False Discovery Versus Familywise Error Rate Approaches to Outlier Detection

Yihuan Xu and Boris Iglewicz

ABSTRACT
Outliers, in general, are observations that deviate sufficiently from a base distribution. This study deals with outlier detection approaches for large samples from continuous univariate distributions. Investigated are the properties of a practical newer outlier detection approach based on use of a false discovery rate (FDR) method in conjunction with a robustly estimated Tukey g-and-h base distribution. Compared are the properties of a boxplot type outlier detection approach that controls the familywise error rate (FWER) with a newer FDR approach. These options are compared in terms of error rates and effects of moving the outliers gradually further from base distribution center while using 5% and 1% FDR or FWERs. Two microarray datasets are used as examples where the assumed null distributions do not fit the data well. In such cases, the proposed estimated Tukey g-and-h null distribution approach leads to superior outlier detection performance. Supplementary materials for this article are available online.

1. Introduction

This article deals with outlier detection approaches for larger univariate samples, which are becoming more common in practice. Outlier detection has a long history with application to multiple applied areas. Barnett and Lewis (1994) provided an excellent and extensive summary of the outlier detection area with a shorter introduction given by Iglewicz and Hoaglin (1993). In general, a base distribution is assumed for the data followed by a search for outliers deviating from this base distribution. Frequently, the normal distribution is used as the base distribution. Barnett and Lewis (1994) discussed almost 50 outlier detection methods from the normal distribution. More recently, a false discovery rate (FDR) approach has become increasingly popular for testing multiple hypotheses. This approach can also be used for outlier detection as an alternative to the commonly used familywise error rate (FWER) methods. In this article, the effectiveness of a popular FWER method is compared to an FDR approach for detecting outliers. Additionally, the quite general Tukey g-and-h distributional family, with estimated parameters, is investigated as a practical alternative to an assumed base distribution for outlier detection.

The Tukey (1977a) basic boxplot is the most popular tool for labeling observations as outliers. The probability properties of the basic boxplot outlier detection procedure were studied by Hoaglin, Iglewicz, and Tukey (1986) and converted to a formal outlier detection procedure, for normally distributed base distributions, by Hoaglin and Iglewicz (1987). That was further generalized for use with a variety of common base distributions by Banerjee and Iglewicz (2007). This relatively simple boxplot procedure, denoted as BP, is used here as the studied FWER approach.

In this article, a simulation comparison is made between the boxplot method (Tukey 1977a), denoted as BP, and a FDR approach (Benjamini and Hochberg 1995), denoted as BH, for detecting potential outliers. An important part of this investigation is to observe the effects on BP and BH of moving the “outliers” gradually further from the bulk of the data. This provides an insight in advantages and disadvantages of using a FWER versus FDR approach for identifying unusual observations.

The standard boxplot outlier detection method labels all observations outside the interval

\[ (Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)) \]

as deserving further investigation, where \( Q_1 \) is the sample lower quartile and \( Q_3 \) the upper quartile, respectively. Here, these outside observations will be labeled as outliers. The popular boxplot labeling rule uses \( k = 1.5 \), though also considers \( k = 3 \) to detect “far out” observations. Banerjee and Iglewicz (2007) obtained the values of \( k \) and its relations to sample size \( n \) for several distributional families and provided a simple solution for \( k \) by controlling the some-outside rate per sample, defined as

\[ \Pr[ \text{at least one } x_i \in \text{region outside the cutoff in (1)} \] under the null distribution] \leq \alpha. \]  

Here, it is assumed that the sample size is sufficiently large so that \( Q_1, M \) (the sample median) and \( Q_3 \) are close to \( F^{-1}(.25), F^{-1}(.5) \) and \( F^{-1}(.75) \), respectively; where \( F \) is the distribution of the data. For symmetric distribution, it is also reasonable to assume that

\[ \Pr[ X < Q_1 - k(Q_3 - Q_1) ] = \Pr[ X > Q_3 + k(Q_3 - Q_1) ]. \]  

ARTICLE HISTORY
Received August 2014 Revised September 2015

KEYWORDS
Boxplot; False discovery rate; Familywise error rate; g-and-h distribution; Microarray; Outlier detection; Robust estimators

CONTACT Yihuan Xu yihuan.xu@lilly.com Eli Lilly and Company, 440 Route 22, Bridgewater, NJ 08807.

Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/SBR

© 2016 American Statistical Association
Then for symmetric base distributions, (2) can be replaced by
\[
\Pr[X_{(n)} > Q_3 + k(Q_3 - Q_1)] \leq \alpha/2, \text{which can be estimated by solving} \quad \Pr[X_{(n)} > F^{-1}(0.75) + k(F^{-1}(0.75) - F^{-1}(0.25))] \leq \alpha/2, \text{where } X_{(n)} \text{is the largest observation. This leads to}
\]
\[
k = \frac{F^{-1}((1 - \alpha/2)^{1/n}) - F^{-1}(-\rightarrow 0.75)}{F^{-1}(0.75) - F^{-1}(-\rightarrow 0.75)}. \tag{4}
\]

Here \(\alpha = 0.05\) is used. Kimber (1990) studied this rule for skewed distribution and replaced (1) by
\[
\left((Q_1 - k(M - Q_1)), (Q_3 + k(Q_3 - M))\right). \tag{5}
\]

