arXiv:1310.7278v2 [stat.AP] 21 Jan 2015

Robust Hypothesis Testing via Lq-Likelihood

Yichen Qin and Carey E. Priebe*

January 23, 2015

Abstract

In this article, we introduce a robust testing procedure — the Lq-likelihood ratio test (LqLR). We derive the asymptotic distribution of our test statistic and demonstrate its robustness properties both analytically and numerically. We further investigate the properties of its influence function and breakdown point. We also propose a method for selecting the tuning parameter q, and demonstrate that, with the $\hat{q}$ selected using our approach, the LqLR attains an excellent efficiency/robustness trade-off compared to the traditional likelihood ratio test (LR) and other robust tests. For the special case of testing the location parameter in the presence of gross error contamination, we show that LqLR dominates the Wilcoxon-Mann-Whitney test and the sign test at different levels of contamination.

Keywords: relative efficiency, robustness, gross error model.

*Yichen Qin is Assistant Professor (E-mail: qinyn@ucmail.uc.edu), Department of Operations, Business Analytics and Information Systems, University of Cincinnati, 529 Carl H. Lindner Hall, 2925 Campus Green Drive, Cincinnati, OH 45221. Carey E. Priebe is Professor (E-mail: cep@jhu.edu), Department of Applied Mathematics and Statistics, Johns Hopkins University, 100 Whitehead Hall, 3400 North Charles Street, Baltimore, MD 21210.

1 INTRODUCTION

The likelihood ratio test (LR) is one of the most frequently used statistical tools in many areas of scientific research. However, only under a collection of strict assumptions does the LR obtain its assumed optimal performance. It is known that its performance degrades significantly in the presence of even a mild violation of model assumptions. In an attempt to overcome this problem, we propose a robust testing procedure — the Lq-likelihood ratio test (LqLR) — using the newly developed concept of Lq-likelihood (Ferrari and Yang (2010)). Under a gross error model, the performance of the LqLR compares favorably to the LR and other tests, such as the Wilcoxon-Mann-Whitney test (Wilcoxon (1945); Mann and Whitney (1947)), the sign test (Arbuthnot (1710)), and Huber's robust version of the likelihood ratio test (Huber (1965)).

Our study of the LqLR focuses on the context of a gross error model $h(x) = (1-\epsilon)f(x;\theta) + \epsilon g(x)$, where f is our "idealized" model with the parameter θ that we are interested in testing, g is the measurement error component (or the contamination component), and ϵ is the contamination ratio. With $\epsilon > 0$, h represents the true data generating process, which is a small deviation from the "idealized" model f. For a data set generated from h, the majority of the data points (i.e., $100(1-\epsilon)\%$ portion) come from f, whereas the rest of the data points (from g) are considered measurement errors or outliers. Common choices for the contamination distribution g are (a) a fat tail distribution and (b) a point mass function.

The measurement error problem has been one of the most practical problems in Statistics. Suppose we have some measurements $X = (X_1, X_2, ..., X_n)$ generated by a scientific experiment. X follows a distribution f_θ with an interpretable parameter θ, our parameter of interest. However, we do not observe X, rather, we observe $X^* = (X_1^*, X_2^*, ..., X_n^*)$, where most of the $X_i^* = X_i$, but there are a few outliers due to human errors or instrument malfunction. In other words, X^* is X contaminated with gross errors. Under such circumstances, using data X^*, we still have θ as the target parameter for our hypothesis testing or estimation (Bickel and Doksum (2007)). To overcome this problem, we introduce the LqLR.

Robust statistics has been well studied for the past 50 years. It addresses the problem of model assumption violation and proposes remedies for this issue. A robust statistical procedure performs nearly optimally when model assumptions are valid and still maintains good performance when the assumptions are violated. A robust procedure should be able to produce a valid conclusion regardless of a few bad or contaminated data points. However, within the subject of robust statistics, there is relatively less research on testing than there is on estimation (Huber and Ronchetti (2009); Hampel et al. (1986)). This is partially because the setting for hypothesis testing is more complex than estimation.

In order to robustify the statistical hypothesis testing procedure, many researchers have proposed methods with desirable properties. Huber (1965) suggested a censored likelihood ratio as $T(\mathbf{x}) = \prod_{i=1}^{n} \max(c', \min(c'', p_1(x_i)/p_0(x_i)))$. The tuning parameters c' and c'' are brought into the equation to address the effect of outliers whose likelihood is exceedingly small and causes the ratio $p_1(x_i)/p_0(x_i)$ to approach zero or infinity. However, hard thresholding using c' and c'' not only causes problems for maximization or minimization, it also induces sensitivity to the thresholds. On the other hand, the LqLR can be considered as a smooth version of the Huberized likelihood ratio test. Meanwhile, Rousseeuw (1984) proposed a "least median of square" approach and corresponding testing procedures. Heritier and Ronchetti (1994) proposed a general class of tests for the bounded influence function. Markatou et al. (1998) proposed a weighted likelihood and Agostinelli and Markatou (2001) further proposed a test based on the weighted likelihood. As for the breakdown point, He et al. (1990) and He (1991) have studied and extended the concept breakdown point for robustness evaluation.

The structure of our article is as follows: We begin with a brief introduction of Lq-likelihood and other preliminaries in Section 2. Then we introduce the LqLR in Section

3 and prove its robustness properties via an analysis of the asymptotic distribution, the influence function, and the breakdown point; we also discuss several related issues such as critical values. Numerical results are presented in Section 4. We discuss the selection of q in Section 5 and demonstrate the superior performance of the LqLR. We provide a discussion and conclusions in Section 6 and relegate the proofs to Section 7.

2 PRELIMINARIES

2.1 Lq-Likelihood and Maximum Lq-likelihood Estimation

A likelihood function measures the likelihood of the observed sample $\mathbf{x} = (x_1, ..., x_n)$ under the hypothesized model. It is defined as $L(\mathbf{x};\theta) = \prod_{i=1}^n f(x_i;\theta)$, where f is the hypothesized model with $\theta \in \Theta \subset \mathbb{R}^d$. Usually it is more convenient to work with the log-likelihood, $l(x;\theta) = \log L(\mathbf{x};\theta) = \sum_{i=1}^n \log f(x_i;\theta)$. Ferrari and Yang (2010) introduced the Lq-likelihood, which is defined as $\sum_{i=1}^n L_q(f(x_i;\theta))$. The $L_q(\cdot)$ function (with a tuning parameter q) is defined as $L_q(u) = (u^{1-q} - 1)/(1-q)$ for $q \neq 1$, and $L_q(u) = \log u$ for $q = 1$. Notice that when $q \to 1$, $L_q(u) \to \log u$. Throughout this article, we assume $0 < q \leq 1$.

To estimate θ, maximum likelihood estimation (MLE) is usually used: $\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^n \log f(x_i;\theta)$. Alternatively, we can use maximum Lq-likelihood estimation (MLqE): $\hat{\theta}_q = \arg\max_{\theta\in\Theta} \sum_{i=1}^n L_q(f(x_i;\theta))$. For ML$q$E, we solve the L$q$-likelihood equation, $0 = \sum[f'_\theta/f_\theta]f_\theta^{1-q}$, which is a weighted version of the likelihood equation with weights being f^{1-q}. When $q < 1$, data points with high likelihoods are assigned large weights. Outliers are usually assigned small weights because of their low likelihoods, which gives MLqE remarkable robustness. As $q \to 1$, the MLqE becomes MLE.

The reason we gain robustness from the Lq-likelihood is that the Lq function is bounded from below for $0 < q < 1$. It is easily seen that $L_q(u) \geq -1/(1-q)$, whereas $\log(x) \to -\infty$

when $x \to 0^+$. In this case, if we have an outlier, say x_1, which gives a very small value of $f(x_1;\theta)$, then $\sum \log f(x_i;\theta)$ approaches $-\infty$, no matter whether θ gives high likelihood for x_2, ... , x_n, i.e., large values of $f(x_2;\theta)$, ... , $f(x_n;\theta)$. On the other hand, since $L_q(u)$ is bounded, it limits the effect of one particular data point on the quantity $\sum L_q(f(x_i;\theta))$. Therefore, the Lq-likelihood surface is much more stable than the log-likelihood surface against a perturbation of a small portion of the data.

2.2 MLqE as the Test Statistic and its Relative Efficiency

As a preview of the advantage of Lq-likelihood, we temporarily use the MLqE of θ, $\hat{\theta}_{q,n}$, as our test statistic, and compare it with $\hat{\theta}_{1,n}$ (i.e., MLE).

