Published in final edited form as:

Comput Stat Data Anal. 2014 July 1; 75: 66–80. doi:10.1016/j.csda.2014.01.003.

Robust Estimation of the Parameters of g - and - h Distributions, with Applications to Outlier Detection

Yihuan Xu,

ImClone LLC, A Wholly Owned Subsidiary of Eli Lilly and Company, 440 Route 22, Bridgewater, NJ 08807

Boris Iglewicz¹, and

Department of Statistics, The Fox School of Business, Temple University, 1810 North 13th Street 06-012, Philadelphia, PA 19122

Inna Chervoneva

Division of Biostatistics, Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, 1015 Chestnut St, Suite M100, Philadelphia, PA 19107

Yihuan Xu: yihuan.xu@imclone.com; Boris Iglewicz: borisi@temple.edu; Inna Chervoneva: Inna.Chervoneva@jefferson.edu

Abstract

The g - and - h distributional family is generated from a relatively simple transformation of the standard normal and can approximate a broad spectrum of distributions. Consequently, it is easy to use in simulation studies and has been applied in multiple areas, including risk management, stock return analysis and missing data imputation studies. A rapidly convergent quantile based least squares (QLS) estimation method to fit the g - and - h distributional family parameters is proposed and then extended to a robust version. The robust version is then used as a more general outlier detection approach. Several properties of the QLS method are derived and comparisons made with competing methods through simulation. Real data examples of microarray and stock index data are used as illustrations.

Keywords

g - and - h distribution; indirect inference; least squares; maximum likelihood; outlier detection; quantiles; robust

1. Introduction

The g - and - h distributional family was introduced by Tukey (1977). The g - and - h distributed random variable is a monotone transformation of the standard normal variable. The p^{th} quantile of g - and - h random variable T is defined as:

© 2014 Elsevier B.V. All rights reserved.

¹Boris Iglewicz, Department of Statistics, Temple University, 1810 North 13th Street 06012, Philadelphia, PA 19122. borisi@temple.edu, fax: 215-204-1501, telephone: 215-204-8637.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

$$F^{-1}(p)=T_{A,B,g,h}(z_p)=A+B\frac{e^{gz_p}-1}{g}e^{hz_p^2/2}, B>0, h \geq 0, g \neq 0, \quad (1)$$

$$F^{-1}(p)=T_{A,B,g,h}(z_p)=A+Bz_p e^{hz_p^2/2}, B>0, h \geq 0, g=0, \quad (1a)$$

where, z_p is the p^{th} quantile of the standard normal distribution, A , B , g and h are parameters, with g controlling skewness and h elongation. This family and its extension have been studied by Martinez and Iglewicz (1984), Hoaglin (1985), Field and Genton (2006) and Rayner and MacGillivray (2002a and 2002b), with the latter providing maximum likelihood estimates of the parameters.

The g - and - h distributional family approximates a rich class of common univariate distributions which could accommodate different levels of skewness and elongation and can approximate a wide variety of commonly used distributional shapes, such as normal, lognormal, Weibull and exponential (Martinez and Iglewicz, 1984). It also approximates most of the Pearson family (Pearson 1895) and many other distributions that do not have finite first four moments, such as Cauchy distribution (Martinez and Iglewicz, 1984). This distributional family has been used in a variety of more recent publications, including Badrinath and Chatterjee (1988, 1991), Dutta and Babbell (2002), Dutta and Babbell (2005), Dutta and Perry (2007), Jiménez and Arunachalam (2011), Field (2004), He and Raghunathan (2006), Kafadar and Phang (2003), Mills (1995), Strandberg and Iglewicz (2013) and Walters and Ruscio (2009).

Hoaglin (1985) described the Tukey (1977) parameter estimation approach, which uses a letter value based sequence of quantiles to estimate the parameters sequentially. In this work, the slightly modified version of the letter value based method (LV) is used with following steps: (1) Estimate parameter A by the sample median; (2) Select $p_i = .005, .01,$

$0.025, .05, .10$, and $.25$ to calculate $g_{p_i} = -\frac{1}{z_{p_i}} \log\left(\frac{x_{1-p_i} - x_{.5}}{x_{.5} - x_{p_i}}\right)$, where x_p is the observed p^{th}

sample quantile; then estimate g by median of the g_{p_i} 's; (3) Regress $\log \log \left(\frac{g(x_{1-p} - x_{.5})}{e^{-gz_p} - 1} \right)$

or $\log \log \left(\frac{g(x_{.5} - x_p)}{1 - e^{gz_p}} \right)$ on $z_p^2/2$ depending on g positive or negative to estimate B and h . By adjusting quantile selections, Hoaglin (1985) demonstrated that a quantile estimation method works well for g and h estimation, especially in exploratory data analysis. Rayner and MacGillivray (2002a) derived numerical maximum likelihood estimation (NMLE) method for the generalized g - h or g - k distributions, which can be modified to provide reliable estimates for the g - and - h parameters. Rayner and MacGillivray (2002b) introduced weighted quantile-based estimation of this family, which is suitable for small sample sizes. Currently available g - and - h parameter estimation methods are not resistant to a moderate amount of contaminated observations or outliers. Meanwhile, NMLE is a computational time consuming method when data sample size is large. One aim of this paper is to provide two versions of a newly proposed and relatively simple iterative approach for estimating the g - and - h parameters, with the second option being more robust and thus more resistant to a moderate proportion of contaminated observations. Some theoretical properties of these two estimating methods are derived and a simulation comparison performed. The simulation studies indicate that these proposed methods have potential for use in estimating the g - and - h parameters, independent of potential outlier detection uses.

The second aim of this paper deals with applying the g - and - h distributional family as a univariate outlier detection method. In general, the detection of outliers deals with deviations from a specific distribution, such as normal. Given a distribution, there are a number of methods for detecting outliers. For example, Barnett and Lewis (1994) discuss almost 50 procedures for detecting outliers from a normal distribution. Some outlier detection methods are based on robust estimators and work for a variety of distributions, once distributional choice is made. A more general approach is suggested by first estimating robustly the g - and - h parameters and then using this fitted g - and - h distribution to detect outliers by utilizing a boxplot based approach first discussed in Hoaglin et al. (1986). The robustly estimated g - and - h method also can be combined with the Benjamini and Hochberg (1995) based false discovery rate approach. The proposed two stage outlier detection approach, of first estimating the g - and - h parameters and then using this estimated distribution and boxplot or false discovery rate to identify outliers, will be investigated through a simulation study and illustrated through examples from several published data sets.

The paper is organized as follows. Section 2 describes the proposed non-robust and robust quantile least squares (QLS) approaches and basic properties of the QLS estimator for the parameters of the g - and - h distributional family. In Section 3, a simulation study is reported comparing the LV, NMLE, QLS and robust QLS (rQLS) methods for estimation of the g - and - h distributional family parameters. Additionally, the estimates of the 95th and 99th percentiles are compared. The simulated random samples come from specific uncontaminated members of the g - and - h distributional family plus an additional 5% of the observations from different distributions. This allows for the comparison of estimation methods plus studying the effect of outliers on the estimates. Section 4 summarizes the boxplot outlier detection procedure and results of the simulation study performed to investigate the use of rQLS estimated distributions for detection of outlier. Several real data sets are used in Section 5 to illustrate g - and - h parameter estimation and outlier detection. Conclusions are found in Section 6.

2. QLS and rQLS Estimation of g - and - h Parameters

2.1 Proposed Quantile Least Squares (QLS) Procedure

Let $X_i, i = 1, \dots, n$, be i.i.d from Tukey g - and - h distribution with parameters $A, B, g, h, B > 0, h > 0$ and density function $f_{gh}(x/A, B, g, h)$. Let $0 < p_1 < \dots < p_m < 1$ be known pre-specified numbers. For any $i = 1, \dots, n$, let us denote by z_{p_i} the p_i^{th} quantile of $N(0,1)$ and by ξ_{p_i} the p_i^{th} quantile of g - and - h distribution,

$$\xi_{p_i} = A + Bg^{-1}\{gz_{p_i} - 1\} \exp\left(\frac{hz_{p_i}^2}{2}\right), \text{ if } g \neq 0;$$

$$\xi_{p_i} = A + Bz_{p_i} \exp\left(\frac{hz_{p_i}^2}{2}\right), \text{ if } g=0.$$

Let $\hat{\xi}_{p_i}$ be the p_i^{th} sample quantile, which may be broadly defined as the n_i^{th} order statistic

$x_{(n_i)}$ which satisfies $\lim_{n \rightarrow \infty} \frac{n_i}{n} = p_i$ (Mosteller, 1946). In particular, $n_i = \lceil np_i \rceil$ is used as the smallest integer not less than np_i . Let us denote the vector of unconstrained parameters by θ

$= [\theta_1, \theta_2, \theta_3, \theta_4]$, so that $A = \theta_1$, $B = \exp(\theta_2)$, $g = \theta_3$, $h = \exp(\theta_4)$ and the parameter space is $\Theta = \mathbb{R}^4$. Respectively, in this parameterization, the quantiles are

$$\xi_{p_i}(\theta) = \theta_1 + e^{\theta_2} \theta_3^{-1} \{ \exp(\theta_3 z_{p_i}) - 1 \} \exp \left(\frac{e^{\theta_4} z_{p_i}^2}{2} \right), \text{ if } \theta_3 = g \neq 0 \quad (2)$$

and $\xi_{p_i}(\theta | \theta_3 = 0) = \lim_{\theta_3 \rightarrow 0} \left\{ \theta_1 + \exp(\theta_2) \theta_3^{-1} \{ \exp(\theta_3 z_{p_i}) - 1 \} \exp \left(\frac{\exp(\theta_4) z_{p_i}^2}{2} \right) \right\}$, so that

$$\xi_{p_i}(\theta | \theta_3 = 0) = \theta_1 + e^{\theta_2} z_{p_i} \exp \left(\frac{e^{\theta_4} z_{p_i}^2}{2} \right). \quad (2a)$$

Further denote the vector of true quantiles by $\xi_m(\theta) = [\xi_{p_1}(\theta), \dots, \xi_{p_m}(\theta)]$ and the vector of sample quantiles by $\hat{\xi}_m = [\hat{\xi}_{p_1}, \dots, \hat{\xi}_{p_m}]$.