When only upper (or lower) outliers are of interest, then the upper (or lower) cutoff is used. Hence, for one-sided upper end outlier detection, the outlier rejection region becomes
\[
(Q_3 + k(Q_3 - M)),
\]
where
\[
k = \frac{F^{-1}((1 - \alpha)^{1/n}) - F^{-1}(0.75)}{F^{-1}(0.75) - F^{-1}(0.5)}.
\]

The BH method was introduced to deal with multiple hypotheses testing, which controls the expected proportion of false positive discoveries. This method can also be used to detect outliers. Usually BH deals with \(n\) null hypotheses with decision rule \(D\) applied to each of these null hypotheses to decide whether to reject or not reject. As a simple illustrative example, giving \(n\) statistics \(z_1, z_2, \ldots, z_n\) to test the null hypotheses \(H_{01}, H_{02}, \ldots, H_{0n}\), the rule \(D\) rejects \(H_{0i}\) if \(|z_i| \geq c\), where \(c\) indicates the rejection boundary. In this case, the rejected observations can be viewed as outliers. More generally, for a random sample from distribution \(F\) with outliers from distribution \(G\), each observation \(z_i\) can be viewed as a test statistic with null hypothesis \(H_{0i}\) that \(z_i\) is from \(F\). Hence, the nonnull \(z_i\) that are rejected by \(D\) can be viewed as outliers.

In typical hypothesis testing, the null distribution is assumed to be known. Such a distribution is further called the theoretical null distribution, which is frequently assumed to be normal, possibly after a transformation. So, if the investigators feel that the null distribution is, say a student \(t\), then \(t\) will be used as the theoretical null distribution. Such theoretical null distributions may not fit the data well as discussed by Efron (2006) and Xu, Iglewicz, and Chervoneva (2014). As an alternative, Efron (2005) suggested use of an empirical normal null estimated distribution. Denote the null density functions as \(f_0(z)\) and \(\pi_0\) the proportion of null cases, where \(\pi_0\) is assumed to be large (\(\pi_0 \geq 0.9\)). The null distribution \(f_0(z)\) is assumed to be normal \((\mu_0, \sigma_0)\). When the theoretical null distribution is another distribution, it needs to be transformed to a normal distribution first. This Efron approach makes the theoretical null distribution more general, as it provides the opportunity to use the data to estimate the unknown parameters \(\mu_0\) and \(\sigma_0\). Such an estimated normal null distribution, with \(\mu_0\) and \(\sigma_0\) estimated from the data, will be called the empirical null distribution.

The approach given by Xu, Iglewicz, and Chervoneva (2014), of using an estimated far richer Tukey (1977b) g-and-h distribution as the base distribution, will be further investigated and compared with use of a theoretical and also empirical null. The four parameter g-and-h distribution is a monotone transformation of standard normal quantiles \(z_p\), for \(h \geq 0\), to g-and-h quantiles \(T_{A,R,g,h}(z_p)\) as
\[
F^{-1}(p) = T_{A,R,g,h}(z_p) = A + B\exp(c(z_p/h)^{1/2}),
\]
\[
B > 0, \ h \geq 0, \ g \neq 0, \tag{6}
\]
\[
F^{-1}(p) = T_{A,R,g,h}(z_p) = A + B\exp(c(z_p/h)^{1/2}),
\]
\[
B > 0, \ h \geq 0, \ g = 0, \tag{7}
\]

with \(A\) controlling location, \(B\) scale, \(g\) skewness, and \(h\) elongation. Consider several symmetric cases of (7) corresponding to \(A = 0, B = 1, g = 0\) and increasing \(h\). For example, \(h = 0\) yields the standard normal distribution; \(h = 0.065\) yields the logistic distribution; \(h = 0.109\) yields the double exponential distribution; and \(h = 0.97\) yields the Cauchy distribution.

Notice, that as \(h\) increases, so does the thickness of the distribution tail areas. For simplicity, the Tukey g-and-h distribution will henceforth be referred to as the g-and-h distribution. By changing the values of \(g\) and \(h\), the g-and-h distribution can approximate a wide range of common distributions including Cauchy distribution (Martinez and Iglewicz 1984), other \(t\) distributions, chi-square distributions, and gamma distributions, among others. Available methods to estimate the g-and-h distribution parameters include: a moment-based method (Martinez and Iglewicz 1984); a letter value (LV) approach (Hoaglin 1985); numerical maximum likelihood method (NMLE) (Rayner and MacGillivray 2002); and two quantile least square methods (Xu, Iglewicz, and Chervoneva 2014). The LV and NMLE methods work well when no outliers or other contaminants exist; however, they are not resistant to a moderate proportion of contaminated observations. The robust quantile least square (rQLS) method for estimating the g-and-h parameters, introduced by Xu, Iglewicz, and Chervoneva (2014), is used here to fit the g-and-h parameters. The rQLS estimator minimizes the sum of squared differences between estimated sample quantiles, based on biweight trimmed sample, and population quantiles. Specifically, let \(x_{(i)}, i = 1, \ldots, n\), be the \(i\)th order statistic of the sample \(x_i, i = 1, \ldots, n\), from the g-and-h distribution. The predicted \(i\)th order statistic is defined as the quantile \(\xi_{p_i}(\theta)\) of g-and-h distribution with parameter vector \(\theta\), where \(p_i = \frac{i - 1/3}{n+1/3}\). Then the weight corresponding to \(x_{(i)}\) is based on the biweight function and is given by
\[
w_i = w(X_{(i)}) = \left\{ \begin{array}{ll}
1 - \left( \frac{X_{(i)} - \xi_{p_i}(\theta)}{c} \right)^2 & \text{for } |X_{(i)} - \xi_{p_i}(\theta)| \leq c, \\
0 & \text{for } |X_{(i)} - \xi_{p_i}(\theta)| > c, \tag{8}
\end{array} \right.
\]
where \(c\) is a constant of the biweight function. Effectively all observations with weight 0 are trimmed from the sample used in estimation. This method reduces the effects of the outliers and estimates g-and-h parameters using the central portion of the data rather than the entire sample.