Suppose we have the observed data $\mathbf{x} = (x_1, ..., x_n)$ from a pdf $f(x;\theta)$. We first define as follows:

Definition 1. $u_q(\theta) = \mathbb{E}_\theta[\hat{\theta}_{q,n}]$, $\psi_q(x;\theta) = \frac{\partial}{\partial\theta} L_q(f(x;\theta))$, and $\psi_q'(x;\theta) = \frac{\partial^2}{\partial\theta^2} L_q(f(x;\theta))$.

Theorem 1. The asymptotic distribution of $\hat{\theta}_{q,n}$ is $\sqrt{n}(\hat{\theta}_{q,n} - u_q(\theta)) \sim N(0, V_q(\theta))$, where $V_q(\theta) = \mathbb{E}[\psi_q(X;\theta)^2]/\mathbb{E}[\psi_q'(X;\theta)]^2$. When $q = 1$, we have $\mathbb{E}[\psi_1(X;\theta)^2] = -\mathbb{E}[\psi_1'(X;\theta)]$. Hence, $\sqrt{n}(\hat{\theta}_{1,n} - u_1(\theta)) \sim N(0, 1/\mathbb{E}[\psi_1(X;\theta)^2])$.

We want to test the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$ with a size of α. To maintain the size of α, we reject H_0 when $\frac{\hat{\theta}_{q,n} - u_q(\theta_0)}{\sqrt{V_q(\theta_0)/n}} \geq C_{q,n}$. Notice that $C_{q,n} \to z_{1-\alpha}$ when $n \to \infty$, where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.

It is straightforward to prove that $\hat{\theta}_{q,n}$ $(0 < q \leq 1)$ satisfies assumptions 1 through 4 from Lehmann and D'Abrera (2006) pp. 371-372 (restated in Section 7). Therefore, the efficacy of $\hat{\theta}_{q,n}$ $(0 < q \leq 1)$ is $c_q = \frac{u_q'(\theta_0)}{\sqrt{V_q(\theta_0)}}$, where u_q' is the derivative of u_q. Furthermore, the limiting power of $\hat{\theta}_{q,n}$ (for for testing against $H_1 : \theta = \theta_0 + \frac{\delta}{\sqrt{n}}$) is $\Pi_{q,n} \to \Phi(c_q\delta - z_{1-\alpha})$, where $\Phi(\cdot)$ is the cdf of $N(0,1)$.

Now we study how the relative efficiency between $\hat{\theta}_{1,n}$ and $\hat{\theta}_{q,n}$ $(0 < q < 1)$ changes as the level of contamination increases. Suppose data follows a gross error model $h(x;\theta,\epsilon) = (1-\epsilon)f(x;\theta) + \epsilon g(x)$, where $f(x,\theta)$ is the idealized model with the location parameter θ.

In this case, Theorem 1 is still valid with V_q adjusted as $V_q(\theta) = \frac{\mathbb{E}_h[\psi_q(X;\theta)^2]}{\mathbb{E}_h[\psi_q'(X;\theta)]^2}$. The null hypothesis $H_0 : \theta = \theta_0$ is tested against the alternative hypothesis $H_1 : \theta > \theta_0$. The relative efficiency between $\hat{\theta}_{1,n}$ and $\hat{\theta}_{q,n}$ is defined as $e_{q,1} = (c_q/c_1)^2 = \frac{V_1(\theta_0)}{V_q(\theta_0)}\left(\frac{u_q'(\theta_0)}{u_1'(\theta_0)}\right)^2$, which is an increasing function of ϵ.

Theorem 2. Suppose f and g are distributions that are symmetric about θ. The relative efficiency between $\hat{\theta}_{1,n}$ and $\hat{\theta}_{q,n}$ is $e_{q,1} = \frac{V_1(\theta_0)}{V_q(\theta_0)}$. The limiting power of $\hat{\theta}_{q,n}$ becomes $\Pi_{q,n} \to \Phi(\frac{\delta}{V_q(\theta_0)} - z_{1-\alpha})$.

Proof. Under the assumptions, we can prove $u_1(\theta) = u_q(\theta) = \theta$ and $u_1'(\theta) = u_q'(\theta) = 1$. Based on the definitions of relative efficiency, the result follows. □

The theorem implies that the relative efficiency only depends on the asymptotic variance of $\hat{\theta}_{q,n}$ which gives us a direction for selecting potential suitable qs (Section 5).

3 Lq-LIKELIHOOD RATIO TEST

3.1 Lq-likelihood Ratio Test Statistic

We now define a Lq-likelihood ratio test. Suppose we have data $\mathbf{x} = (x_1, ..., x_n)$. The null and alternative hypotheses are $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_0 is the null parameter space and Θ_1 is the alternative parameter space. The traditional likelihood ratio test statistic is $D(\mathbf{x}) = -2\sup_{\theta\in\Theta_0}\{\sum_{i=1}^n \log f(x_i,\theta)\} + 2\sup_{\theta\in\Theta_0\cup\Theta_1}\{\sum_{i=1}^n \log f(x_i,\theta)\}$. Naturally, we define the Lq-likelihood ratio test (LqLR) as

$$D_q(\mathbf{x}) = -2\sup_{\theta\in\Theta_0}\Big\{\sum_{i=1}^n L_q(f(x_i,\theta))\Big\} + 2\sup_{\theta\in\Theta_0\cup\Theta_1}\Big\{\sum_{i=1}^n L_q(f(x_i,\theta))\Big\}$$

$$= -2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_{q,0})) + 2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_q)),$$

where $\hat{\theta}_{q,0}$ and $\hat{\theta}_q$ are MLqE of θ within the parameter spaces Θ_0 and $\Theta_0 \cup \Theta_1$, respectively. We reject the null hypothesis when we have a large $D_q(\mathbf{x})$. Note that when $q = 1$, the LqLR becomes the LR.

3.2 Asymptotic Distribution

In this section, we derive the asymptotic distribution of the LqLR test statistic. We assume a parametric model f_θ with the parameter $\theta \in \Theta \subset \mathbb{R}^p$. Suppose we are interested in testing whether the first r dimensions of θ are 0s. We partition the parameter as $\theta = (\alpha, \beta)$, where α is a $r \times 1$ vector representing the the first r dimensions and β is a $(p - r) \times 1$ vector representing the remaining $p - r$ dimensions. Our null and alternative hypotheses are $H_0 : \alpha = 0$, $H_1 : \alpha \neq 0$. (Note that, if we are interested in testing whether r linearly estimable functions of θ are 0, we can simply reparametrize θ to get the form above.) The test statistic is $D_q(\mathbf{x}) = -2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_{q,0})) + 2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_q))$, where $\hat{\theta}_{q,0} = (0, \hat{\beta}_{q,0})$ is the constrained MLqE of β with α fixed at 0, and $\hat{\theta}_q = (\hat{\alpha}_q, \hat{\beta}_q)$ is the unconstrained MLqE.

First we have the following.

Definition 2. Define $S = \mathbb{E}[\psi_q(X;\theta)]$, $S_n = \sum_{i=1}^{n} \psi_q(x_i;\theta)$, $A = \mathbb{E}[\psi_q(X;\theta)\psi_q(X;\theta)^T]$,

$$B = -\mathbb{E}[\psi_q'(X;\theta)] = \begin{pmatrix} B_{\alpha\alpha} & B_{\alpha\beta} \\ B_{\beta\alpha} & B_{\beta\beta} \end{pmatrix}, \text{ and } B^* = \begin{pmatrix} 0 & 0 \\ 0 & B_{\beta\beta}^{-1} \end{pmatrix},$$

where $\psi_q(x_i, \theta)$, S, and S_n are $p \times 1$ vectors, and $\psi_q'(x_i, \theta)$, A, and B are $p \times p$ symmetric matrices.

Theorem 3. Under the null hypothesis, the asymptotic distribution of $D_q(\mathbf{x})$ is given by

$$2D_q(\mathbf{x})\Big|H_0 \overset{d}{\to} \sum_{j=1}^{r} \lambda_j Z_j^2,$$

where Z_j's are independent random variables $N(0,1)$ and λ_j's are r positive eigenvalues of $A(B^{-1} - B^*)$. When $q = 1$, we have $\lambda_j = 1$ and D_1 follows a regular Chi-square distribution with r degrees of freedom.