The QLS estimator $\hat{\theta}_{QLS}$ of θ is defined as

$$\hat{\theta}_{QLS} = \operatorname{argmin}_{\theta} \{ [\hat{\xi}_m - \xi_m(\theta)]^T [\hat{\xi}_m - \xi_m(\theta)] \} = \operatorname{argmin}_{\theta} \sum_{i=1}^m \{ \hat{\xi}_{p_i} - \xi_{p_i}(\theta) \}^2. \quad (3)$$

Theorem 1—Let X_i , $i = 1, \dots, n$, be a random sample from Tukey g - and h -distribution $f_{gh}(x|\theta_0)$ with the true parameter vector $\theta_0 = [\theta_1^0, \theta_2^0, \theta_3^0, \theta_4^0] \in \Theta = \mathbb{R}^4$ such that $A = \theta_1^0$, $B = \exp(\theta_2^0)$, $g = \theta_3^0$, $h = \exp(\theta_4^0)$. Under assumptions listed in the Appendix, the QLS estimator defined in (3) is consistent for estimating θ_0 and as $n \rightarrow \infty$

$$\sqrt{n}(\hat{\theta}_{QLS} - \theta_0) \xrightarrow{D} N_4 \left(0, \left(\sum_{\theta_0}^T \sum_{\theta_0} \right)^{-1} \sum_{\theta_0}^T V_{\theta_0}^{\xi_m} \sum_{\theta_0} \left(\sum_{\theta_0}^T \sum_{\theta_0} \right)^{-1} \right);$$

where $[\sum_{\theta_0}]_{ij} = \frac{\partial \xi_{p_i}(\theta)}{\partial \theta_j} \Big|_{\theta = \theta_0}$ are given in the Appendix and matrix $V_{\theta_0}^{\xi_m}$ has the entries

$$[V_{\theta_0}^{\xi_m}]_{jk} = \frac{\min(p_j, p_k) [1 - \max\{p_j, p_k\}]}{f_{gh}(\xi_{p_j}|\theta_0) f_{gh}(\xi_{p_k}|\theta_0)}. \quad (4)$$

The proof is given in the Appendix.

Since the QLS estimator may be viewed as an indirect estimator (Gourieroux et al., 1993), its robust properties are inherited from the robust properties of the auxiliary parameter vector of sample quantiles $\hat{\xi}_m$ (Genton and De Luna, 2000; Genton and Ronchetti, 2003). Theorem 1 in Genton and de Luna (2000) yields the influence function at the probability measure ν defining the contamination process F_ε

$$IF(v, \hat{\theta}_{QLS}, \theta_0) = \lim_{\varepsilon \rightarrow 0+} \frac{\hat{\theta}_{QLS}(F_\varepsilon) - \hat{\theta}_{QLS}(F_{gh})}{\varepsilon} = \left(\sum_{\theta_0}^T \sum_{\theta_0} \right)^{-1} \sum_{\theta_0}^T IF(v, \hat{\xi}_m, \theta_0),$$

where $IF(v, \hat{\xi}_m, \theta_0) = [IF(v, \hat{\xi}_{p_1}, \theta_0), \dots, IF(v, \hat{\xi}_{p_m}, \theta_0)]$ is the influence function of $\hat{\xi}_m$. For the Hampel's original definition of influence function with $v = \Delta_x$ (the measure with the point mass at x) and $F_\varepsilon = (1 - \varepsilon)F_{gh} + \varepsilon\Delta_x$, the influence function of $\hat{\xi}_{p_i}$ is given (Shao, 2003) by

$$IF(x, \hat{\xi}_{p_i}, \theta_0) = \left[F_{gh}(\xi_{p_i}|\theta_0) - I_{(-\infty, \xi_{p_i}]}(x) \right] / f_{gh}(\xi_{p_i}|\theta_0),$$

which is bounded for any fixed $p_1 > 0$ and $p_m < 1$. Consequently, the corresponding influence function of θ_{QLS} is bounded.

Furthermore, since the breakdown point of the indirect estimator is the same as the breakdown point of the auxiliary estimator (Genton and De Luna, 2000), the asymptotic breakdown point of θ_{QLS} is the $\min\{p_1, 1 - p_m\}$.

In the QLS estimating procedure, all information of the unknown parameters is captured through the sample quantile $\hat{\xi}_{p_i}$. Therefore, the selection of p_i and sample quantile $\hat{\xi}_{p_i}$ will determine the quality of QLS estimator. The following general rules are recommended for the selection:

- a. Both sides of median should have adequate p_i 's;
- b. There should be some p_i 's allocated within quarter range and in both sides of tails;
- c. Asymptotically, any consistent sample quantile estimator $\hat{\xi}_{p_i}$ works equivalently.

In this paper, $p_i = \frac{i-1/3}{m+1/3}$ (Hoaglin et al., 1983); the sample quantile $\hat{\xi}_{p_i}$ is estimated by order statistics $x_{(k_i)}$, where $k_i = \lceil np_i \rceil$. It is demonstrated in the simulations (Section 3) that this selection works well.

The selection of p_i here is different from LV method, because the emphasis of LV is on exploratory data analysis and on tail behavior, whereas QLS method focuses on the behavior of majority of the observations. The impact of such selection will be further discussed in Section 3.

In the proposed QLS estimator, quantiles serve as the auxiliary parameters (Gourieroux et al., 1993). Heggland and Frigessi (2004) showed that asymptotically, the highest efficiency is achieved when dimension of original and auxiliary parameter vectors are the same, e.g. $m = 4$. However, the simulation results did not support this asymptotic result when sample sizes are between 100 and 10,000, which cover most practical situations. Genton and Ronchetti (2003) mention that the Akaike information criterion (AIC) or Bayesian information criterion (BIC) may be used to select the number of auxiliary parameters. In this paper, it is selected to minimize AIC given by $AIC = n \log \left(\frac{SSE_{QLS}}{n} \right) + 2(m+1)$ (Burnham and Anderson, 1998), where $SSE_{QLS} = \sum_{i=1}^n \{\hat{\xi}_{p_i} - \xi_{p_i}(\theta)\}^2$, $p_i = \frac{i-1/3}{n+1/3}$ and $4 \leq m \leq 20$. Since extreme quantiles are easily impacted by outliers, $m > 20$ is not considered to improve robustness of the QLS estimator. Besides using the AIC for selection of m , simulations results using various a priori chosen m are also provided in supplemental Table 1.

Any effective non-linear optimization algorithm would work well when minimizing the non-linear function $\sum_{i=1}^m \{\hat{\xi}_{p_i} - \xi_{p_i}(\theta)\}^2$ to compute the QLS estimator. In this paper, the optim function in R with option "Nelder-Mead" is used for the minimization. The Nelder-Mead (Nelder and Mead, 1965) algorithm is a "simplex" direct search method, which doesn't

require derivatives, is easy to implement, and “has been consistently one of the most used and cited methods for unconstrained optimization” (Wright, 2012). In the Nelder-Mead minimization procedure, LV estimates of θ are used as the initial values.

2.2 Proposed Robust QLS Procedure

When data are contaminated with a fair number of outliers or observations from other distributions, these outliers will cause bias in estimating the parameters from the base distribution. Consequently, a robust estimation method is desired to fit g - and - h parameters well for the uncontaminated portion of the data. The QLS method is derived by minimizing the quantile least squares function (2) and (2a) with respect to parameter set θ . The robust version applies the QLS method to robustly estimated sample quantiles derived by trimming the outliers. This trimming procedure uses the Tukey biweight weighing function to trim the observations with weight 0.