The g-and-h distribution has been applied in many areas because it fits a great variety of unimodal distributions well and is easy to use in simulations. Among more recent references are Badrinath and Chatterjee (1988, 1991), Dutta and
Babbel (2002, 2005), Kafadar and Phang (2003), Field (2004), He and Raghunathan (2006), Dutta and Perry (2007), Walters and Ruscio (2009), Jimenez and Arunachalam (2011), and Strandberg and Iglewicz (2013).

The purpose of this article is to further study a number of aspects of outlier detection for larger samples including the effect of varied proportion of outliers on outlier detection quality, the performance of the BH method, the effect of outlier distribution location on BP and BH performance, and effect of choice of base distribution on outlier detection. The article is organized as follows. Section 2 summarizes an extensive simulation study. Two relatively large microarray datasets are considered in Section 3. These are analyzed using BP and BH with several choices for base distributions, leading to further insight in determining the number of expressed genes. Conclusions are presented in Section 4.

2. Simulation Study

A simulation study is first performed to compare BH methods based on estimated g-and-h, theoretical and empirical normal null distributions. As the student t distribution appears commonly in microarray literature, this distribution is used here as a base distribution. So, random samples generated from a g-and-h approximation of t distribution with 10 degrees of freedom (t10), given by g-and-h parameters (0, 1.02, 0, 0.052), and its variants are used in this simulation study. The following scenarios were simulated to evaluate how well the BH method performed in identifying true outliers (null hypothesis is correctly rejected) and false outliers (true null hypothesis is rejected by chance). The simulation was repeated 1000 times for each scenario based on using n = 10,000 or 1000 random observations, as specified.

Four sample generating scenarios are considered in Table 1. In SM1, the theoretical null distribution is assumed to be t distributed with 10 degrees of freedom (t10) and random samples \( x_1, x_2, \ldots, x_n \) \((n = 10,000)\) generated from g-and-h distribution with parameters \((0, 1.02, 0, 0.052)\) as a g-and-h approximation to \( t(10) \). In scenario SM2, it is assumed that the theoretical t distribution is not valid and the sample distribution is skewed and more/less dispersed than a t distribution. To simulate such a scenario, random samples \( x_1, x_2, \ldots, x_n \) \((n = 10,000)\) were generated from the g-and-h distribution \((0, 0.7, 0.15, 0.052)\). This distribution has the same elongation as the \( t(10) \) approximation in SM1 but is right skewed and less dispersed. Scenario SM3 consists of the same \( n = 10,000 \) random observations from SM2 plus contamination with 500 (5%) additional observations from \( N(5.82, 0.5) \), which is located at the 99.9968% quantiles of the base distribution. Scenario SM4 again consists of 10,000 random observation from SM2, with 500 additional observations, 250 observations from \( N(-3.19, 0.5) \), and 250 observations from \( N(5.82, 0.5) \), which are located at the quantiles of 0.0032% and 99.9968% quantiles of the base distribution, respectively.

The theoretical null distributions in SM1–SM4 are assumed to be \( t \) distributed with 10 degrees of freedom. The g-and-h parameters are estimated using the rQLS method on the original generated sample. Due to the simplicity of calculations, the theoretical and empirical normal null distributions are estimated from the transformed sample to compare with the g-and-h null distribution. Hereafter, they are referred to as the theoretical null distribution and the empirical null distribution, respectively. The transformation, \( z_i = \Phi^{-1}(F_{t(10)}(x_i)) \), where \( F_{t(10)}(\cdot) \) is the t10 cumulative distribution function and \( \Phi \) the standard normal distribution function, is used to transform \( x_1, x_2, \ldots, x_n \) from assumed t10 distribution to \( z_1, z, \ldots, z_n \) of the standard normal distribution. The FDR and local FDR results were obtained using the R package “locfdr” that allows estimating \( \pi_0 \) for theoretical null distribution \( N(0, 1) \) and \((\mu_0, \pi_0, \pi_1)\) for empirical null distribution from transformed random sample \( z_1, z, \ldots, z_n \).