Proof. See Section 7 for proof. □

3.3 Robust Properties of Lq-likelihood Ratio Test

We will further discuss the robust properties of LqLR and explain how the asymptotic distribution changes when there is contamination in the data. In particular, we discuss the situation where data are actually generated from a gross error model $h = (1-\epsilon)f + \epsilon g$, where f is the assumed model and g is the contamination component. In this case, the results in Theorem 3 are still valid but the expectation is taken under h (i.e., $A = \mathbb{E}_h[\psi_q(X;\theta)\psi_q(X;\theta)^T]$, $B = -\mathbb{E}_h[\psi_q'(X;\theta)]$). For the purpose of illustration, we divide this section into two parts: univariate parameter and multivariate parameter cases.

3.3.1 Univariate Parameter Case

Under the univariate parameter case, A and B are scalars, and the asymptotic distribution of $D_q(\mathbf{x})$ is

$$D_q(\mathbf{x})\Big|H_0 \overset{d}{\to} \frac{A(\epsilon,q)}{B(\epsilon,q)}\chi_1^2,$$

where χ_1^2 is a Chi-square distribution with 1 degree of freedom. When $\epsilon = 0$ and $q = 1$, we have $D_1(\mathbf{x})\Big|H_0 \overset{d}{\to} \chi_1^2$, which is the case of the LR test using data with no contamination.

When $\epsilon > 0$ and $q = 1$, we have $D_1(\mathbf{x})$ follows a distorted Chi-square distribution with the "distortion" captured by the ratio $\frac{A(\epsilon,q)}{B(\epsilon,q)}$. Looking closely at the ratio $\frac{A(\epsilon,q)}{B(\epsilon,q)}$, we have:

Theorem 4. Assuming that g has a relatively fat tail compared to f in the sense that $\mathbb{E}_g[f''_\theta(X,\theta)/f(X,\theta)] > 0$ and assuming the regularity conditions for f (assumption A1 and A2 in Section 7), it holds that

$$\frac{A(\epsilon, q=1)}{B(\epsilon, q=1)} > 1 \text{ for } \epsilon > 0$$
$$\frac{\partial}{\partial \epsilon}\left[\frac{A(\epsilon, q=1)}{B(\epsilon, q=1)}\right] > 0 \text{ for } \epsilon \geq 0.$$

When f is a normal distribution, the condition $\mathbb{E}_g[f''_\theta/f] > 0$ becomes $\sigma_g^2 > \sigma_f^2$.

Proof. See Section 7 for proof. □

Remarks: For the condition in Theorem 4 ($0 < \mathbb{E}_g[f''_\theta/f] = \int f''_\theta \cdot g/f dx$), please note that when $g = f$, we have $0 = \mathbb{E}_f[f''_\theta/f] = \int f''_\theta dx$. Therefore, this condition means that the ratio g/f inflates the quantity $\int f''_\theta dx$ to be positive. When g has a fat tail distribution compared to f, then g/f is greater than 1 when $|x|$ is large and g/f is less than 1 when $|x|$ is small. Meanwhile, f usually has a bell shape, therefore, f''_θ takes positive values at large $|x|$ and negative values at small $|x|$.

Theorem 4 implies that as ϵ increases away from 0, the discrepancy between $A(\epsilon, q = 1)$ and $B(\epsilon, q = 1)$ increases, and $A(\epsilon, q = 1)/B(\epsilon, q = 1)$ increases above 1. Therefore, the LR test statistic D_1 with contaminated data follows an "inflated" Chi-square distribution under the null hypothesis $(D_1|H_0)$. The same phenomenon occurs for the asymptotic distribution under the alternative hypothesis (i.e., an "inflated" non-central Chi-square distribution) for $D_1|H_1$. Such an inflation is aggravated when more contamination is brought to the data. As the inflation becomes more serious, the null and alternative distributions become flatter; therefore, the overlap between these distributions will become larger (see Figure 3 for more details). This explains the degradation of power when contamination is brought into the data. In order to control the degradation of power, we need to control the inflation of the asymptotic distribution. The following theorem illustrates how we can control the inflation with $0 < q < 1$.

Theorem 5. Under the assumptions as in Theorem 4 and an additional assumption (A3 in Section 7), for $\epsilon > 0$ there exists a $q < 1$, such that

$$\Big|\frac{A(\epsilon,1)}{B(\epsilon,1)} - 1\Big| > \Big|\frac{A(\epsilon,q)}{B(\epsilon,q)} - 1\Big|.$$

Proof. See Section 7 for proof. □

Theorem 5 implies that by setting $q < 1$ we can alleviate the inflation and pull the ratio $A(\epsilon,q)/B(\epsilon,q)$ towards 1. In other words, the effect of $q < 1$ on the ratio $A(\epsilon,q)/B(\epsilon,q)$ can offset the inflation effect of contamination $\epsilon > 0$ on the ratio $A(\epsilon,q)/B(\epsilon,q)$. Therefore, we avoid the increasing overlap between null and alternative distributions and create protection for the power of the test.

In summary, we have proved that the divergence between $A(\epsilon,q)$ and $B(\epsilon,q)$ due to contamination is much more serious for $q = 1$ than $q < 1$. Even though we have $A(\epsilon = 0, q = 1) = B(\epsilon = 0, q = 1)$ at zero contamination, the loss of power at $\epsilon > 0$ due to the divergence between $A(\epsilon > 0, q = 1)$ and $B(\epsilon > 0, q = 1)$ is not avoidable for any likelihood-based statistical tests. On the other hand, by setting $q < 1$ we lose the exact equality at zero contamination, that is, $A(\epsilon = 0, q < 1) \neq B(\epsilon = 0, q < 1)$, but the divergence between A and B is much less, and hence its power is greatly preserved. We want to reiterate that, by setting $q < 1$, we trade the exact equality of $A = B$ at $\epsilon = 0$ for much less divergence between A and B at heavy contamination $\epsilon > 0$. In the following section, we will illustrate our findings through numerical examples.

For illustration, we simulate the ratio $A(\epsilon,q)/B(\epsilon,q)$ as a function of ϵ and q and plot

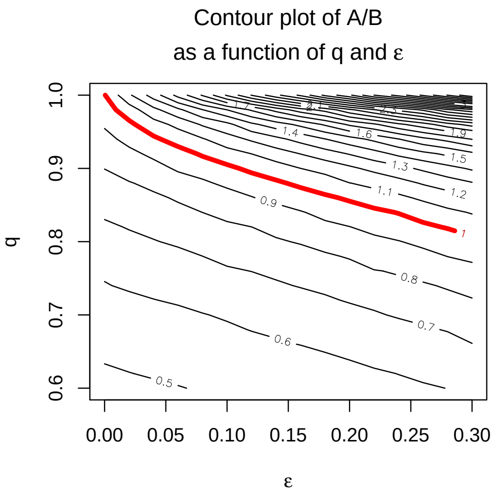

Figure 1: Contour plot of $A(\epsilon, q)/B(\epsilon, q)$ as a function of ϵ and q. In the figure, by setting $q < 1$ we can always decrease the ratio A/B and pull it back toward 1.

contours in Figure 1. The assumed model f is a normal distribution with unknown mean and known variance, whereas the true data generating gross error model is $h(x) = (1 - \epsilon)\varphi(x; 0, 1) + \epsilon\varphi(x; 0, 10)$. We highlight the level of 1 in red (i.e., not inflated). As we can see, when we stand at $q = 1$, the ratio A/B increases as ϵ increases. However, by decreasing q below 1, we can always find a value of q such that the ratio A/B is closer to 1.

3.3.2 Multivariate Parameters Case

Let us move to the multivariate parameters case, where A and B are $p \times p$ matrices. In this case, the distortion is no longer captured by the ratio A/B as in the univariate parameter case, but by all the eigenvalues λ_j of $A(B^{-1} - B^*)$. For simplicity, let us assume $r = p$. That is, we restrict to the simple null hypothesis. Then the distortion is captured by all the eigenvalues of AB^{-1}. We denote all (sorted) eigenvalues of a matrix $M_{r\times r}$ by $\lambda_j(M)$ for $j = 1, ..., r$ with $\lambda_1(M) \geq ... \geq \lambda_r(M)$.

We have the asymptotic distribution of D_q

$$D_q(\mathbf{x})\Big| H_0 \xrightarrow{d} \sum_{j=1}^{r} \lambda_j(A(\epsilon, q)B(\epsilon, q)^{-1})\chi^2_{1,j},$$

where $\chi^2_{1,j}$ are r independent random variables following Chi-square distribution with 1 degree of freedom. Note that when $\epsilon = 0$ and $q = 1$, $A(\epsilon, q) = B(\epsilon, q)$ and $\lambda_j(A(\epsilon, q)B(\epsilon, q)^{-1}) = 1$ for $j = 1...r$, which means D_q follows a Chi-square distribution with r degree of freedom.