The robust QLS (rQLS) estimator $\hat{\theta}_{rQLS}$ is defined as

$$\hat{\theta}_{rQLS} = \operatorname{argmin}_{\theta} \{ [\hat{\xi}_m^{TR} - \xi_m(\theta)]^T [\hat{\xi}_m^{TR} - \xi_m(\theta)] \} = \operatorname{argmin}_{\theta} \sum_{i=1}^m \{ \hat{\xi}_{p_i}^{TR} - \xi_{p_i}(\theta) \}^2,$$

Where $\hat{\xi}_m^{TR} = [\hat{\xi}_{p_1}^{TR}, \dots, \hat{\xi}_{p_m}^{TR}]$ and $\hat{\xi}_{p_1}^{TR}, \dots, \hat{\xi}_{p_m}^{TR}$ are sample quantiles described in section 2.1 but using the trimmed subset of order statistics. The trimming is performed using the Tukey biweight weights applied to the differences between the observed and predicted order statistics. More specifically let $x_{(i)}$, $i = 1, \dots, n$, be the i^{th} order statistic of the sample X_i , $i = 1, \dots, n$, from Tukey g - and - h distribution. The predicted i^{th} order statistic is defined as the quantile $\xi_{p_i}(\theta)$ of Tukey g - and - h distribution with parameter vector θ , where $p_i = \frac{i-1/3}{n+1/3}$. Then the weight corresponding to $x_{(i)}$ is given by

$$w_i = w(x_{(i)}) = \begin{cases} \left[1 - \left(\frac{x_{(i)} - \xi_{p_i}(\theta)}{c} \right)^2 \right]^2 & \text{for } |x_{(i)} - \xi_{p_i}(\theta)| \leq c \\ 0 & \text{for } |x_{(i)} - \xi_{p_i}(\theta)| > c \end{cases}, \quad (5)$$

where c is a constant of the Tukey biweight function. The order statistics with non-zero weights defined by (5) are included in the trimmed subset of order statistics $\{x_{(i)}, 1 \leq i \leq n'\}$, where n' is the number of observations with non-zero weights. This subset is then viewed as the full set of order statistics for the purpose of computing the sample quantiles

$\hat{\xi}_{p_1}^{TR}, \dots, \hat{\xi}_{p_m}^{TR}$ for $0 < p_1 < \dots < p_m < 1$. When the constant c is appropriately chosen, the extreme data values will be down weighted to zero under the true parameter θ . Therefore the sample quantiles can be estimated robustly using trimmed order statistics with non-zero weights. For a given constant c , rQLS estimator $\hat{\theta}_{rQLS}$ is computed using the following iterative procedure:

- I. Estimate the Tukey g - and - h parameters using a non-robust method, such as LV or QLS to obtain $\hat{\theta}_{rQLS}^{(0)}$.
- II. The following steps are repeated until convergence of estimates of θ .
- III. For a current estimate $\hat{\theta}_{rQLS}^{(i)}$, $i = 0$, compute the weights for order statistics using (5) with $\hat{\theta}_{rQLS}^{(i)}$.

- IV. Compute the sample quantiles $\hat{\xi}_{p_1}^{TR}, \dots, \hat{\xi}_{p_m}^{TR}$ using computed weights.
- V. Obtain new estimates $\hat{\theta}_{rQLS}^{(i+1)}$ using (4).

Since outliers may impact AIC, the empirical results suggest that using fixed $m = 10$ yields more robust rQLS estimates. Since there is no universal constant c in Tukey biweight function appropriate for a general g - and - h distribution, it is proposed to select the constant c adaptively. An appropriately chosen constant c should imply zero weight for outliers and non-zero weights for the regular observations, i.e. the observations at the end should have lower weight and observations in the middle have heavier weight. In practice, the population distribution is unknown; moreover, the sample might come from a mixed distribution or contaminated by outliers or extreme values. Consequently, it is not realistic to use a single fixed constant c to cover various $(A, B, g, h)'$ combinations. The simulation studies indicate that when outliers exist, a suitable c is in a range of $(a, b/2)$, where a is the smallest constant resulting in 50% observations with non-zero weight, and b is the smallest constant resulting in 100% observations with non-zero weight. Once this range is determined, a grading procedure is used to select the candidate values for c . The following procedure is used to adaptively select the constant c , which is used to compute the rQLS estimator.

1. Apply rQLS method using a few trial values for constant c ; check the number of trimmed observations when the procedure converges. Find a and b such that a is the smallest constant resulting in $< 50\%$ observations trimmed, and b is the smallest constant resulting in no observations trimmed. A large constant can be used here as a starting point and decreased until a and b are found. The heavier the tail, the larger a starting point is needed. For samples with LV estimated $h \approx 0, 0.1, 0.4$ and 1 , starting values can be $5, 10, 30$ and 1500 respectively. Here a and b need not be precise. For starting point 1500 , the searching process can use $1500, 1400$ and etc.
2. Define $c_i = b/2 - (i-1) \times v$, where $i=1, \dots, \left\lfloor \frac{b-a}{v} \right\rfloor + 1$; v is the desired grading interval, $10^{\lfloor \log_{10}(b/2) \rfloor - 1}$ is used here.
3. Use rQLS procedure to compute the robust estimator $\hat{\theta}_{rQLS}$ using $c = c_1$.
4. Repeat step 2 using $c = c_i, i=1, \dots, \left\lfloor \frac{b-a}{v} \right\rfloor + 1$ until stopping criteria are met.

The stopping criteria are set so that the weights at the lower/upper end should be smaller than in the middle. They are defined as:

1. Denote w_i the weight of order statistics $x_{(i)}$. Stop if $w_i \geq 0.8, i = 1, \dots, n$;
2. Otherwise: Stop if both (a) and (b) are met.
 - a. Denote $s_1 = \max(i, w_i < 0.7 \text{ and } i < \frac{n}{2})$ and $s_2 = \min(i, w_i > 0.8 \text{ and } i < \frac{n}{2})$; if $s_1 < s_2$ and $w_1 = \min\{w_i, i=1, \dots, \frac{n}{2}\}$;
 - b. Denote $s_3 = \max(i, w_i > 0.8 \text{ and } i > \frac{n}{2})$ and $s_4 = \min(i, w_i < 0.7 \text{ and } i > \frac{n}{2})$; if $s_3 < s_4$ and $w_n = \min\{w_i, i=\frac{n}{2}, \dots, n\}$.
3. Stop if $c < a$.

Note: if the sample is skewed or only one side is of interest, then one condition (a) or (b) applied to left or right skewed sample should be sufficient. The different cutoffs of 0.7 and 0.8 are chosen to avoid excluding too many regular observations; because the weights of regular observations could jump back and forth around 0.8 .

3. Simulation Study of QLS and rQLS Procedures

3.1 Comparison of LV, NMLE, QLS and rQLS performance in fitting uncontaminated data

Simulations are used in this section to compare the accuracy of g - and - h parameter estimation for four studied methods, LV, NMLE, QLS, and rQLS. Without loss of generality, it is assumed that $A = 0$ and $B = 1$ and the notation (g, h) is used to specify considered g - and - h distributions. Simulations of six uncontaminated g - and - h distributions $(0, 0)$, $(0, 0.1)$, $(0, 0.4)$, $(0.1, 0)$, $(0.4, 0)$, $(0.2, .02)$ were carried out. The selection of these combinations covers the most commonly used distributions with moderate tails. For each parameter combination, 1000 standard normal random samples were generated with distinct seeds using 'rnorm' function in R, and then transformed to g - and - h distribution sample using formula (1). In this simulation study, sample sizes $n = 100$, $n = 1,000$ and $n = 10,000$ were considered; Table 1 shows comparison of results for uncontaminated samples fitted by LV, NMLE, QLS and rQLS for $n = 100$ and $n = 1,000$. Selected results for $n = 10,000$ fitted by rQLS are presented in Table 3 as part of outlier detection study.

In Table 1, for each (g, h) case, the upper entries represent the average biases between the estimated and true population parameters, while the entries below, in parenthesis, represent their standard errors. Thus, for $n = 100$, estimating g using the LV method and the $(0, 0)$ case, standard normal, the average bias = -0.0044 and standard error = 0.0031 . The third column for each procedure gives percent error for estimating the 95th and 99th percentiles placed above and below, respectively. The percentiles are computed as $\frac{|x_q - y_q|}{x_q} \times 100\%$, where x_q is the q^{th} quantile of underlying distribution, while y_q is the estimated value based on the fitted g - and - h distribution. For QLS, m is selected by minimizing AIC as described in section 2.1; $m = 10$ is used for rQLS.

From Table 1 it is clear that all procedures perform well in estimating the (g, h) parameters for uncontaminated samples, with NMLE generally having the lowest standard errors. Regarding the estimation of the 95th and 99th percentiles, QLS is competitive with NMLE, with both being superior to the LV method, which was used to obtain the starting values for the NMLE and QLS iterative procedures.

When sample size increases from 100 to 1,000, the patterns are quite similar, but with improved efficiency. Consider, as an example, the estimation of the (g, h) parameters for $(0, 0)$. Then the (g, h) standard errors are $(0.0029, 0.0012)$ for QLS and $n = 100$, while they become $(.0009, .0003)$ for $n = 1,000$. Supplementary Table 3 indicates that the magnitude of bias and standard errors for estimation of A and B by QLS and rQLS have similar pattern as the estimation of g and h .

To ensure one-to-one transformation, h was chosen to be a non-negative number, but some commonly used distributions can be approximated using g - and - h distributions with negative h (Martinez and Iglewicz, 1984). When h is negative, the distributions are not proper and any approximations to distributions can be only partial. The observations at the tail may not be fitted properly. Since QLS method does not use the observations at the very tail, like the LV method, the QLS can be modified to allow negative estimates of h by using parameterization $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]$, so that $A = \theta_1$, $B = \exp(\theta_2)$, $g = \theta_3$, $h = \theta_4$. The corresponding simulation results (supplemental Table 2) indicate that using this parameterization provides less biased estimates of $h = 0$ as compared to using the QLS approach with assumption $h > 0$. This property makes QLS more flexible to fit some commonly used distributions such as Chi-squared distribution.