The following steps are used to identify outliers for g-and-h estimated BH method:

1. For every observation, the z-score based on rQLS estimated g-and-h parameters is calculated by solving for \( z_p \) in Equation (6) or (7).
2. Convert the z-score to corresponding p-value \( P_i, i = 1, \ldots, n \). Find the largest k such that \( P_{(k)} \leq \frac{k}{n} \alpha \), where \( \alpha \) is an a priori selected discovery rate; here \( \alpha = 0.05 \) or \( \alpha = 0.01 \) and \( P_{(k)} \)is the kth ranked p-value.
3. Declare outliers the observations corresponding to \( P_{(i)}, i = 1, \ldots, k \).

This step is implemented using the “p.adjust” function in R with “BH” option.

Table 1 presents the comparison of BH decision rule with discovery rate 1% and 5% (denoted as BH1 and BH5 hereafter, respectively) based on g-and-h null distribution estimated from original sample, plus theoretical and empirical null distributions from transformed sample. This table contains the number of false outliers and the number of true outliers for the BH1 and BH5 methods. The results indicate that when the theoretical null distribution is valid, the g-and-h estimated null distribution approach performs roughly as well as the assumed theoretical null for both the BH1 and BH5 methods. However, when the theoretical null distribution is not valid, that is, the theoretical null distribution is assumed to be t10 when the true null distribution is skewed and less dispersed, the estimated g-and-h null distribution performs better than the theoretical and the empirical null distribution in terms of control of FDR. In this uncontaminated case (SM2), the BH5 detects 0.1 and 24.3 false outliers on average based on the estimated g-and-h null distribution and the empirical null distribution, respectively. That is, the empirical null method detects noticeably more false outliers as compared to the estimated g-and-h null approach (24.3 vs. 0.1). When outliers are on both sides (SM4), the BH5 detects 490.5 true outliers with 29.4 false outliers based on the estimated g-and-h null distribution compared to 250.5 true outliers with 2.3 false outliers based on the theoretical null and 490.5 true outliers and 118.9 false outliers based on the empirical null distribution. The one-sided outlier scenario (SM3) shows a similar pattern as the two-sided outlier case, especially for the empirical null approach. Overall, the estimated g-and-h null approach works well for all Table 1 cases, while both the theoretical and empirical null do not.

Figure S1 in the online supplementary materials illustrates the histogram and the density curve of \( \pi_0 f_0(x) \) for the estimated g-and-h distribution (lower panel), the theoretical (upper left), and the empirical (upper right) null distributions from one of
the simulated examples of SM2, where the true null distribution is skewed and less dispersed. The estimated g-and-h distribution fits the data well on both sides. Since the sample is less dispersed than its theoretical distribution, $N(0, 1)$ does not fit the transformed sample on both sides. The empirical null distribution fits the data better than does $N(0, 1)$ but does not take account of the skewness.

A sensitivity analysis is summarized in Table 2 where BP and BH are compared as the outlier distribution is moved further from the base distribution. Here the g-and-h base distribution is $(0, 1, 0.01)$ or $(g, h) = (0, 0.1)$. This distribution has slightly lighter tails than the double exponential with $(0, 0.109)$. The outlier generating normal distribution, $N(\mu, \sigma)$, has the mean ranging from 8.9 to 81.1, with $\sigma = 1$ or $\sigma = 0.25$. For $n = 10,000$ and $m = 500$ outliers from $N(8.9, 1)$, BP identifies 0 false outliers, but only 3.4 true outliers out of 500. This is clearly unsatisfactory. BH5, on the other hand, identifies all 500 true outliers at a cost of 26.3 false outliers. The performance of BH1 falls in between, with procedure identifying 490.5 true outliers out of 500 at a cost of identifying only 5.3 false outliers. For outliers coming from $N(81.1, 1)$, all three procedures identify the 500 true outliers but at a cost of 0.5, 4.9, and 26.5 false outliers for BP, BH1, and BH5, respectively. In conclusion, BH is superior at identifying true outliers when the outlier generating distribution is a moderate distance from the base distribution at a modest cost of identifying false outliers, while BP does well when the outlier distribution is far away from the base distribution. Similar results are observed for the $\sigma = 0.25$ case and when $n = 1000$ with 50 added outliers. Two similar tables with standard normal and $(g, h) = (0.1, 0)$ base distributions are given in the table.

### Table 1. Comparison of false and true outliers, using BH methods, identified based on g-and-h, theoretical and empirical null distributions ($n = 10,000$) in 1000 simulations.

| Scenario | Regular obs ($n = 10,000$) | Outliers ($n = 500$) | Method | g-and-h null distribution | Theoretical null distribution | Empirical null distribution |
|----------|-----------------------------|----------------------|--------|--------------------------|------------------------------|-----------------------------|
|          |                             |                      |        |                          |                              |                             |
|          |                             |                      | # of false outliers | # of true outliers | # of false outliers | # of true outliers | # of false outliers | # of true outliers |
| SM1      | (0, 1.02, 0)                | —                    | —                  | —                  | 0.0              | 0.0              | 0.0              | 0.0              |
|          | 0.0526                      | —                    | 0.1                | 0.0                | 0.0              | 0.0              | 0.0              | 0.0              |
| SM2      | (0.07, 0.15, 0.0526)        | —                    | 0.0                | 0.0                | 0.0              | 0.0              | 0.0              | 0.0              |
| SM3      | (0.07, 0.15, N(5.82, 0.5))  | BH5                  | 10.6              | 499.7             | 7.4             | 496.2            | 47.5             | 500.0            |
|          | (0.0526)                    | BH5                  | 39.7              | 500.0             | 4.5             | 499.9            | 119.1            | 500.0            |
| SM4      | (0.07, 0.15, N(5.82, 0.5))  | BH5                  | 6.1               | 461.3             | 0.0             | 0.0              | 45.3             | 443.9            |
|          | 0.0526                      | BH5                  | 29.4              | 494.0             | 2.3             | 250.5            | 118.9            | 490.5            |