For eigenvalues $\lambda_j(A(\epsilon, q)B(\epsilon, q)^{-1})$, we have similar theorems as in the univariate parameter case.

Theorem 6. Under the matrix version of the assumptions in Theorem 4 (assumption A1 and A2 in Section 7), for $\epsilon > 0$ and $q = 1$, the eigenvalues of AB^{-1} are greater than 1, that is, $\lambda_j(A(\epsilon, 1)B(\epsilon, 1)^{-1}) > 1, j = 1, ..., r$.

Proof. See Section 7 for proof. □

The theorem implies that, as the contamination increases, the divergence between A and B increases, and $\lambda_j(AB^{-1})$ increases away from 1 too, causing the inflation of the asymptotic distribution. The original Chi-square distribution with r degrees of freedom (under $\epsilon = 0$, $q = 1$) becomes a sum of r inflated Chi-square distribution with 1 degree of freedom, where each inflation is captured by $\lambda_j(AB^{-1})$ $j = 1, ..., r$. Therefore, the overlap between the null and alternative distributions becomes larger as well. Hence, the power of the test degrades. In order to preserve the test's power, we need to control the inflation of the eigenvalues. We present the following theorem.

Theorem 7. Under the same assumptions in Theorem 6 and an additional assumption A3 (in Section 7), for $\epsilon > 0$ there exists a $q < 1$ such that

$$|\lambda_j(A(\epsilon, q)B(\epsilon, q)^{-1}) - 1| < |\lambda_j(A(\epsilon, 1)B(\epsilon, 1)^{-1}) - 1| \text{ for } j = 1, ..., r.$$

Proof. See Section 7 for proof. □

This theorem means that, by setting $q < 1$, we can effectively shrink these eigenvalues $\lambda_j(AB^{-1})$ back toward 1, and avoid the inflation of distributions as well as the increasing overlap of null and alternative distributions, hence protecting the power of the test. In fact, we can show that $\max_{j=1\ldots r}[\lambda_j(A(\epsilon,q)B(\epsilon,q)^{-1})] \leq \lambda_1(A(\epsilon,q))/\lambda_r(B(\epsilon,q))$. We can further show that $\lambda_1(A(\epsilon,q))/\lambda_r(B(\epsilon,q))$ is bounded when $\epsilon < \epsilon_q^*$ where ϵ_q^* is the breakdown point of LqLR (Section 3.5).

The importance of the Theorems introduced above is twofold: (1) It implies that LqLR makes the approximiation $A \approx B$ more robust against model misspecification. By setting $q < 1$, many statistical inferences based on $A = B$ can be still valid even when the model is misspecified. 2) It provides us a tool for identifying model misspecification. As we can see, setting $q < 1$ can effectively eliminate the influence of outliers. On the other hand, it is not hard to prove that setting $q > 1$ can magnify the effect of outliers. Therefore, the difference between A and B for $q > 1$ will be more sensitive to model misspecification. When $q = 1$, A is essentially Fisher's information matrix. Many model misspecification tests are based on testing $A = B$ (e.g., White (1982)). The results shown above provide a new approach to model misspecification detection.

3.3.3 Simulation Study for Multivariate Parameters Case

We first plot the eigenvalues of $A(\epsilon,q)B(\epsilon,q)^{-1}$ as a function of ϵ for $q = 1$ and $q = 0.97$. The assumed model f is a two dimensional multivariate Gaussian distribution with known variance. The true data generating gross error model is $h = (1-\epsilon)f + \epsilon g$ where g is another multivariate Gaussian with $\mu_g = \mu_f$ and $\Sigma_g = 30\Sigma_f$. In Figure 2, we can see that, when ϵ increases, the eigenvalues increase away from 1 much faster for $q = 1$ than for $q = 0.97$.

Next, we simulate the asymptotic null and alternative distributions under $\epsilon = 0, 0.05, 0.1$ and $q = 1, 0.97, 0.8$. The assumed model f is a three dimensional multivariate Gaussian

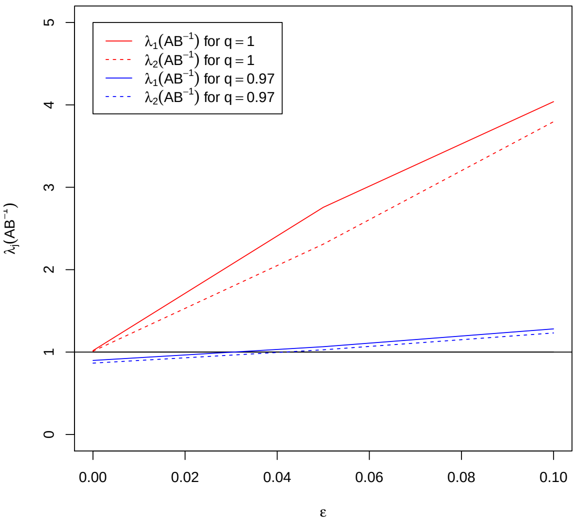

Figure 2: Comparison of eigenvalues of AB^{-1} for $q = 1$ and $q = 0.97$ at difference levels of contamination.

distribution with known variance. The true data generating gross error model is $h = (1 - \epsilon)f + \epsilon g$, where g is the another multivariate Gaussian with $\mu_g = \mu_f$ and $\Sigma_g = 30\Sigma_f$. We simulate the null and alternative distributions of D_q using $\mu_f = [0, 0, 0]^T$ and $\mu_f = [0.15, 0.15, 0.15]^T$.

In Figure 3, in the first row ($q = 1$), we see that as the contamination increases, the null and alternative distributions become flatter and the overlap between these distributions becomes larger, which results in power degradation. In the second row ($q = 0.97$), we see that, instead of having the inflated Chi-square distribution, the null and alternative distributions are less affected by the contamination. This is because the eigenvalues of AB^{-1} are pulled back toward 1 by setting $q < 1$. In the third row ($q = 0.8$), the distributions are much less affected. They hardly change as the contamination increases. However, it is worth noting that, in the lower left figure ($q = 0.8$ and $\epsilon = 0$), the null and alternative distributions overlap more than they do in the upper left figure ($q = 1$ and $\epsilon = 0$), which means that by setting $q < 1$ we lose some of the test's power at zero contamination. Figure 3 illustrates how

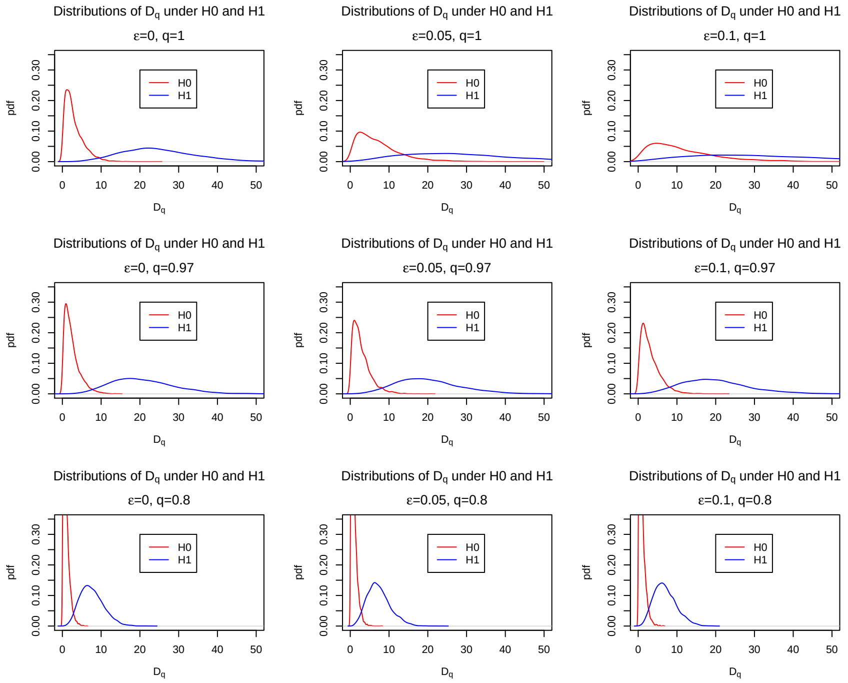

Figure 3: Comparison of asymptotic null and alternative distributions for $\epsilon = 0, 0.05, 0.1$ and $q = 1, 0, 97, 0.8$ for testing the mean of a three dimensional Gaussian distribution.

we gain robustness using the Lq-likelihood with $q < 1$ and make a trade-off for robustness by giving up a little power at zero contamination.