3.2 Comparison of LV, NMLE, QLS and rQLS performance in fitting contaminated data

To study the QLS and rQLS procedure in sample with moderate amount of outliers in high end tail, for each of the above uncontaminated cases described in section 3.1, additional 5% random observations consisting of $N(\mu, 0.5)$, where the 0.5 represents the standard deviation, were added. The values for μ are 5 for the (0, 0) case, 17.5 for the (0, 0.1) case, 742 for the (0, 0.4) case, 6.5 for the (0.1, 0) case, 16.5 for the (0.4, 0) case, and 105 for the (0.2, 0.2) case, respectively. Each of these μ represents the corresponding upper tail area about 2.9×10^{-7} . In these simulations, $m = 10$ is used for both QLS and rQLS. The simulation results comparing all four methods are shown in Table 2. Some other simulations with outliers at both low end and high end were also explored; the results are similar to the simulations with high end outliers and not presented.

Table 2 indicates considerably smaller average bias of rQLS estimates as compared to the other three methods. Similarly, the percent errors in estimating the 95th and 99th percentiles are considerably smaller than those based on the other three methods. This suggests that rQLS works well in estimating the g - and - h parameters when a moderate proportion of outlying observations are present. The other three methods treat the entire data set as representing the true distribution and fit the parameters accordingly.

Table 2 also demonstrates the effects of the different emphasis of LV, NMLE, QLS and rQLS methods, especially when outliers exist. Since QLS and rQLS use a selected subset of data quantiles from the body of data, they are more resistant to the outliers in the tail; LV approach, using data quantiles from the quartiles out, with ever-decreasing ‘probability’ steps, fits the tail better; NMLE uses all observations to fit the distribution and works well when no outliers exist.

The simulations were carried out also to explore the breakdown point of the rQLS method numerically. For a given random sample with $n = 10,000$, different percentage of observations were moved to infinity (10,000 used here); then rQLS method was used to fit the g - and - h distribution. The results from different g - and - h distributions suggest that the actual breakdown point is near 50%.

Among the considered methods, the LV method requires the least computing time and the LV estimates were used as starting values for the other three procedures. When sample size is 1,000, using a PC with Intel Core i3-330M dual-core processor and 4GB RAM, the QLS approach requires on average about 0.1 second per sample, compared to about 2 minutes for NMLE. When sample size increases to 10,000, the running time for QLS and NMLE increases to 0.2 second and 15 minutes per sample, respectively. The NMLE computing time is approximately proportional to the sample - size. This property makes QLS a better approach for large samples. The rQLS usually takes 5–10 times as long as the QLS method, depending on the amount and location of the contaminants.

4. Outlier Detection Using rQLS Estimated g - and - h Distribution

4.1 Outlier Detection Procedure for g - and - h Distributional Family

The detection of outliers has a long history and many practical applications. A detailed reference is Barnett and Lewis (1994) and a far shorter introduction given by Iglewicz and Hoaglin (1993). This work deals with univariate data. In general one assumes that the bulk of the sample comes from a specific distribution with the rest of the observations, if any, deviating sufficiently from these observations to be labeled as outliers. Often it is assumed that the base distribution is normal and then uses this distribution to identify outliers in the sample. Thus, Barnett and Lewis (1994) lists almost 50 procedures for identifying outliers from normal distribution, with a fair number of these procedures dealing with the

identification of one or two outliers. Of interest here are procedures that have the ability to identify multiple outliers. The approach is to first use the proposed robust version, rQLS, to estimate the g - and - h parameters and then use this estimated distribution as the base distribution for identifying potential outliers. For this purpose the simple robust Banerjee and Iglewicz (2007) procedure will be used, although other methods can also be considered.

For a given relatively large sample $X_1, \dots, X_n$, from a univariate distribution $F(\cdot)$, the boxplot outlier detection procedure (Hoaglin et al., 1986; Hoaglin and Iglewicz, 1987) defines the upper outlier cutoff for right skewed distribution as

$$Q_3 + k(Q_3 - M), \quad (6)$$

and lower and upper outlier cutoff for symmetric distributions as

$$Q_1 - k(Q_3 - Q_1) \quad \text{and} \quad Q_3 + k(Q_3 - Q_1), \quad (7)$$

respectively; where M is the sample median and Q_1 and Q_3 are the first and third sample quartiles respectively.

When sample size is sufficiently large, Banerjee and Iglewicz (2007) provided the solution for k ,

$$k = \frac{F^{-1}\left((1-\alpha)^{1/n}\right) - F^{-1}(.75)}{F^{-1}(.75) - F^{-1}(.5)}, \quad (8)$$

$$k = \frac{F^{-1}\left((1-\alpha/2)^{1/n}\right) - F^{-1}(.75)}{F^{-1}(.75) - F^{-1}(.25)}, \quad (9)$$

for right skewed and symmetric distribution respectively; where $F^{-1}(p)$ is the p^{th} quantile of a given distribution (Banerjee and Iglewicz, 2007). This boxplot based procedure will be here denoted by BP. For BP, $\alpha = .05$ will be used. Once the g - and - h parameters A , B , g and h are estimated for a given sample, the constant k in (6) and (7) is easily calculated using (8), (9) and (1) to obtain the outlier cutoff.

4.2 Simulation Results

Table 3 reports simulation results for the same cases as in Table 1 and 2, but with $n = 10,000$, since the real data examples involve relatively large samples. Thus, the simulated data include 10,000 random observations for uncontaminated cases and 10,000 random uncontaminated observations plus 500 “upper outliers” for contaminated cases. A comparison is made between applications of boxplot detection procedure utilizing three possibilities of computing quantiles in (8): (a) Using the fitted g - and - h distribution with rQLS parameter estimates; (b) using the normal distribution with parameter estimates from data normally distributed in (8); (c) Using the true g - and - h distribution. Table 3 shows the average number of regular observations labeled as outliers, the average number of contaminated observations labeled as outliers for contaminated cases, and the estimated probability that at least one regular observation in a sample is labeled as an outlier (some-outside rate per sample). In Table 3 the base distribution results can be viewed as ideal for purposes of making comparisons.

A simulation study, summarized in Table 3 indicates that the number of real outliers (contaminated observations) identified by outlier labeling procedure based on rQLS estimated g - and - h distribution is comparable to that based on known distribution for all simulation scenarios. Meanwhile, the g - and - h method successfully controls the some-outside rate per sample to around 5% in studied scenarios, which is comparable to that based on using the known distribution. When the true distributions is not normal, the procedure based on normal distribution does well at detecting the real outliers, but at the cost of detecting a fair number of regular observations as outliers. This indicates the drawback of assuming that data comes from normal distribution with possibly some outliers that need to be detected, when the observations actually come from a distribution that differs considerably from the normal. Overall, Table 3 lends support for using the proposed outlier detection rule consisting of first fitting a g - and - h distribution using rQLS to estimate parameters and then using formula (6) to identify outliers. This proposed approach works well for relatively large samples.

The above outlier detection approach is based on control of the rather conservative familywise error rate, Hochberg and Tamhane (1987), where family is defined as “any collection of inferences for which it is meaningful to take into account some combined measure of error”. More recently, methods based on control of the false discovery rate (FDR), introduced by Benjamini and Hochberg (1995), have become popular, especially in identifying expressed genes in microarray experiments. FDR is designed to control the expected proportion of incorrectly rejected null hypotheses (type I errors). Among the large number of papers on this topic are Reiner et al. (2003), Qian and Huang (2005) and Kauffmann and Huber (2010). This procedure can be modified to use for outlier detection. The following steps are used in this paper: (1) For every observation, the z -score based on rQLS estimated g - and - h parameters is calculated by solving for z_p in formula (1); (2) Convert the z -score to original p - value and adjust these p - values using “p.adjust” function in R with “BH” option; (3) Declare the observations with adjusted p - value $< .01$ or $.05$ as outliers. For a brief study of the false discovery approach, the (0, 0.1) case from Table 3 was repeated by first estimating the (g , h) parameters using rQLS and then using the Benjamini and Hochberg (1995) procedure (Call this procedure BH). Then for the 500 contaminated observations, BH identifies all of them but at the cost of identifying, on average, 24.3 regular observations as outliers, versus 0.03 by BP. Realizing this, p - values $< .01$, instead of $.05$ are frequently used with false discover rate procedure, see, e.g. Kauffmann and Huber (2010). Denote the 1% version as BH1 and the 5% version as BH5, respectively. If BH5 is now replaced by BH1, then again on average, all 500 contaminated observations were labeled as outliers, at a cost of identifying, on average, 4.7 regular observations as outliers. Both BH and BP outlier identification methods will be illustrated with three published data sets in Section 5. In summary, rQLS can be used to estimate a g - and - h distribution and then effectively use this estimated distribution as the basis for identifying outliers either using the familywise or false discovery rate approaches.

5. QLS and rQLS Applications

In this section three data sets are used to illustrate application of the proposed QLS and rQLS estimation procedures, two involving gene expression microarray data and the third stock market index data. All three data sets involve between 6,000 and 18,000 observations. So, in that sense they can be viewed as large data sets. These data sets will be fitted by LV, NMLE, QLS and rQLS, using the bootstrap to estimate each estimated parameter’s bias, standard error, and confidence interval, plus p - value of Kolmogorov-Smirnov (KS) test for goodness of fit. These results are summarized in Table 4. Outlier detection using both the BP and BH procedures will be investigated.

The first data set, called ApoAI here, is from Callow et al. (2000) who investigated the gene expression profile in HDL deficient mice with microarray technique. This study compared the effects of ApoAI deficiency on gene expression in the liver in order to investigate the molecular pathway of ApoAI in regulating HDL metabolism. The raw data contained the measurements of Cy5 and Cy3 signal from 6384 genes on 16 microarrays. The data set used here consists of the estimated mean differences in log ratio of Cy5 to Cy3 for each gene after normalization using LIMMA software package (Smyth, 2005).