NOTE: The random samples were generated from g-and-h distribution approximating student’s $t$ distribution with degrees of freedom 10 (SM1); g-and-h distribution with same elongation as SM1 but right skewed and less dispersion (SM2-SM4); additional outliers were generated from normal distributions $N(5.82, 0.5)$ and/or $N(-3.19, 0.5)$ in SM3 and SM4. The average number of false outliers and true outliers were presented for BH1 and BH5 methods.

### Table 2. Sensitivity analysis of BP and BH methods based on g-and-h distribution ($0, 0.1$).

| Scenario | Regular obs ($n = 10,000$) | Outliers ($n = 500$) | Method | $\sigma = 1$ | $\sigma = 0.25$ |
|----------|-----------------------------|----------------------|--------|--------------|-----------------|
|          |                             |                      |        | # of regular obs. identified as outlier | # of contaminated obs. identified as outliers | Some-outside rate per sample (%) | # of regular obs. identified as outlier | # of contaminated obs. identified as outliers | Some-outside rate per sample (%) |
|          |                             |                      |        |              |                  |                             |                             |                             |                             |
|           |                             |                      |        |              |                  |                             |                             |                             |                             |
| Regular obs | (0, 0.1) | N(8.9, $\sigma$) | BP | 0.0 | 3.4 | 3.7 | 0.0 | 3.9 |
|           |           | [3.2e-05]                | BH1 | 5.3 | 490.5 | 97.8 | 6.5 | 99.5 |
|           |           | (0.1)                     | BHS | 26.3 | 500.0 | 100.0 | 30.6 | 500.0 |
|           |           | (N(17.5, $\sigma$) | BP | 0.0 | 493.9 | 3.6 | 0.0 | 495.6 |
|           |           | (2.9e-07)                  | BH5 | 5.4 | 500.0 | 98.0 | 5.6 | 98.1 |
|           |           | (9.9e-10)                  | BH5 | 26.5 | 500.0 | 100.0 | 27.3 | 500.0 |
| Outliers  | (0.01)     | N(36.3, $\sigma$) | BP | 0.0 | 500.0 | 3.6 | 0.0 | 500.0 |
|           |           | [9.9e-10]                   | BH1 | 5.4 | 500.0 | 98.0 | 5.6 | 98.1 |
|           |           | (1.3e-12)                  | BH1 | 26.5 | 500.0 | 100.0 | 27.3 | 500.0 |
|           |           | (81.1, $\sigma$)          | BP | 0.0 | 500.0 | 3.6 | 0.0 | 500.0 |
|           |           | [1.3e-12]                  | BH5 | 5.4 | 500.0 | 98.0 | 5.6 | 98.1 |
|           |           | (0.01)                     | BH5 | 26.5 | 500.0 | 100.0 | 27.3 | 500.0 |

NOTE: Additional 5% outliers were generated from $N(\mu, \sigma)$, where $\mu$s are 8.9, 17.5, 36.3, and 81.1, respectively; $\sigma = 1$ or 0.25, respectively. The number of false outliers, true outliers, and some-outside rate based on rQLS robustly estimated g-and-h null distribution are summarized.
Table 3 also reports the “some-outside rate,” defined as the probability that a sample containing \( n \) random null observations leads to the identification of at least one false outlier. The some-outside rate probability is controlled by the BP method for random samples with no outliers. So, in simulating, say, 1000 samples of \( n = 10,000 \) random observations, BP leads to about 50 of the 1000 samples containing at least one false outlier. As these roughly 50 samples will contain few false outliers, this simulation of 1000 times \( 10,000 = 10,000,000 \) observations will lead to roughly 100 false outliers. In that sense, BP is a very conservative procedure. From Table 2, \( n = 10,000 \), it is observed that the some-outside rate is between 3.5% and 3.9% for BP, while it is 100% for BH5. Thus, BH5 identifies far more false outliers among the 10,000,000 sampled uncontaminated observations.

Table 3 contains results based on different proportion of outliers with samples generated from the g-and-h distribution \((g, h) = (0.1, 0)\), \( n = 10,000 \) regular observations and outliers coming from \( N(\mu, 1) \), with \( \mu = 4.9 \) and 10.1, respectively. The number of outliers are, 0, 50, 100, 250, and 500 representing 0.5%, 1%, 2.5%, and 5%, respectively. The case with \( N(10.1, 1) \) outlier distribution results in the detection of all outliers, making BP the best of the three considered methods for this distant outlier case. The more interesting case deals with outliers generated from \( N(4.9, 1) \). Then, BH5 is a better choice as it identifies the highest proportion of true outliers. That leads to: 26.3 out of 50 identified outliers, or 52.6%; 67.6 out of 100, or 67.6%; 211.6 out of 250, or 84.6%; and 386.2 out of 500, or 77.2%. So the proportion of true outliers detected by BH5 depends on the number of outliers in sample, with typically a higher proportion detected as outliers increasing up to a point, after which the detected proportion of true outliers seems to decrease.