3.4 Bootstrap Estimation of the Critical Value

In the previous section, we discussed the variation of the null distribution of D_q at different levels of contamination. From our research we find that that the null distribution depends on the magnitude of contamination. However, in practice, we hardly ever know the contamination ratio ϵ and other properties of the contamination component g (e.g., variance); therefore, we do not know the exact null distribution or its critical values for different sizes. In order to solve this problem, we need to estimate the critical value from the sample. We propose a bootstrap method for estimating the critical value; it is described as follows. (Suppose we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, where θ is the location parameter.)

Step 1: Given a sample $\mathbf{x} = (x_1, ..., x_n)$, we estimate the mean using a robust procedure, e.g., MLq estimate of the sample mean, $\hat{\theta}_q$.

Step 2: Transform the sample by shifting the entire sample by $\theta_0 - \hat{\theta}_q$ and get $\mathbf{x'} = (x_1 - \hat{\theta}_q + \theta_0, ..., x_n - \hat{\theta}_q + \theta_0)$.

Step 3: Perform a bootstrap using $\mathbf{x'}$ and get bootstrap samples $\mathbf{x'_b}$ for $b = 1, ..., B$.

Step 4: Calculate $D_q(\mathbf{x'_b})$ for each bootstrap sample and denote each as D_q^b.

Step 5: Calculate the $1 - \alpha$ quantile of D_q^b. Denote it as $\widehat{CV}_\alpha$.

$\widehat{CV}_\alpha$ is our final estimate for the critical value. The rationale behind our bootstrap method is that since we are interested in the null distribution under $H_0 : \theta = \theta_0$, we need to shift the observed sample $\mathbf{x}$ so that it has a mean of θ_0. With this new sample $\mathbf{x'}$, we can use the bootstrap to mimic the null distribution. However, since there are usually outliers in the sample, we need to use a robust estimation for the mean. In our case, we adopt the MLqE of the sample. This robust mean helps us to mimic the null distribution.

3.5 Influence Function and Breakdown Point

We now turn to the analysis of the influence function and the breakdown point. The influence function (Hampel et al. (1986)) investigates the infinitesimal behavior of the real valued functional ($\hat{\theta}_q$ in our case). Ronchetti (1979, 1982a,b) have extended the influence function to hypothesis testing by defining the level influence function (LIF) and power influence function (PIF). They show how the asymptotic level and power are influenced by a small amount of contamination at a particular point. The LIF of $\hat{\theta}_q$ is

$$\mathrm{LIF}(x; \hat{\theta}_q, F) = \frac{\phi(\Phi^{-1}(1 - \alpha_0))\mathrm{IF}(x; \hat{\theta}_q, F)}{\sqrt{\int \mathrm{IF}(x; \hat{\theta}_q, F)^2 dF(x)}},$$

where α_0 is the nominal level of the test. We see that $\mathrm{LIF}(x; \hat{\theta}_q, F)$ is proportional to $\mathrm{IF}(x; \hat{\theta}_q, F)$. From the previous section, we know that $\mathrm{IF}(x; \hat{\theta}_q, F)$ is proportional to ψ_q. It is easy to prove that, for most of the models that satisfy the regularity conditions, ψ_q is bounded. Therefore, both IF and LIF are bounded. Similarly, we have bounded PIF. These properties give LqLR the stability of level and power.

As for the breakdown point, we have the following theorem:

Theorem 8. The breakdown point ϵ_q^* for the level and power of LqLR is the same as the breakdown point of the maximum Lq-likelihood estimator, $\hat{\theta}_q$.

Proof. See Section 7 for proof. □

4 NUMERICAL RESULTS AND VALIDATION

4.1 Simulation

Let us assume f is a normal distribution with an unknown mean θ and an unknown variance σ_f^2 (i.e., 2 dimensional parameter space). We want to test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. We simulate data with the sample size $n = 50$ from $h(x; \theta, \epsilon) = (1 - \epsilon)\varphi(x; \theta, 1) + \epsilon\varphi(x; \theta, 50)$, where $\varphi(x; \theta, 1)$ corresponds to f. We apply the LqLR (with $q = 1, 0.9, 0.6$), the Wilcoxon test, the sign test, and Huber's censored likelihood ratio (with $c' = 0.1$, $c'' = 10$) on the data. Note that $q = 1$ is essentially the LR (or equivalently, the t test). At different levels of ϵ, we use $h(x; \theta = 0, \epsilon)$ to generate the data 3000 times and calculate the size and then use $h(x; \theta = 0.34, \epsilon)$ to generate the data and calculate the power. The results are shown in Figure 4.

In Figure 4, let us first note that the sizes of all tests are successfully controlled at 0.05. At $\epsilon = 0$, the LqLR with $q = 1$ (LR) has the highest power; as we decrease q to 0.9 and 0.6, the power decreases. The Wilcoxon test and Huber's test also have high powers. The sign test has the lowest power. As contamination becomes more serious, i.e., ϵ increases away from 0, the LqLR with $q = 1$ (LR) degrades much faster than any of the other tests.

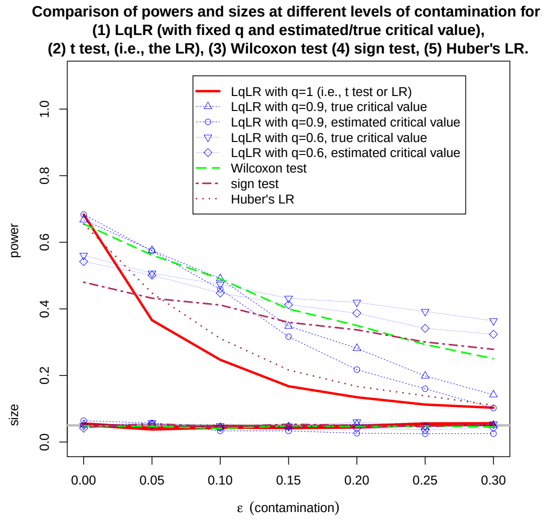

Figure 4: Comparison of powers and sizes for the LqLR for $q = 1$, $q = 0.9$, and $q = 0.6$, the Wilcoxon test, the sign test and Huber's censored likelihood ratio test at different levels of contamination. For the purpose of illustration, we show the power and size of LqLR for both estimated critical values and true critical values (Since we know the true data generating process h, we can obtain the true critical values). However, in practice, it is impossible to know the true critical value.

Huber's test degrades relatively slowly compared to LqLR for $q = 1$. The LqLRs for $q = 0.9$ and $q = 0.6$ degrade at much slower rates. The Wilcoxon test also degrades slowly. Among all tests, the LqLR with $q = 0.6$ and the sign test have the slowest degradation rates (i.e., flattest curves). By adjusting the tuning parameter q to 0.9, we can beat the Wilcoxon test at mild contamination ($\epsilon < 0.05$). If we change q to 0.6, we can beat the Wilcoxon test at heavy contamination ($\epsilon > 0.15$). Meanwhile, the LqLR with $q = 0.6$ uniformly dominates the sign test at all levels of contamination. Last but not least, the figure also shows that our estimated critical values work well. We only slightly overestimate the critical values; therefore, the powers obtained from the estimated critical values are slightly below the powers obtained from the true critical values.

We see that remarkable robustness can be obtained by using the LqLR. The figure also implies that, with an appropriately selected q, it is possible that the performance of LqLR can be generally obtained (see Section 5 for details).

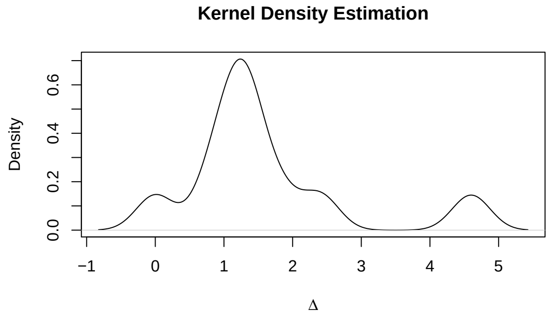

Figure 5: Kernel density estimation of the difference in sleep hours gained for the two drugs.

4.2 Real Data

We use a real data example to demonstrate the effectiveness of our proposed method. The data was first presented in Cushny and Peebles (1905) and later used in Staudte and Sheather (1990). Cushny and Peebles (1905) conducted the experiment to illustrate the effects of optimal isomers of hyoscyamine hydrobromide in producing sleep. There were 10 patients in total. Each patient was given two types of drugs in a randomized order and was asked to record their average sleeping hours gained for the two drugs. Furthermore, the differences in sleeping hours gained for the two drugs, Δ, are calculated: 1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4. We want to test the null hypothesis that two drugs have the same effect, i.e., $H_0 : u_\Delta = 0$, $H_1 : u_\Delta > 0$, where $u_\Delta = \mathbb{E}[\Delta]$.