The second data set, called Swirl here, used zebrafish as a model to identify genes with differential expression in the Swirl mutant zebrafish compared to wild-type zebrafish. The data illustrated here comes from package LIMMA (Smyth, 2005). Briefly, their final data, consisting of 8448 observations, was transformed from raw data by normalization, linear model and empirical Bayes analysis. This univariate data was theoretically t distributed with 7 degrees freedom. An independent analysis indicates that a t distribution with 7 degrees of freedom does not fit so well, so (g, h) estimated distribution will be relied on.

The third data set comes from Strandberg and Iglewicz (2013) and consists of 17,310 modified daily percent changes, in full five day weeks, of Dow Jones Industrial Average (DJIA) values from October 1, 1928 to December 18, 2009. Supplemental web figures are provided to illustrate the histogram of these three datasets.

The QLS, rQLS, LV and NMLE methods are used to estimate the g - and - h parameters for all three datasets. Bootstrapping was applied to re-sample from original data 1000 times with replacement to obtain the estimated bias, standard error and 95% confidence interval for each estimated parameter. Additionally, the Kolmogorov-Smirnov Goodness-of-fit test (KS test), here used as an exploratory tool, was performed for each estimated distribution. Table 4 summarizes the results. Additionally, Figure 1 provides Q-Q plots of NMLE, QLS quantiles versus sample quantiles, and rQLS estimated quantiles versus trimmed sample quantiles for DJIA and Swirl data. Q-Q plot for all examples with LV, NMLE, QLS and rQLS fitted quantiles are provided as web supplements. For ApoAI data, all considered estimators resulted in very similar Q-Q plots (Supplemental Figure 2). It was observed from Table 4 that all four g - and - h estimated distributions fit nicely the ApoAI sample data, as seen from KS tests, but with slightly elevated NMLE estimate of h . This occurs because NMLE method uses all observations to estimate g - h parameters; whereas LV, QLS and rQLS only use selected quantiles. Consequently, NMLE is not resistant to outliers. Furthermore, use of QLS and rQLS provide reasonable distributional fits for both the Swirl and DJIA cases, while neither LV nor NMLE were able to do so for the DJIA case and LV failed to provide a reasonable fit for the Swirl case. As seen from Table 4, these three data sets seem to consist of mixtures of mostly regular observations that are contaminated with outliers. This is especially true for the Swirl and DJIA cases, where both QLS and rQLS provided reasonably fitted distributions for the data sets. Note that the rQLS method fits majority of observations (about 95%) in the middle better than in the tails. Figure 1 demonstrates that rQLS fits the trimmed sample well. Note that only 0.46% of the Swirl data and 0.83% of the DJIA data were actually trimmed for the purpose of computing the final rQLS estimates. Q-Q plots of QLS for both Swirl and DJIA data look worse than NMLE, because the Q-Q plot cannot reflect the robustness of QLS, as illustrated in supplemental Figure 5.

Both the BP and BH outlier identification procedures are used to identify lower and upper ends outliers for the three data sets. For BH approach both the 5% and 1% significance levels are used, while only 5% is used for the BP approach. Eight outliers are identified by both the BP and BH approaches for the ApoAI data with the observations, following removal of these eight outliers, fitting nicely a normal distribution. The number of outliers

found for the other two data sets differ depending on method. For the Swirl data set, 14 outliers are found using BP, 22 using BH1 and 43 based on BH5; here the BP outliers are included in BH1 outliers. Similarly, the numbers of outliers found for the DJIA case are 21, 27, 52 for BP, BH1 and BH5, respectively. This time the BH1 and BP outliers are both subsets of BH5 outliers; BP identifies 4 lower end outliers ($< -35.5\%$) corresponding to daily percent change during the Great Depression in 1929; BH1 only identified positive changes, despite BH5 identified both ends outliers. As illustrated for the Swirl and DJIA data sets, the number of outliers identified depends on whether BP or BH are used and whether 1% or 5% are used for BH.

6. Conclusions and discussion

A quantile least square method (QLS) and a robust version, rQLS, are proposed to estimate the g - and - h parameters. Meanwhile, the same estimation approaches may be used to estimate parameters of any distribution with known quantile functions. Similar estimators may be also developed for any parametric distribution (with unknown quantile functions) using the framework of simulation-based indirect inference (Gourieroux and Monfort, 1996). The multivariate extension of g - and - h distribution was recently developed by Field and Genton (2006). They used the estimation procedure similar to QLS (based on empirical multivariate quantiles) for estimating multivariate versions of g and h parameters after estimating location and scale parameters using the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985), but did not address the asymptotic properties of the resulting estimators.

The QLS is an efficient and rapidly convergent estimator of the Tukey g - and - h distributional parameters and rQLS provides a robust version that works well when contaminated observations are present. These two approaches are compared with two alternative procedures of LV and NMLE. The simulation results indicate that these two methods are competitive for data coming from a random sample, with both QLS and rQLS providing similar estimates of the parameters.

The QLS and rQLS approaches also enable the robustly estimated g - and - h distribution as a basis for identifying outliers in the data. When data includes some contaminated observations, QLS does well at fitting the distribution of the entire sample, while rQLS does well at fitting the observations coming from the uncontaminated bulk of the data set. The rQLS method is further used to provide a more general version of an outlier identification procedure, by first fitting the g - and - h parameters and then using this fitted distribution as a means of identifying outliers. This more general outlier identification approach is investigated, through simulation, both for a familywise error rate method, BP, and the BH false discovery rate approach. The QLS and rQLS methods are further illustrated with three data sets, two involving microarray data where expressed genes are of interest, and one dealing with stock market index data. These provide additional evidence that the QLS and rQLS estimation methods can be effectively used to estimate the g - and - h parameters and rQLS can also be used as the basis for obtaining an estimated base distribution that can be successfully utilized to identify potential outliers.

The proposed robust version of QLS may be viewed as a robust indirect estimator (Genton and de Luna, 2000; Genton and Ronchetti, 2003). However, the asymptotic properties of rQLS estimator do not follow immediately from the general theory of indirect inference because the consistency and asymptotic normality of the sample quantiles based on trimmed order statistics has not been established. Deriving such asymptotic distribution is far beyond the scope of this work and will be a subject of a separate investigation. As an alternative to using trimmed order statistics, it is possible to adapt M-quantiles (Breckling and Chambers,

1988) as a robust auxiliary parameter. However, the M-quantiles are monotone M-estimators and may be expected to be less robust as compared to using redescending Tukey biweight. Also, the joint asymptotic distributions of the vector of M-quantiles, which would enable derivation of asymptotic distribution of corresponding indirect estimator, has not been established either. For another potentially highly robust alternative to QLS, one may consider using the least median of squares (Rousseeuw, 1984) instead of the least sum of squares in (3) as an objective function to be minimized. This direction was not pursued here because computing such least median of squares estimator for large data sets is expected to be really challenging (e.g. Steele and Steiger, 1986; Stromberg, 1993).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank the Editor, Associate Editor, and reviewers for their careful review and valuable suggestions. In addition, the authors wish to thank David Hoaglin for valuable editing the pre-submission draft and Zhigen Zhao for making useful suggestions. Inna Chervoneva was partially supported by the grant from the National Institutes of Health (NIH/NCI R21CA159816).

References

- Badrinath SG, Chatterjee S. On Measuring Skewness and Elongation in Common Stock Return Distributions: The Case of the Market Index. *Journal of Business*. 1988; 61(4):451–472.
- Badrinath SG, Chatterjee S. A Data-Analytic Look at Skewness and Elongation in Common-Stock-Return Distributions. *Journal of Business & Economic Statistics*. 1991; 9(2):223–233.
- Banerjee S, Iglewicz B. A simple Univariate Outlier Identification Procedure Designed for Large Samples. *Communications in Statistics - Simulation and Computation*. 2007; 36:249–263.
- Barnett, V.; Lewis, T. *Outliers in Statistical Data*. 3. New York: Wiley; 1994.
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of Royal Statistical Society, B*. 1995; 57(1):289–300.
- Breckling J, Chambers R. M-quantiles. *Biometrika*. 1988; 75(4):761–771.
- Burnham, KP.; Anderson, DR. *Model Selection and Inference*. Springer-Verlag; NY: 1998.
- Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*. 2000; 10:2022–2029. [PubMed: 11116096]
- Dutta, KK.; Babbal, DF. On measuring skewness and kurtosis in short rate distributions: the case of the US dollar London Inter-Bank Offer Rates. University of Pennsylvania, Wharton School Financial Institutions Center; 2002. Working Paper 02–25
- Dutta KK, Babbal DF. Extracting probabilistic information from the prices of interest rate options: tests of distributional assumptions. *Journal of Business*. 2005; 78(3):841–870.
- Dutta, K.; Perry, J. A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital. Federal Reserve Bank of Boston working paper 06–13. 2007. <http://www.bos.frb.org/economic/wp/index.htm>
- Field CA. Using the *gh* distribution to model extreme wind speeds. *Journal of Statistical Planning and Inference*. 2004; 122(1):15–22.
- Field C, Genton MG. The multivariate g-and-h distribution. *Technometrics*. 2006; 48:104–111.
- Genton MG, de Luna X. Robust simulation-based estimation. *Statistics and Probability Letters*. 2000; 48:253–259.
- Genton MG, Ronchetti E. Robust indirect inference. *Journal of the American Statistical Association*. 2003; 98:67–76.
- Gourieroux, C.; Monfort, A. *Statistics and Econometric Models*. Vol. 2. Cambridge University Press; London: 1996.