Table 3 contains results based on different proportion of outliers with samples generated from the g-and-h distribution \((g, h) = (0.1, 0)\), \( n = 10,000 \) regular observations and outliers coming from \( N(\mu, 1) \), with \( \mu = 4.9 \) and 10.1, respectively. The number of outliers are, 0, 50, 100, 250, and 500 representing 0.5%, 1%, 2.5%, and 5%, respectively. The case with \( N(10.1, 1) \) outlier distribution results in the detection of all outliers, making BP the best of the three considered methods for this distant outlier case. The more interesting case deals with outliers generated from \( N(4.9, 1) \). Then, BH5 is a better choice as it identifies the highest proportion of true outliers. That leads to: 26.3 out of 50 identified outliers, or 52.6%; 67.6 out of 100, or 67.6%; 211.6 out of 250, or 84.6%; and 386.2 out of 500, or 77.2%. So the proportion of true outliers detected by BH5 depends on the number of outliers in sample, with typically a higher proportion detected as outliers increasing up to a point, after which the detected proportion of true outliers seems to decrease.

Although this study deals with larger sample cases, a limited investigation was performed for smaller sample sizes. Table S3 in the online supplementary materials provides some results for \( n = 100 \). These indicate that the BH5 method, based on estimated g-and-h base distribution, performs reasonably well even for \( n = 100 \), by identifying roughly 4.5 true outliers out of 5, on average, at a cost of less than one false outlier out of 100 regular observations. In general, the rQLS estimates of the g-and-h parameters are less efficient for smaller sample sizes, which are expected to make the set FDRs somewhat less reliable. Similarly, the smaller sample boxplot FWER approach will be affected by the replacement of the sample quartiles by population quartiles, in addition to the less efficient estimates of the g-and-h parameters. This may not be a problem in exploratory investigations. For example, an interim data investigation of 300 human temperature readings may include an outlier value of 139, leading to extra care in data recording. In summary, the discussed FDR outlier g-and-h method works well for larger samples and should also provide reasonable results for more moderate sample-size cases.

### 3. Data Examples

Two datasets are used in this section to illustrate the application of outlier detection based on g-and-h null distribution to relatively large sample size microarray experiments. Here, the “outliers” are genes of interest, whose expression are significantly increased or decreased as compared to other genes. The first set of observations is called the Swirl dataset. The Swirl experiment (Smyth 2005) used zebrafish as a model to identify genes with differential expression in the Swirl mutant zebrafish compared to wild-type zebrafish. The raw data contain 8448 gene expression measurements from four Swirl mutant and four wild-type zebrafish. The raw data were transformed by normalization, then linear model and empirical Bayes analysis were applied to normalized data and the final output is a univariate variable, which theoretically should be \( t \) distributed with seven degrees of freedom. The second dataset is HIV data (Efron 2013), which is also microarray data consisting of \( n = 7680 \) genes comparing four HIV-positive subjects with four healthy controls. The data used here contain the two-sample \( t \) statistics with six degrees of freedom.

For both datasets, the “lofcdr” function is used to estimate the empirical null distribution following the transformation of the \( t \) statistics to \( z \) values by \( z_i = \Phi^{-1}(F_{t(df)}(x_i)) \), where \( df \) is the
degrees of freedom 7 and 6, respectively. Since the rQLS method has the restriction $h > 0$, the original student $t$ statistics are used to estimate the g-and-h null distribution parameters.

Figure 1 shows the histograms of the Swirl data and $\pi_0 f_0 (z)$ for the theoretical null distribution $N(0, 1)$, the empirical null distribution, and the estimated g-and-h null distribution. It is apparent that the actual null distribution is more dilated than $N(0, 1)$ and left skewed. The empirical null fits the tail of data poorly and the estimated $\pi_0$ is greater than 1, which does not make practical sense. On the other hand, the estimated g-and-h distribution fits the data fairly well. From Table 4 with BH5 method, it is observed that use of the theoretical null leads to detection of 156 expressed genes; this number changes to 0 for empirical null and 43 for the estimated g-and-h distribution approach. Note that the empirical null method did not find any differentially expressed genes for this dataset.