The importance of this data set is that many statisticians have examined it assuming the normal distribution (including William S. Gosset with his Student's t test). However, the value $\Delta_9 = 4.6$ raises some questions about the normality assumption. A kernel density estimation of Δ (with bandwidth of 0.2755) is presented in Figure 5, where we see $\Delta_9 = 4.6$ clearly casts doubt on the normality assumption. For a level 0.05 test, we can reject H_0 using the t test (equivalent to the LR). If we were to replace the value of 4.6 with 16, then the t test would no longer reject H_0 at the level of 0.05. One may argue that $\Delta_9 = 16$ is an obvious outlier; however, it is counterintuitive that more extreme evidence is favorable to

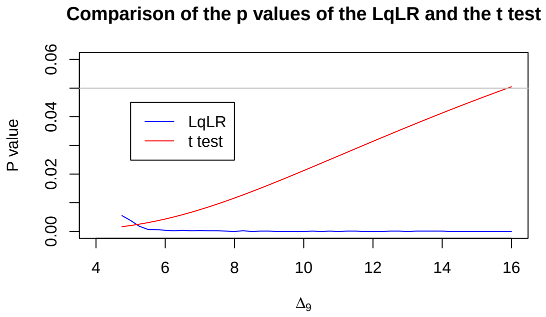

Figure 6: Comparison of p-values of the LqLR and the t test as functions of Δ_9.

the null hypothesis, that is, no difference in two drugs.

Meanwhile, we apply the LqLR on the data set with $q = 0.85$. In Figure 6, we plot the p-value as a function of Δ_9 (which goes from 4.6 to 16) for both the LqLR and the t test. As we can see from the figure, the p-value of the t test gradually increases above 5% as Δ_9 increases. On the other hand, the p-value of the LqLR is well controlled and decreases to 0 as Δ_9 increases. Therefore, the LqLR successfully rejects the null hypothesis with the p-value being consistent with the evidence. Even though the t test (LR) is a special case of the LqLR, by setting $q < 1$, we preserve the efficiency and attain remarkable robustness.

5 SELECTION OF q

So far we have assumed the tuning parameter q to be known, but we never know the optimal q in practice. In this section, we propose a method for adaptively selecting q. The optimal q we propose is defined as $q_{\text{opt}} = \arg\max_q \Pi$, where Π is the limiting power of the test (i.e., asymptotic power). When testing for the location parameter in the symmetric distribution, we have $\Pi = \Phi(\frac{\delta}{V_q(\theta_0)} - z_{1-\alpha})$. Since this is a monotonic function in $V_q(\theta_0)$, our optimal q is given by $q_{\text{opt}} = \arg\min_q V_q(\theta_0)$.

In Figure 7, we plot the relationship between $V_q(\theta_0)$ and q at different levels of contamination using the same setup as in the previous section. We can clearly see that the optimal

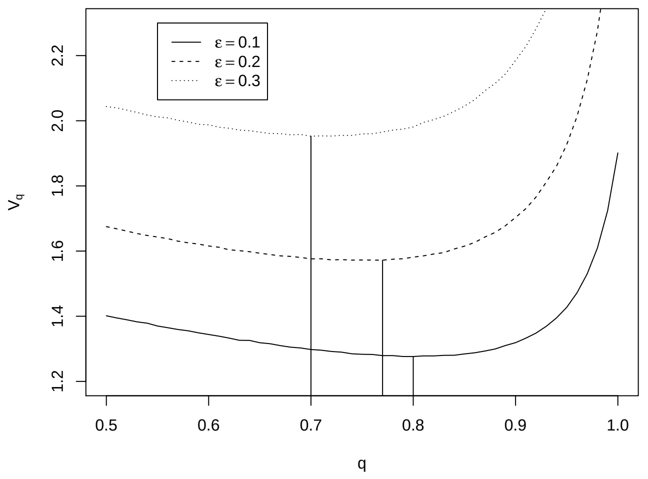

Figure 7: $V_q(\theta_0)$ as a function of q at different levels of contamination ϵ.

q is between 0.6 to 0.9 for these contamination levels. As expected, the more serious the contamination, the lower the optimal q.

In practice, we do not have $V_q(\theta_0)$. We can replace it with the empirical version of this quantity. The data-adaptive estimation for the tuning parameter is given by

$$\hat{q} = \arg\min_q \frac{\frac{1}{n}\sum_{i=1}^n \psi_q(x_i;\hat{\theta}_q)^2}{[\frac{1}{n}\sum_{i=1}^n \psi_q'(x_i;\hat{\theta}_q)]^2}.$$

We now provide a simulation study of the LqLR using estimated q and estimated critical value. We adopt the same setup from the previous section (Section 4.1). By setting θ to 0 and 0.34 and using 2000 Monte Carlo iterations, we compare the sizes and powers of the LqLR, the LR, the Wilcoxon test, the sign test and the Huber's robust version of likelihood ratio test at different levels of contamination. The results are presented in Figure 8. We can clearly see the advantage of the LqLR (with estimated q and estimated critical value) over other tests. Not only does the LqLR degrade very slowly, it also holds the highest power among all other tests. Note that the sizes have been successfully controlled at 5%. In Figure 8, at zero contamination (i.e., $\epsilon = 0$), the LR has the highest power. The LqLR has almost

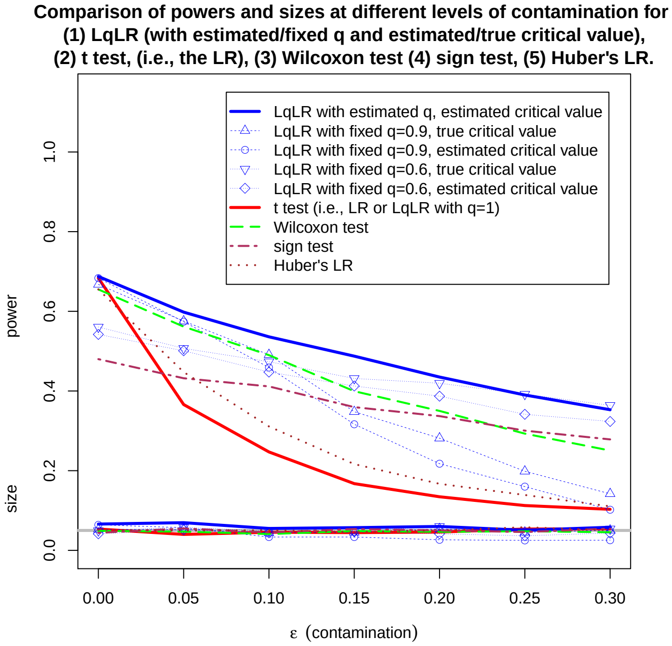

Figure 8: Comparison of the powers and sizes of (1) the LqLR with the estimated q and the estimated critical value; (2) the t test, i.e., the LR; (3) the Wilcoxon test; (4) the sign test; and (5) Huber's robust likelihood ratio test at different levels of contamination.

the same power (only slightly less than the LR). The Wilcoxon and the sign tests have the third and the fourth highest powers, but not comparable to the two likelihood ratio tests. As the contamination becomes more serious (i.e., ϵ increases away 0), the log-likelihood degrades the fastest. Its power quickly drops below all other tests. The Wilcoxon test and the sign test both show good robustness and their powers degrade at much slower rates. However, the LqLR shows a remarkable robustness. It degrades slower than the Wilcoxon test (i.e., the blue curve is flatter than the green curve) and only slightly faster than the sign test (i.e., the blue curve is steeper than the maroon curve). Since the power of the LqLR at $\epsilon = 0$ is above that of the Wilcoxon test and the sign test, the power of the LqLR dominates both the Wilcoxon test and the sign test at all levels of contamination. This implies that, not only can Lq-likelihood preserve efficiency almost perfectly at $\epsilon = 0$, it also obtains robustness comparable to these nonparametric tests, which are known to be very robust. We conclude

