theta}}_{mle,n})\} \\ &= -2U^0_n(\hat{\boldsymbol{\theta}}^*_2)(\hat{\boldsymbol{\theta}}^0_{male,n} - \hat{\boldsymbol{\theta}}^0_{mle,n}) + 2U_n(\hat{\boldsymbol{\theta}}^*_3)(\hat{\boldsymbol{\theta}}_{male,n} - \hat{\boldsymbol{\theta}}_{mle,n}) = o_p(1), \end{aligned}$$

where $\hat{\boldsymbol{\theta}}^*_2 \in (\hat{\boldsymbol{\theta}}^0_{male,n}, \hat{\boldsymbol{\theta}}^0_{mle,n})$ and $\hat{\boldsymbol{\theta}}^*_3 \in (\hat{\boldsymbol{\theta}}_{male,n}, \hat{\boldsymbol{\theta}}_{mle,n})$ and $U^0_n(\hat{\boldsymbol{\theta}}^*_2)$ is the partial derivative function of $L_n(\boldsymbol{\theta})$ in the restricted parameter space Ω_0. The last equation follows from the fact $1/\sqrt{n}U^0_n(\boldsymbol{\theta})$ and $\sqrt{n}U_n(\boldsymbol{\theta})$ are bounded for $\boldsymbol{\theta}$ in a "small" neighborhood of $\hat{\boldsymbol{\theta}}^0_{mle,n}$ and $\hat{\boldsymbol{\theta}}_{mle,n}$, respectively.

By plugging I and II back into Eq. (A.5), we have $D_n = D^*_n + o_p(1)$, which implies that D_n and D^*_n have the same asymptotic distribution. The distribution of D^*_n was derived in Self and Liang (1987). The rest of the proof follows from Theorem 3 in Self and Liang (1987) as well as its application to cases 4–6 in that paper. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2015.06.002.

References

Bolker, B.M., 2008. Ecological Models and Data in R. Princeton University Press.
Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16, 1190–1208.
Corwin, E., Logar, A., 2004. Sorting in linear time variations on the bucket sort. J. Comput. Sci. Coll. 20, 197–202.
Degen, M., Embrechts, P., Lambrigger, D.D., 2007. The quantitative modeling of operational risk: between g-and-h and EVT. Astin Bull. 37, 265–291.
Dutta, K.K., Babbel, D.F., 2002. On measuring skewness and kurtosis in short rate distributions: The case of the US dollar London inter bank offer rates, 1 Technical Report. The Wharton School, University of Pennsylvania.
Field, C., 2004. Using the *gh* distribution to model extreme wind speeds. J. Statist. Plann. Inference 122, 15–22.
Field, C., Genton, M.G., 2006. The multivariate *g*-and-*h* distribution. Technometrics 48, 104–111.
He, Y., Raghunathan, T.E., 2006. Tukey's gh distribution for multiple imputation. Amer. Statist. 60, 251–256.
He, Y., Raghunathan, T.E., 2012. Multiple imputation using multivariate gh transformations. J. Appl. Stat. 39, 2177–2198.
Hoaglin, D.C., 1985. Summarizing shape numerically: The *g*-and-*h* distributions. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), Data Analysis for Tables, Trends and Shapes: Robust and Exploratory Techniques. Wiley, New York.
Hoaglin, D.C., 2010. Extreme-value distributions as *g*-and-*h* distributions: An empirical view, Technical Report. JSM.
Jiménez, J.A., Arunachalam, V., 2011. Using Tukey's g and h family of distributions to calculate value-at-risk and conditional value-at-risk. J. Risk 13, 95–116.
Jones, M.C., 2015. On families of distributions with shape parameters (with discussions). Internat. Statist. Rev. in press.
Ledolter, J., Hogg, R.V., 2009. Applied Statistics for Engineers and Physical Scientists, third ed. Pearson/Prentice Hall.
MacGillivray, H.L., 1992. Shape properties of the g-and-h and Johnson families. Comm. Statist. Theory Methods 21, 1233–1250.
Martinez, J., Iglewicz, B., 1984. Some properties of the Tukey *g* and *h* family of distributions. Comm. Statist. Theory Methods 13, 353–369.
Morgenthaler, S., Tukey, J.W., 2000. Fitting quantiles: doubling, HR, HQ, and HHH distributions. J. Comput. Graph. Statist. 9, 180–195.
Rayner, G.D., MacGillivray, H.L., 2002a. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. Stat. Comput. 12, 57–75.
Rayner, G.D., MacGillivray, H.L., 2002b. Weighted quantile-based estimation for a class of transformation distributions. Comput. Statist. Data Anal. 39, 401–433.
R Development Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0, URL: http://www.R-project.org.
Self, S.G., Liang, K.Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Amer. Statist. Assoc. 82, 605–610.
Tukey, J.W., 1977. Modern techniques in data analysis. In: NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.
Xu, Y., Iglewicz, B., Chervoneva, I., 2014a. Robust estimation of the parameters of g-and-h distributions, with applications to outlier detection. Comput. Statist. Data Anal. 75, 66–80.
Xu, G., Wang, S., 2010. A goodness-of-fit test of logistic regression models for case-control data with measurement errors. Biometrika 98, 877–886.
Xu, G., Wang, S., Huang, J.Z., 2014b. Focused information criterion and model averaging based on weighted composite quantile regression. Scand. J. Statist. 41, 365–381.
Xu, G., Xiang, Y.B., Wang, S., Lin, Z.Y., 2012. Regularization and variable selection for infinite variance autoregressive models. J. Statist. Plann. Inference 142, 2545–2553.
Zhang, B., 2001. An information matrix test for logistic regression models based on case-control data. Biometrika 88, 921–932.