Table 4. Parameter estimates of the theoretical $(0, 1, \pi_0)$, the empirical $(\mu_0, \sigma_0, \pi_0)$, and the g-and-h $(A, B, g, h)$ null distributions and corresponding number of discoveries for Swirl and HIV data.

| SeqID | Gene name | Moderated $t$ | Adjusted $p$-value | Rank | Adjusted $p$-value | Rank |
|-------|-----------|---------------|-------------------|------|-------------------|------|
| 8295  | 20-L12    | 14.08         | 0.0027            | 1    | 0.0021            | 5    |
| 7036  | 7-O14     | 13.53         | 0.0027            | 2    | 0.0021            | 6    |
| 515   | 27-E17    | 13.41         | 0.0027            | 3    | 0.0021            | 7    |
| 5075  | 18-G18    | 13.40         | 0.0027            | 4    | 0.0021            | 8    |
| 7907  | 24-H18    | 13.24         | 0.0027            | 5    | 0.0021            | 9    |
| 4032  | 18-N24    | 12.85         | 0.0027            | 6    | 0.0021            | 12   |
| 3721  | BMP2-20   | -21.07        | 0.0027            | 7    | 0.0004            | 1    |
| 4546  | 18-G13    | 12.77         | 0.0027            | 8    | 0.0021            | 14   |
| 683   | 6-E18     | 12.44         | 0.0027            | 9    | 0.0022            | 15   |
| 1609  | BMP2-20   | -20.29        | 0.0027            | 10   | 0.0004            | 2    |
| 1697  | 3-L20     | 12.35         | 0.0027            | 11   | 0.0022            | 16   |
| 7491  | 3-O11     | 12.32         | 0.0027            | 12   | 0.0022            | 17   |
| 3723  | D3x3-20   | -20.01        | 0.0027            | 13   | 0.0004            | 3    |
| 1611  | D3x3-19.64| 0.0027        | 14                |      | 0.0004            | 4    |
| 4380  | 6-G21     | 12.02         | 0.0027            | 15   | 0.0022            | 19   |
| 7602  | 9-L23     | 11.70         | 0.0034            | 16   | 0.0022            | 23   |
| 3790  | 2-N16     | 11.61         | 0.0034            | 17   | 0.0022            | 25   |
| 7542  | 6-D23     | 11.65         | 0.0052            | 18   | 0.0030            | 26   |
| 1146  | 2-I23     | 10.22         | 0.0096            | 19   | 0.0043            | 29   |
| 315   | 18-A9     | 10.19         | 0.0096            | 20   | 0.0043            | 31   |
corresponding numbers of differentially expressed genes for HIV data example were 18, 107, and 13 for theoretical null, empirical null, and estimated g-and-h distribution approaches, respectively. Table 4 also presents the number of outliers found by BP and BH1 methods. These results indicate that the normal distribution assumptions cannot efficiently detect outliers in the Swirl dataset for BP method. Additionally, Table 5 shows the top-ranked 20 genes identified by BH5 based on estimated g-and-h null distribution and their corresponding ranks based on theoretical null distribution. Note that the skewness estimated by g-and-h distribution changed the rank of gene “BMP2” from number 1 based on theoretical null distribution to number 7 based on estimated g-and-h distribution. It also results in changing ranks 2–10 and ranks 3–13 among other reorderings. In summary, the choice of null distribution can have a profound effect on the importance rankings of the expressed genes.

Although the illustrated examples deal with two microarray datasets, there are many other types of situations where the proposed outlier testing methods can be fruitful. As an example, the proposed methods can be applied to larger clinical trials, especially when studying none endpoint measurements, where outliers can indicate incorrect recordings, laboratory readings, or patients with more extreme reactions to treatment.

4. Conclusions

This article deals with relatively simple outlier detection methods designed for larger univariate samples. Such methods have the potential for use in a variety of statistical applications. Investigated are the FWER BP outlier detection method and also the far less common false discovery BH outlier detection approach. In addition, the g-and-h distribution, estimated by the robust rQLS method, is proposed to be used as the base distribution for larger sample datasets. This is shown to be a superior approach as data often do not follow the assumed theoretical null distribution, or even a symmetry assumption. When the data show skewness or heavier or lighter tails than normal, the theoretical null or empirical normal null distribution is no longer valid. We have seen that a robustly estimated g-and-h distribution produces superior results in such situations. The flexibility of the fitted g-and-h distribution makes it an appropriate choice as a practical null distribution. In addition, the robust estimation method, rQLS, reduces the impact of outliers when estimating
the g-and-h null distribution parameters. The studied robustly estimated g-and-h parameter null distribution approach is especially useful when the null distribution is unknown.

The simulation results suggest that the location of outliers has an impact on the performance of the BP and BH methods. In larger samples, more nonoutliers may appear as potential outliers. In such cases, BH will identify more true outliers with controlled number of false outliers. On the other hand, the BP method identifies a reasonable number of true outliers and fewer false outliers. In some cases, where outliers are observed to be far away from the bulk of data, the BP method identifies a similar number of true outliers as the BH false discovery method but with fewer false outliers. Overall, the BH5 method is superior in practice to the FWER BP method, except when the outliers are far away from the true null distribution.

These conclusions are reinforced by the two relatively large sample microarray data examples discussed in this article. The graphs in Figures 1 and 2 show that the theoretical and empirical null distributions do not fit the data well, especially at the tails, while the robustly estimated g-and-h distribution provides a far better fit. Thus, using the estimated g-and-h as a base distribution is expected to yield more accurate identification of relevant genes that may be of special interest for further scientific study.

Although the proposed BH5 approach coupled with an estimated g-and-h base distribution is designed for larger sample sizes, it should perform reasonably well for more moderate values of \( n \), especially when used in exploratory data investigations. Although a limited smaller sample-size study is reported here, the more moderate sample-size issue needs further investigation.

Supplementary Materials

The online supplementary materials contain an additional figure and tables.