- Gourieroux C, Monfort A, Renault E. Indirect inference. *J Appl Economet*. 1993; 8:S85–S118.
- He YL, Raghunathan T. Tukey's g and h distribution for multiple imputation. *American Statistician*. 2006; 60(3):251–256.
- Heggland K, Frigessi A. Estimating functions in indirect inference. *J R Statist Soc B*. 2004; 66:447–62.
- Hoaglin, DC. Summarizing Shape Numerically: The g - and - h Distributions. In: Hoaglin, DC.; Mosteller, F.; Tukey, JW., editors. *Exploring Data Tables, Trends, and Shapes: Robust and Exploratory Technique*. John Wiley & Sons Inc; New York: 1985. p. 461-513.
- Hoaglin DC, Iglewicz B. Fine tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*. 1987; 82:1147–1149.
- Hoaglin DC, Iglewicz B, Tukey JW. Performance of some resistance rules for outlier labeling. *Journal of American Statistical Association*. 1986; 81:991–999.
- Hoaglin, DC.; Mosteller, F.; Tukey, JW. *Understanding robust and exploratory data analysis*. John Wiley and Sons; New York: 1983.
- Hochberg, Y.; Tamhane, AC. *Multiple Comparison Procedures*. John Wiley and Sons; New York: 1987.
- Iglewicz, B.; Hoaglin, DC. *How to detect and handle outliers*. ASQC Quality Press; Milwaukee, Wisconsin: 1993.
- Jiménez JA, Arunachalam V. Using Tukey's g and h family of distributions to calculate value-at-risk and conditional value-at-risk. *Journal of Risk*. 2011; 13(4):95–116.
- Kafadar K, Phang T. Transformations, background estimation, and process effects in the statistical analysis of microarrays. *Computational Statistics and Data Analysis*. 2003; 44:313–338.
- Kauffmann A, Huber W. Microarray data quality control improves the detection of differentially expressed genes. *Genomics*. 2010; 95:138–142. [PubMed: 20079422]
- Martinez J, Iglewicz B. Some properties of the Tukey g and h Family of distributions. *Communications in Statistics - Theory and Methods*. 1984; 13(3):353–369.
- Mosteller F. On some useful "inefficient" statistics. *Ann Math Statist*. 1946; 17(4):377–408.
- Mills TC. Modeling Skewness and Kurtosis in the London Stock Exchange FT-SE Index Return Distributions. *The Statistician*. 1995; 44(3):323–332.
- Nelder JA, Mead R. A simplex method for function minimization. *Computer Journal*. 1965; 7:308–313.
- Pearson K. Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous materia. *Philosophical Transactions of the Royal Society of London*. 1895; 186:343–414.
- Qian H, Huang S. Comparison of false discovery rate methods in identifying genes with differential expression. *Genomics*. 2005; 86(4):495–503. [PubMed: 16054333]
- Rayner GD, MacGillivray HL. Numerical maximum likelihood estimation for the g -and- k and generalized g -and- h distributions. *Statist Comput*. 2002a; 12:57–75.
- Rayner GD, MacGillivray HL. Weighted quantile-based estimation for a class of transformation distributions. *Computational Statistics and Data Analysis*. 2002b; 29(4):401–433.
- Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*. 2003; 19(3):368–375. [PubMed: 12584122]
- Rousseeuw PJ. Least median of squares regression. *Journal of the American statistical association*. 1984; 79(388):871–880.
- Rousseeuw PJ. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*. 1985; B:283–297.
- Shao, J. *Mathematical Statistics*. 2. Springer; New York: 2003.
- Steele JM, Steiger WL. Algorithms and complexity for least median of squares regression. *Discrete Applied Mathematics*. 1986; 14(1):93–100.
- Strandberg AG, Iglewicz B. A Nonparametric Test for Deviation from Randomness with Applications to Stock Market Index Data. *Communications in Statistics - Simulation and Computation*. 2013; 42:686–697.
- Stromberg AJ. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM Journal on Scientific Computing*. 1993; 14(6):1289–1299.

- Smyth, GK. Limma: linear models for microarray data. In: Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W., editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer; New York: 2005. p. 397-420.
- Tukey, JW. Modern techniques in data analysis. NSF-sponsored regional research conference; North Dartmouth, MA: Southeastern Massachusetts University; 1977.
- Walters GD, Ruscio J. To sum or not to sum: Taxometric analysis with ordered categorical assessment items. *Psychological Assessment*. 2009; 21(1):99–111. [PubMed: 19290770]
- Wright MH. Nelder, Mead, and the other simplex methods. *Documenta Mathematica Extra Volume: Optimization Stories*. 2012:271–276.

APPENDIX

1. Assumptions

It is assumed that $m \geq 4$ and $0 < p_1 \dots < p_m < 1$ are selected so that

A1 The mapping $\theta \rightarrow \xi_m(\theta)$ defined by (2) is locally injective ($\theta_1 \neq \theta_2 \Rightarrow \xi_m(\theta_1) \neq \xi_m(\theta_2)$) in some neighborhood of the true parameter θ_0 .

A2 The $m \times 4$ Jacobian matrix Σ_θ of the partial derivatives $[\Sigma_\theta]_{ij} = \frac{\partial \xi_{p_i}(\theta)}{\partial \theta_j}$ has column rank 4 at the true parameter vector θ_0 .

2. Proof of Theorem 1

Since g - and h distribution satisfies Conditions 1 and 2 in Mosteller (1946), his Theorem 2 provides that under the true parameter θ_0 and $n \rightarrow \infty$, vector $\hat{\xi}_m$ is consistent for estimating $\xi_m(\theta_0)$ and

$$n^{1/2} \{ \hat{\xi}_m - \xi_m(\theta_0) \} \xrightarrow{D} N_m(0, V_{\theta_0}^{\xi_m}) \text{ as } n \rightarrow \infty,$$

where the entries of matrix $V_{\theta_0}^{\xi_m}$ are given in (4).

Hence, the QLS estimator $\hat{\theta}_{QLS}$ may be viewed as an indirect estimator (Gourieroux, et al., 1993; Gourieroux and Monfort 1996) with the auxiliary parameter vector $\xi_m(\theta)$, $\sqrt{n}$ -consistent estimator $\hat{\xi}_m$ of $\xi_m(\theta)$, and explicit binding function given by (2) for $\theta_3 = g = 0$ and by (2a) when $\theta_3 = g \neq 0$. The indirect estimator is computed as

$$\hat{\theta} = \arg \min_{\theta} \{ (\hat{\xi}_m - \xi_m^*(\theta, S))^T (\hat{\xi}_m - \xi_m^*(\theta, S)) \}, \quad (\text{a.1})$$

where $\xi_m^*(\theta, S)$ is a consistent estimator of $\xi_m(\theta)$ obtained using a large number S simulated data sets under parameter θ (the mean of estimated $\hat{\xi}_m(\theta)$ for S data sets simulated from $f_{gh}(\theta)$). For a fixed S , the consistency and asymptotic normality of indirect estimator $\hat{\theta}$ follows from the consistency and asymptotic normality of $\hat{\xi}_m$ (Gourieroux et al., 1993). Since we have the explicit expressions (2) and (2a) for the binding function (not generally the case for indirect inference), we can view the objective function $R(\theta, \hat{\xi}_m) = (\hat{\xi}_m - \xi_m(\theta))^T (\hat{\xi}_m - \xi_m(\theta))$ as a limit of (a1) as $S \rightarrow \infty$ with $\xi_m(\theta)$ computed directly for each θ using (2) and (2a).

The elements of the Jacobian matrix Σ_θ for $\theta_3 = g = 0$, are given by

$$\begin{aligned}\frac{\partial \xi_{p_i}}{\partial \theta_1} &= 1 \\ \frac{\partial \xi_{p_i}}{\partial \theta_2} &= e^{\theta_2} \frac{\exp(\theta_3 z_{p_i}) - 1}{\theta_3} \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \\ \frac{\partial \xi_{p_i}}{\partial \theta_3} &= e^{\theta_2} \frac{\theta_3 z_{p_i} \exp(\theta_3 z_{p_i}) - \theta_3 z_{p_i} + 1}{\theta_3} \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \\ \frac{\partial \xi_{p_i}}{\partial \theta_4} &= e^{\theta_2} \frac{\exp(\theta_3 z_{p_i}) - 1}{\theta_3} \frac{1}{2} z_{p_i}^2 \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right)\end{aligned}$$

and for $\theta_3 = g = 0$,

$$\begin{aligned}\frac{\partial \xi_{p_i}}{\partial \theta_1} &= 1 \\ \frac{\partial \xi_{p_i}}{\partial \theta_2} &= \lim_{\theta_3 \rightarrow 0} \left\{ e^{\theta_2} \frac{\exp(\theta_3 z_{p_i}) - 1}{\theta_3} \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \right\} e^{\theta_2} z_{p_i}^2 \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \\ \frac{\partial \xi_{p_i}}{\partial \theta_3} &= \lim_{\theta_3 \rightarrow 0} \left\{ e^{\theta_2} \frac{\theta_3 z_{p_i} \exp(\theta_3 z_{p_i}) - \theta_3 z_{p_i} + 1}{\theta_3} \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \right\} = \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \\ \frac{\partial \xi_{p_i}}{\partial \theta_4} &= \lim_{\theta_3 \rightarrow 0} \left\{ e^{\theta_2} \frac{\exp(\theta_3 z_{p_i}) - 1}{\theta_3} z_{p_i}^2 \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right) \right\} = \frac{1}{2} e^{\theta_2} z_{p_i}^2 \exp\left(\frac{e^{\theta_4} z_{p_i}^2}{2}\right)\end{aligned}$$

Thus, the Jacobian matrix Σ_{θ} of the binding function exist and has column rank 4 at the true parameter vector θ_0 under assumption **A2**. Hence, Proposition 4.2 in Gourieroux and Monfort (1996) implies that the indirect estimator (a.1) is consistent for estimating θ_0 and asymptotically normal under assumptions (A.1)-(A.8) in the Appendix 4A of Gourieroux and Monfort (1996). Assumptions (A1)-(A4) and (A5)-(A7) in Gourieroux and Monfort (1996) are needed to ensure consistency of the auxiliary parameter estimator, and they are satisfied for the sample quantile estimator ξ_m . Assumption (A5) in Gourieroux and Monfort (1996) is equivalent to our assumption **A1**.