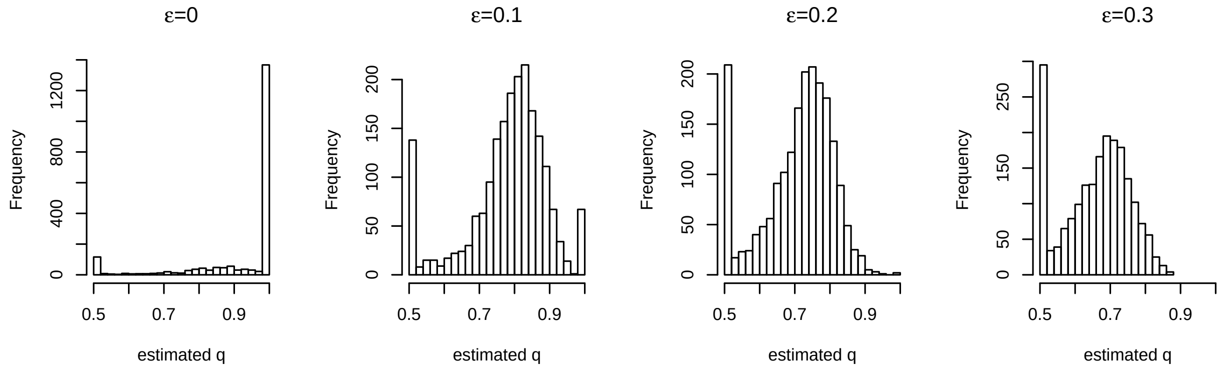

Figure 9: Histogram of the estimated q at different levels of contamination.

that, by losing a little efficiency at $\epsilon = 0$, we have traded for great robustness at $\epsilon > 0$. Our LqLR can be considered as a combination of the LR (at $\epsilon = 0$) and the nonparametric tests (at $\epsilon > 0$). The reason our test beats nonparametric tests uniformly is that we can control the amount of information used by selecting q, whereas the Wilcoxon test always uses the rank information and the sign test always uses the information about whether each data point is below or above the hypothesized mean.

Meanwhile, we also plot the histograms of the estimated q at different levels of contamination in Figure 9. We see that as we get more serious contamination, the estimated q tends to be smaller. In our experiment, we limit the smallest q to be 0.5, which is very similar to the case of testing based on minimum Hellinger distance (Beran (1977)). Whenever our estimated q drops below 0.5, we use 0.5 instead. The reason for this censoring is that we have not understood the case of $q < 0.5$ very well, which is an interesting topic for future research.

Lastly, we test the LqLR against a point mass contamination. We adopt the same setup from the previous simulation and only change the contamination component to the point mass distribution. That is, the true data generating process is $h(x; \theta, \epsilon) = (1-\epsilon)\varphi(x; \theta, 1) + \epsilon\varphi(x; -5, 0.0001)$, where $\varphi(x; -5, 0.0001)$ can be considered as an approximation to the point mass density. We repeat the same experiment above and present the results in Figure 10.

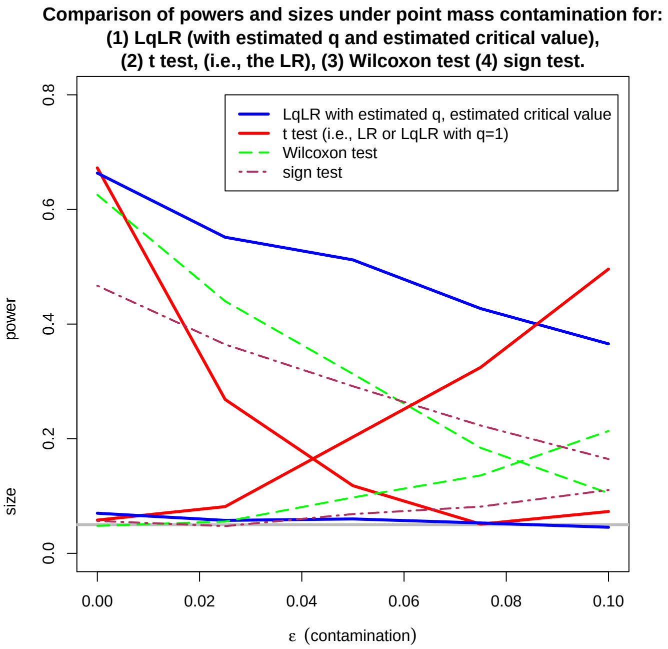

Figure 10: Comparison of the powers and sizes of (1) the LqLR with the estimated q and the estimated critical value; (2) the t test, i.e., the LR; (3) the Wilcoxon test; and (4) the sign test at different levels of **asymmetric** contamination.

In Figure 10, we first notice that as the contamination increases, the sizes of the LR test (t test), Wilcoxon test, and sign test all increase above 5%. This phenomena does not happen when we have symmetric contamination in Figure 8. This is because these tests are more vulnerable against the asymmetric contamination. Among these tests, the LR test is the least robust test and the sign test is more robust than Wilcoxon test. When ϵ increases, the center of the data is dragged to the left (to the contamination at -5), which makes the null hypothesis ($H_0 : \theta = 0$) easier to reject. However, the LqLR is very successful in protecting size and perfectly controls size at different contamination levels. As for the power, we can see a similar phenomenon as in Figure 8. The LqLR is the best among all the tests because it degrades the most slowly and remains the highest. From this figure, we can see the obvious advantage of the LqLR over all other tests.

6 CONCLUSION

In this article, we have proposed a robust testing procedure — the Lq-likelihood ratio test (LqLR) — and demonstrated its advantage over the LR, the Wilcoxon test, the sign test, and Huber's censored likelihood ratio test under the gross error model for testing the location parameter of a symmetric distribution. We prove the LqLR's robustness advantages by deriving the asymptotic distribution, the influence function, and the breakdown point. We further accompany our analytical study with numerical comparisons.

Our LqLR can be considered as a bridge connecting the LR and the nonparametric tests such as the Wilcoxon test and the sign test. By changing the tuning parameter q, we can control the information used in the hypothesis testing. The LR uses the full information of all data points and gives all data points equal weights. The Wilcoxon test takes only the rank information, and therefore becomes extremely robust at the cost of wasting much information. Our LqLR gives each data point a weight as a function of its likelihood and q. Therefore, the data points that are consistent with the "idealized" model are given higher weights whereas data points that are inconsistent with the "idealized" model are partially ignored.

To the extent that the robustness of the Wilcoxon test (minimum asymptotic relative efficiency (ARE) of the Wilcoxon test vs. the t test is 0.864) suggests that the Wilcoxon test should be the default test of choice (so rather than "use Wilcoxon if there is evidence of non-normality," the default position should be "use Wilcoxon unless there is good reason to believe the normality assumption"), these new results in this article suggest that the LqLR test has the potential to become the new default go-to test for practitioners.

Even though our test shows remarkable robustness over other tests, there are still many directions for future research. For example, the investigation of the LqLR's properties under the asymmetric distribution is an important topic. Meanwhile, better estimation procedures for the critical value and q are possible. We have shown that our estimate of the critical value performs decently, but there is clearly a gap between the power obtained from the true critical values and the power obtained from the estimated critical values. Filling in that gap is a challenging task for the future. The estimation of q also leaves many directions for future research; we could develop a more robust procedure for selecting q. Finally, the divergence of A and B mentioned in Section 3.3.2 also indicates a new approach to model misspecification detection.

7 APPENDIX

7.1 Assumptions from Lehmann and D'Abrera (2006) pp. 371-372:

1) The function u_q is differentiable at θ_0 with the derivative $u_q'(\theta_0) \neq 0$;

2) The standard deviation of $\hat{\theta}_{q,n}$ is of order $1/\sqrt{n}$;

3) For a sequence of alternative $\theta_n \to \theta_0$, the distribution of $[\hat{\theta}_{q,n} - u_q(\theta_n)]/\sqrt{V_q(\theta_n)/n}$ tends to the standard normal distribution, where $\theta_n \to \theta_0$ as $n \to \infty$;

4) $V_q(\theta_n)/V_q(\theta_0) \to 1$.

7.2 Assumptions in the Proof

- **A1**: f satisfies the regularity conditions of maximum likelihood estimation.
- **A2**: g has a relatively fat tail compared to f in the sense that $\mathbb{E}_g[\frac{f_\theta''(X,\theta)}{f(X,\theta)}]$ has all positive eigenvalues.

  Remarks for A2: When f is a normal distribution, the assumption A2 becomes $\Sigma_g \Sigma_f^{-1} - I$ has all positive eigenvalues where Σ_g and Σ_f are their covariance matrices. Under the univariate parameter case, the assumption A2 becomes $\mathbb{E}_g[\frac{f_\theta''(X,\theta)}{f(X,\theta)}] > 0$.