References

Badrinath, S. G., and Chatterjee, S. (1988), “On Measuring Skewness and Elongation in Common Stock Return Distributions: The Case of the Market Index,” Journal of Business, 61, 451–472. [144]

——— (1991), “A Data-Analytic Look at Skewness and Elongation in Common-Stock-Return Distributions,” Journal of Business & Economic Statistics, 9, 223–233. [144]

Banerjee, S., and Iglewicz, B. (2007), “A Simple Univariate Outlier Identification Procedure Designed for Large Samples,” Communications in Statistics—Simulation and Computation, 36, 249–263. [143]

Barnett, V., and Lewis, T. (1994), Outliers in Statistical Data (3rd ed.), New York: Wiley. [143]

Benjamini, Y., and Hochberg, Y. (1995), “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” Journal of the Royal Statistical Society, Series B, 57, 289–300. [143]

Dutta, K., and Perry, J. (2007), A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital, Working Paper 06–13, Boston, MA: Federal Reserve Bank of Boston. Available at http://www.bos.frb.org/economic/wp/index.htm. [145]

Dutta, K., and Babbel, D. F. (2002), On Measuring Skewness and Kurtosis in Short Rate Distributions: The Case of the US Dollar London Inter-Bank Offer Rates, Working Paper 02–25, Philadelphia, PA: Wharton School Financial Institutions Center. [145]

——— (2005), “Extracting Probabilistic Information from the Prices of Interest Rate Options: Tests of Distributional Assumptions,” Journal of Business, 78, 841–870. [145]

Efron, B. (2005), “Local False Discovery Rates,” Available at http://statweb.stanford.edu/~ckirby/badpapapers/2005LocalFDR.pdf [144]

——— (2006), “Microarrays, Empirical Bayes, and the Two-Groups Model.” Available at http://statweb.stanford.edu/~ckirby/badpapapers/2006TwoGroupsModel.pdf [144]

——— (2013), Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Institute of Mathematical Statistics Monographs), Cambridge, UK: Cambridge University Press. [147]

Field, C. A. (2004), “Using the gh Distribution to Model Extreme Wind Speeds,” Journal of Statistical Planning and Inference, 122, 15–22. [145]

He, Y. L., and Raghunathan, T. (2006), “Tukey’s gh Distribution for Multiple Imputation,” The American Statistician, 60, 251–256. [145]

Hoaglin, D. C. (1985), “Summarizing Shape Numerically: the g-and-h Distributions,” in: Exploring Data Tables, Trends, and Shapes: Robust and Exploratory Technique, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley, pp. 461–513. [144]

Hoaglin, D. C., and Iglewicz, B. (1987), “Fine-Tuning Some Resistant Rules for Outlier Labeling,” Journal of the American Statistical Association, 82, 1147–1149. [143]

Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986), “Performance of Some Resistance Rules for Outlier Labeling,” Journal of the American Statistical Association, 81, 991–999. [143]

Iglewicz, B., and Hoaglin, D. (1993), How To Detect and Handle Outliers, Milwaukee, WI: ASQC Quality Press. [143]

Jiménez, J. A., and Arunachalam, V. (2011), “Using Tukey’s g and h Family of Distributions to Calculate Value-At-Risk and Conditional Value-At-Risk,” Journal of Risk, 13, 95–116. [145]

Kafadar, K., and Phang, T. (2003), “Transformations, Background Estimation, and Process Effects in the Statistical Analysis of Microarrays,” Computational Statistics and Data Analysis, 44, 313–338. [145]

Kimber, A. C. (1990), “Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions,” Applied Statistics, 39, 21–30. [144]

Martinez, J., and Iglewicz, B. (1984), “Some Properties of the Tukey g and h Family of Distributions,” Communications in Statistics—Theory and Method, 13, 353–369. [144]

Rayner, G. D., and MacGillivray, H. L. (2002), “Numerical Maximum Likelihood Estimation for the g-and-k and Generalized g-and-h Distributions,” Statistics and Computing, 12, 57–75. [144]

Smyth, G. K. (2005), “Limma: Linear Models for Microarray Data,” in Bioinformatics and Computational Biology Solutions Using R and Bioconductor, eds. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, New York: Springer, pp. 397–420. [147]

Strandberg, A. G., and Iglewicz, B. (2013), “A Nonparametric Test for Deviation from Randomness with Applications to Stock Market Index Data,” Communications in Statistics—Simulation and Computation, 42, 686–697. [145]

Tukey, J. W. (1977a), “Exploratory Data Analysis,” Boston, MA: Addison-Wesley. [143]

——— (1977b), “Modern Techniques in Data Analysis,” in Proceedings of the NSF-Sponsored Regional Research Conference at Southeastern Massachusetts University, North Dartmouth, MA. [144]

Walters, G. D., and Ruscio, J. (2009), “To Sum Or Not to Sum: Taxometric Analysis With Ordered Categorical Assessment Items,” Psychological Assessment, 21, 99–111. [145]

Xu, Y., Iglewicz, B., and Chernoveneva, I. (2014), “Robust Estimation of the Parameters of g-and-h Distributions, With Applications to Outlier Detection,” Computational Statistics and Data Analysis, 75, 66–80. [144]