Since (a.1) is a continuous function of $\xi_m^*(\theta, S)$ and $\xi_m^*(\theta, S)$ is consistent for estimating $\xi_m(\theta)$, (a.1) converges to (3) in probability as $S \rightarrow \infty$. Hence, as $n \rightarrow \infty$, $\hat{\theta}_{QLS}$ is consistent for estimating θ_0 and has asymptotic normal distribution with the covariance matrix

$(\sum_{\theta_0}^T \sum_{\theta_0})^{-1} \sum_{\theta_0}^T V_{\theta_0}^{\xi_m} \sum_{\theta_0} (\sum_{\theta_0}^T \sum_{\theta_0})^{-1}$, which is obtained as a limit as $S \rightarrow \infty$ of the asymptotic covariance matrix of θ as derived in Proposition 4.2 of Gourieroux and Monfort (1996).

Highlights

- Introduced simple methods to estimate the g-and-h distributional parameters.
- Proved consistency and asymptotic normality.
- Effective robust version introduced.
- Robust version used to obtain base distribution for outlier detection.
- Illustrated use of proposed methods to multiple fields.

Figure 1. The quantile-quantile plot of Swirl and DJIA daily percent change data fitted by different methods.

Table 1

Comparisons of Bias and (Standard Error) for g - and h - distribution parameter estimates from methods including LV, NMLE, QLS and rQLS. The random samples were generated from standard normal distribution and transformed to g - and h - distribution using the selected parameters g and h . The error rate (%) at 95th (upper row) and 99th (lower row) quantiles were calculated by $\frac{|x_p - y_p|}{x_p} \times 100\%$, where x_p is the exact p^{th} quantile point, y_p is the estimated p^{th} quantile point using the fitted g - and h - distribution.

Sample Size (g, h)	LV			NMLE			QLS			robust QLS			
	g	h	Error (%)	g	h	Error (%)	g	h	Error (%)	g	h	Error (%)	
n=100	(0, 0)	-0.0044(0.0031)	-0.0513(0.0020)	3.13 9.85	-0.0083(0.0029)	0.0119(0.0008)	1.38 0.27	-0.0036(0.0029)	0.0213(0.0012)	0.55 3.17	-0.0062(0.0037)	0.0421(0.0020)	0.85 6.34
	(0, 0)	-0.0044(0.0037)	-0.0604(0.0027)	4.36 12.10	-0.0075(0.0036)	-0.0156(0.0021)	1.19 3.71	-0.0022(0.0038)	0.0044(0.0025)	0.74 1.05	-0.0035(0.0044)	0.0161(0.0034)	0.25 2.14
	(0, 0)	-0.0050(0.0059)	-0.0868(0.0052)	6.76 17.35	-0.0021(0.0058)	-0.0193(0.0038)	1.24 4.05	-0.0039(0.0068)	0.0140(0.0046)	1.17 3.24	-0.0001(0.0069)	0.0072(0.0056)	0.43 1.19
	(0.1, 0)	-0.0092(0.0032)	-0.0498(0.0021)	3.53 10.22	-0.0042(0.0028)	0.0120(0.0008)	1.06 0.24	-0.0066(0.0029)	0.0218(0.0012)	0.42 2.95	-0.0111(0.0038)	0.0443(0.0021)	0.82 6.41
	(0.4, 0)	-0.0239(0.0034)	-0.0418(0.0029)	4.57 10.87	0.0099(0.0030)	0.0106(0.0007)	0.07 1.76	-0.0209(0.0034)	0.0278(0.0015)	0.31 2.86	-0.0220(0.0042)	0.0504(0.0024)	0.98 6.76
	(0.2, 0.2)	-0.0143(0.0044)	-0.0706(0.0037)	5.61 14.80	-0.0085(0.0044)	-0.0181(0.0028)	1.26 4.20	0.0009(0.0048)	0.0034(0.0034)	1.34 1.60	-0.0045(0.0052)	0.0077(0.0044)	0.13 0.77
n=1,000	(0, 0)	-0.0002(0.0010)	-0.0100(0.0007)	0.21 1.58	0.0005(0.0009)	0.0050(0.0002)	0.08 0.60	0.0003(0.0009)	0.0069(0.0003)	0.13 1.07	0.0001(0.0012)	0.0152(0.0007)	0.7 2.79
	(0, 0)	0.0002(0.0012)	-0.0089(0.0010)	0.45 1.67	0.0006(0.0011)	-0.0010(0.0007)	0.02 0.15	0.0001(0.0012)	0.0040(0.0008)	0.26 0.79	0.0005(0.0014)	0.0028(0.0013)	0.15 0.54
	(0, 0)	0.0012(0.0020)	-0.0080(0.002)	0.75 1.82	0.0014(0.0018)	-0.0001(0.0012)	0.11 0.11	0.0012(0.0023)	0.0145(0.0017)	1.28 3.31	0.0016(0.0022)	0.0040(0.0018)	0.33 0.92
	(0.1, 0)	-0.0007(0.0010)	-0.0029(0.0008)	0.43 0.88	0.0020(0.0008)	0.0048(0.0002)	0.02 0.71	-0.0010(0.0009)	0.0071(0.0004)	0.07 0.97	-0.0009(0.0012)	0.0155(0.0007)	0.67 2.76
	(0.4, 0)	-0.0024(0.0011)	-0.0039(0.0010)	0.51 1.17	0.0058(0.0009)	0.0039(0.0002)	0.28 1.10	-0.0053(0.0011)	0.0100(0.0005)	0.10 1.18	-0.0041(0.0014)	0.0173(0.0008)	0.63 2.79
	(0.2, 0.2)	-0.0007(0.0015)	-0.0077(0.0015)	0.57 1.67	0.0007(0.0014)	-0.0006(0.0008)	0.03 0.04	0.0009(0.0016)	0.0071(0.0012)	0.61 1.61	0.0009(0.0017)	0.0037(0.0015)	0.26 0.78

Table 2

of Bias and (Standard Error) for g - and h distribution parameter estimates from methods including LV, NMLE, QLS and rQLS. The observations were generated from standard normal distribution and transformed to g - and h distribution using the selected parameters g and h . Observations were from normal distribution with different mean and standard deviation. The error rate (%) at 95th (upper row) and 99th (lower row) were calculated by $\frac{|x_p - y_p|}{x_p} \times 100\%$, where x_p is the exact p^{th} quantile point, y_p is the estimated p^{th} quantile point using the fitted g - and h

Regular Obs	Outliers	LV				NMLE				QLS				robust QLS			
		g	h	Error(%)	g	h	Error(%)	g	h	Error(%)	g	h	Error(%)	g	h	Error(%)	Error (%)
(0, 0)	N (5, 0.5)	0.2894 (0.0027)	0.1694 (0.0023)	56.55 116.47	0.3148 (0.0027)	0.1398 (0.0020)	65.16 123.45	0.1784 (0.0040)	0.1442 (0.0037)	45.40 86.52	0.0018 (0.0043)	0.0519 (0.0029)	45.40 86.52	0.0018 (0.0043)	0.0519 (0.0029)	2.42 9.78	
(0, 0.1)	N (17.5, 0.5)	0.5377 (0.0037)	0.4376 (0.0028)	133.49 417.55	0.5019 (0.0041)	0.3236 (0.0027)	137.77 347.19	0.2664 (0.0052)	0.2048 (0.0052)	66.52 140.55	-0.0037 (0.0044)	0.0162 (0.0034)	66.52 140.55	-0.0037 (0.0044)	0.0162 (0.0034)	0.25 2.14	
(0, 0.4)	N (74.2, 0.5)	1.2532 (0.0065)	1.2377 (0.0048)	695.28 7519.18	0.8448 (0.0069)	0.7270 (0.0044)	456.37 2034.01	0.5193 (0.0095)	0.4252 (0.0101)	155.99 459.08	-0.0001 (0.0069)	0.0065 (0.0057)	155.99 459.08	-0.0001 (0.0069)	0.0065 (0.0057)	0.41 1.07	
(0.1, 0)	N (6.5, 0.5)	0.3223 (0.0027)	0.1977 (0.0024)	66.97 145.02	0.3584 (0.0026)	0.1505 (0.0020)	77.03 149.67	0.1988 (0.0042)	0.1580 (0.004)	50.95 99.86	-0.0072 (0.0042)	0.0494 (0.0027)	50.95 99.86	-0.0072 (0.0042)	0.0494 (0.0027)	1.49 8.03	
(0.4, 0)	N (16.5, 0.5)	0.4430 (0.0032)	0.3063 (0.0029)	111.12 290.19	0.4999 (0.0027)	0.1702 (0.0021)	123.33 256.41	0.2646 (0.0049)	0.2104 (0.0052)	71.98 155.98	-0.0219 (0.0042)	0.0507 (0.0024)	71.98 155.98	-0.0219 (0.0042)	0.0507 (0.0024)	1.04 6.88	
(0.2, 0.2)	N (105, 0.5)	0.8438 (0.0045)	0.7598 (0.0037)	303.40 1568.59	0.7028 (0.0051)	0.4587 (0.0034)	261.78 825.55	0.3930 (0.0069)	0.3086 (0.0076)	108.51 272.97	-0.0025 (0.0054)	0.0097 (0.0046)	108.51 272.97	-0.0025 (0.0054)	0.0097 (0.0046)	0.49 1.49	
(0, 0)	N (5, 0.5)	0.2897 (0.0010)	0.1668 (0.0007)	62.99 124.87	0.3140 (0.0008)	0.1369 (0.0006)	65.15 122.73	0.1832 (0.0013)	0.1195 (0.0013)	42.82 77.85	-0.0005 (0.0012)	0.0150 (0.0007)	42.82 77.85	-0.0005 (0.0012)	0.0150 (0.0007)	0.58 2.62	
(0, 0.1)	N (17.5, 0.5)	0.5382 (0.0014)	0.4209 (0.0010)	144.59 430.81	0.4988 (0.0013)	0.3187 (0.0008)	137.40 343.25	0.2661 (0.0017)	0.1819 (0.0016)	62.04 127.27	0.0005 (0.0014)	0.0028 (0.0013)	62.04 127.27	0.0005 (0.0014)	0.0028 (0.0013)	0.13 0.52	
(0, 0.4)	N (74.2, 0.5)	1.3125 (0.0031)	1.1776 (0.0021)	762.59 7787.73	0.8348 (0.0021)	0.7254 (0.0013)	453.41 2007.4	0.5078 (0.0029)	0.3820 (0.003)	138.73 389.87	0.0016 (0.0022)	0.0040 (0.0018)	138.73 389.87	0.0016 (0.0022)	0.0040 (0.0018)	0.33 0.92	
(0.1, 0)	N (6.5, 0.5)	0.3263 (0.0010)	0.1894 (0.0007)	74.01 153.28	0.3579 (0.0008)	0.1470 (0.0006)	76.95 148.57	0.2026 (0.0013)	0.1345 (0.0013)	48.07 90.54	-0.0021 (0.0012)	0.0151 (0.0007)	48.07 90.54	-0.0021 (0.0012)	0.0151 (0.0007)	0.44 2.42	
(0.4, 0)	N (16, 0.5)	0.4640 (0.0012)	0.2803 (0.0010)	121.40 299.90	0.5002 (0.0008)	0.1639 (0.0006)	122.66 252.70	0.2643 (0.0015)	0.1887 (0.0017)	67.35 142.28	-0.0046 (0.0014)	0.0172 (0.0008)	67.35 142.28	-0.0046 (0.0014)	0.0172 (0.0008)	0.53 2.65	
(0.2, 0.2)	N (105, 0.5)	0.8801 (0.0021)	0.7171 (0.0015)	329.77 1613.07	0.6984 (0.0016)	0.4526 (0.0010)	260.29 812.42	0.3857 (0.0022)	0.2809 (0.0023)	98.99 242.16	0.0009 (0.0017)	0.0037 (0.0015)	98.99 242.16	0.0009 (0.0017)	0.0037 (0.0015)	0.26 0.78	

Table 3

The comparisons of upper outliers detection based on rQLS estimated g - and h distribution, normal distribution and corresponding base distribution (n = 10,000)

Regular Obs N = 10,000	Outliers (N=500)	rQLS estimation		rQLS estimated g - and h distribution				Normal Distribution				Base distribution			
		g	h	# of regular obs. labeled as outlier	# of contaminated obs. labeled as outliers	Some- outside rate per sample (%)	# of regular obs. labeled as outlier	# of contaminated obs. labeled as outliers	Some- outside rate per sample (%)	# of regular obs. labeled as outlier	# of contaminated obs. labeled as outliers	Some- outside rate per sample (%)	# of regular obs. labeled as outlier	# of contaminated obs. labeled as outliers	Some- outside rate per sample (%)
(0, 0)	NA	0.0007 (0.0004)	0.0043 (0.0002)	0.03	-	3.1	0.05	-	4.9	-	-	-	-	-	-
(0, 0)	N (5, 0.5)	0.0009 (0.0004)	0.0045 (0.0002)	0.00	222.4	0.3	0.00	296.8	0.4	-	-	-	-	-	-
(0, 0.1)	NA	0.0008 (0.0004)	-0.0006 (0.0004)	0.07	-	6.3	16.98	-	100	0.04	-	100	0.04	-	4.4
(0, 0.1)	N (17.5, 0.5)	0.0003 (0.0004)	-0.0006 (0.0004)	0.04	494.1	3.6	9.45	500.0	100	0.02	500.0	100	0.02	500.0	2.2
(0, 0.4)	NA	0.0011 (0.0007)	-0.0009 (0.0006)	0.05	-	5.1	196.46	-	100	0.05	-	100	0.05	-	4.6
(0, 0.4)	N (742, 0.5)	0.0011 (0.0007)	-0.0009 (0.0006)	0.04	500.0	3.6	149.65	500.0	100	0.03	500.0	100	0.03	500.0	3.2
(0.1, 0)	NA	0.0007 (0.0004)	0.0043 (0.0002)	0.03	-	3.3	0.92	-	61.6	0.05	-	61.6	0.05	-	4.9
(0.1, 0)	N (6.5, 0.5)	0.0009 (0.0004)	0.0045 (0.0002)	0.01	273.6	0.8	0.19	498.5	17	0.01	364.0	17	0.01	364.0	0.8
(0.4, 0)	NA	0.0008 (0.0004)	-0.0006 (0.0004)	0.04	-	3.6	28.73	-	100	0.04	-	100	0.04	-	4.4
(0.4, 0)	N (16, 0.5)	0.0003 (0.0004)	-0.0006 (0.0004)	0.01	440.8	1.3	13.61	500.0	100	0.01	500.0	100	0.01	500.0	1.2
(0.2, 0.2)	NA	0.0011 (0.0007)	-0.0009 (0.0006)	0.05	-	5	108.87	-	100	0.05	-	100	0.05	-	4.5
(0.2, 0.2)	N (105, 0.5)	0.0011 (0.0007)	-0.0009 (0.0006)	0.04	500.0	3.6	74.59	500.0	100	0.03	500.0	100	0.03	500.0	2.7

Table 4

Estimates of g - and h - h distribution parameters for ApoAI, Swirl and DJIA data using methods of LV, NMLE, QLS and rQLS. The Bootstrapping estimated bias, standard error, 95% confidence interval of the estimated g and h and P-value of KS test were provided.

	Estimate	Bias	Standard Error	95% CI	Number of outliers	P-value (KS test)	
ApoAI							
g	LV	-0.042	0.001	0.012	(-0.068, -0.018)	8	0.903
	QLS	-0.032	-0.001	0.015	(-0.060, -0.002)	8	0.913
	rQLS	-0.030	-0.001	0.001	(-0.032, -0.030)	8	0.925
	NMLE	-0.066	-	-	-	7	0.154
h	LV	0.007	0.001	0.011	(-0.015, 0.028)	-	-
	QLS	0.000	0.003	0.006	(-0.014, 0.008)	-	-
	rQLS	0.000	0.002	0.001	(0.002, 0.002)	-	-
	NMLE	0.048	-	-	-	-	-
Swirl							
g	LV	-0.073	-0.001	0.013	(-0.097, -0.046)	0	0.029
	QLS	-0.103	0.000	0.016	(-0.134, -0.072)	11	0.560
	rQLS	-0.106	-0.000	0.001	(-0.108, -0.106)	14	0.460
	NMLE	-0.076	-	-	-	0	0.110
h	LV	0.119	-0.005	0.020	(0.086, 0.163)	-	-
	QLS	0.072	-0.002	0.015	(0.045, 0.103)	-	-
	rQLS	0.069	-0.002	0.001	(0.065, 0.069)	-	-
	NMLE	0.145	-	-	-	-	-
DJIA							
g	LV	-0.161	-0.001	0.012	(-0.184, -0.136)	0	0.028
	QLS	-0.156	0.000	0.012	(-0.182, -0.134)	9	0.289
	rQLS	-0.152	0.007	0.001	(-0.146, -0.144)	21	0.069
	NMLE	-0.156	-	-	-	0	0.024
h	LV	0.170	0.000	0.010	(0.151, 0.189)	-	-

	Estimate	Bias	Standard Error	95% CI	Number of outliers	P-value (KS test)
QLS	0.104	0.000	0.011	(0.085, 0.127)	-	-
rQLS	0.089	-0.004	0.000	(0.085, 0.086)	-	-
NMLE	0.178	-	-	-	-	-