When f is a one dimensional normal distribution, the assumption A2 further becomes $\sigma_g^2 > \sigma_f^2$.

- **A3**: For $q < 1$, $|\frac{\partial}{\partial\theta}(\frac{f'}{f} f^{1-q})|$ is a bounded function in x.

7.3 Proof of Theorem 3

The original hypotheses can be further generalized to the following hypotheses: $H_0 : \theta = g(\gamma), H_1 : \theta \neq g(\gamma)$, where $g(\gamma) = (g_1(\gamma), ...g_p(\gamma))$ is a continuously differentiable function from $\mathbb{R}^{p-r}$ to $\mathbb{R}^p$ with a full rank $g'_\gamma(\gamma) = \partial g(\gamma)/\partial\gamma$. $\gamma = (\gamma_1, ..., \gamma_{p-r})$ is a $(p-r) \times 1$ parameter and $g'_\gamma(\gamma)$ is a $p \times (p-r)$ matrix. In other words, we are testing whether the parameter θ lies in the $p-r$ dimensional subspace and can be expressed by $\theta = g(\gamma)$.

When we set $g_1(\gamma) = ... = g_r(\gamma) = 0$, $g_{r+1}(\gamma) = \gamma_1$, ... $g_p(\gamma) = \gamma_{p-r}$ and $\gamma = \beta$, we have recovered the original null hypothesis $H_0 : \alpha = 0$ and $H_1 : \alpha \neq 0$.

Using Taylor expansion, we have

$$2[\ell_q(\hat{\theta}_q) - \ell_q(\theta_0)] = (\hat{\theta}_q - \theta_0)^T nB(\theta_0)(\hat{\theta}_q - \theta_0) + o_p(1),$$

where $\ell_q(\theta) = \sum_{i=1}^n L_q(f(x_i, \theta))$. Since we know

$$\sqrt{n}B(\theta_0)(\hat{\theta}_q - \theta_0) = \frac{1}{\sqrt{n}} S_n(\theta_0) + o_p(1),$$

Substitute in, we have

$$2[\ell_q(\hat{\theta}_q) - \ell_q(\theta_0)] = \frac{1}{n} S_n(\theta_0)^T B(\theta_0)^{-1} S_n(\theta_0) + o_p(1),$$

We repeat the above procedure with the constrain of $\alpha = 0$ for H_0. Note that $\theta_0 = g(\gamma_0)$ in this case. We denote our estimate $\hat{\gamma}_q$. Similarly, we have

$$2[\ell_q(g(\hat{\gamma}_q)) - \ell_q(g(\gamma_0))] = (\hat{\gamma}_q - \gamma_0)^T n g'_\gamma(\gamma_0)^T B(g(\gamma_0)) g'_\gamma(\gamma_0)(\hat{\gamma}_q - \gamma_0) + o_p(1),$$

and

$$-\sqrt{n} g'_\gamma(\gamma_0)^T B(g(\gamma_0)) g'_\gamma(\gamma_0)(\hat{\gamma}_q - \gamma_0) = \frac{1}{\sqrt{n}} g'_\gamma(\gamma_0)^T S_n(g(\gamma_0)) + o_p(1),$$

Substitute in, we have

$$2[\ell_q(g(\hat{\gamma}_q)) - \ell_q(\theta_0)] = \frac{1}{n} S_n(\theta_0)^T g'_\gamma(\gamma_0)[g'_\gamma(\gamma_0)^T B(\theta_0) g'_\gamma(\gamma_0)]^{-1} g'_\gamma(\gamma_0)^T S_n(\theta_0) + o_p(1),$$

Combining the above results, we obtain

$$\begin{aligned} 2D_q(\mathbf{x}) &= 2[\ell_q(\hat{\theta}_q) - \ell_q(g(\hat{\gamma}_q))] \\ &= \frac{1}{n} S_n(\theta_0)^T \Big[B(\theta_0)^{-1} - g'_\gamma(\gamma_0)[g'_\gamma(\gamma_0)^T B(\theta_0) g'_\gamma(\gamma_0)]^{-1} g'_\gamma(\gamma_0)^T \Big] S_n(\theta_0) + o_p(1) \\ &= \frac{1}{n} S_n(\theta_0)^T \Big[B(\theta_0)^{-1} - B^*(\theta_0) \Big] S_n(\theta_0) + o_p(1), \end{aligned}$$

The last step is because we know $g'_\gamma(\gamma) = [\mathbf{0}, I_{(p-r)\times(p-r)}]^T$ where $\mathbf{0}$ is a $(p-r) \times r$ matrix with all 0s and $I_{(p-r)\times(p-r)}$ is a $(p-r) \times (p-r)$ identity matrix. Furthermore, we know $\frac{1}{\sqrt{n}} S_n(\theta_0) \xrightarrow{d} N(0, A(\theta_0))$. It follows that under H_0

$$2D_q(\mathbf{x}) \Big| H_0 \xrightarrow{d} \sum_{j=1}^{r} \lambda_j Z_j^2,$$

where λ_j's are the r positive eigenvalues of $A^{1/2}[B^{-1} - B^*]A^{1/2}$ (or $A[B^{-1} - B^*]$).

7.4 Proof of Theorem 4

Theorem 4: Under assumption A1 and A2, it holds that

$$\frac{A(\epsilon, q=1)}{B(\epsilon, q=1)} > 1 \text{ for } \epsilon > 0$$
$$\frac{\partial}{\partial \epsilon}\left[\frac{A(\epsilon, q=1)}{B(\epsilon, q=1)}\right] > 0 \text{ for } \epsilon \geq 0.$$

Proof. First, we know that $B(\epsilon, q=1) = -\mathbb{E}_h\left[\frac{f''_\theta}{f}\right] + A(\epsilon, q=1) = -\epsilon\mathbb{E}_g\left[\frac{f''_\theta}{f}\right] + A(\epsilon, q=1)$, where we use the fact that $\mathbb{E}_f[f''_\theta/f] = 0$ in the last step. Since $\mathbb{E}_g[f''_\theta/f] > 0$, we have $A(\epsilon, q=1) > B(\epsilon, q=1) > 0$, and hence, $\frac{A(\epsilon,q=1)}{B(\epsilon,q=1)} > 1$. When f is a normal distribution, $\mathbb{E}_g[f''_\theta/f] > 0$ becomes $0 < \mathbb{E}_g\left[\frac{f''_\theta}{f}\right] = \mathbb{E}_g\left[\frac{(x-\theta)^2}{\sigma^4} - \frac{1}{\sigma^2}\right] = \frac{\sigma_g^2}{\sigma_f^4} - \frac{1}{\sigma_f^2}$, which is equivalent to $\sigma_f^2 < \sigma_g^2$. Note that this condition does not require g to be a normal distribution.

Since $h = (1-\epsilon)f + \epsilon g$, we have $\frac{\partial}{\partial \epsilon}\Big(A(\epsilon, q=1) - B(\epsilon, q=1)\Big) = \mathbb{E}_g[\psi_q(X;\theta)^2] + \mathbb{E}_g[\psi'_q(X;\theta)] - \Big(\mathbb{E}_f[\psi_q(X;\theta)^2] + \mathbb{E}_f[\psi'_q(X;\theta)]\Big)$, which is a function that does not involve ϵ, that is, a constant function in ϵ. Furthermore, we know $A(\epsilon=0, q=1) - B(\epsilon=0, q=1) = 0$ (from Definition 2), and $A(\epsilon, q=1) - B(\epsilon, q=1) > 0$ for $\epsilon > 0$. Therefore, we conclude that $\frac{\partial}{\partial \epsilon}(A-B)|_{q=1} > 0$, further, we have $\frac{\partial}{\partial \epsilon}\left[\frac{A}{B}\right]\Big|_{\epsilon=0,q=1} > 0$. □

7.5 Proof of Theorem 5

This theorem is a special case of Theorem 7. Please see Section 7.7 for the proof of the multivariate parameter case.

7.6 Proof of Theorem 6

Theorem 6: Under assumptions A1 and A2, for $\epsilon > 0$ and $q = 1$, the eigenvalues of AB^{-1} are greater than 1, that is, $\lambda_j(A(\epsilon, 1)B(\epsilon, 1)^{-1}) > 1$, $j = 1, ..., r$.

Proof. Based on the definitions of A and B, we have

$$\begin{aligned} AB^{-1} &= \mathbb{E}_h[\frac{f'f'^T}{f^2}(f^{1-q})^2]\Big[\mathbb{E}_h[(\frac{\partial}{\partial\theta}(-\frac{f'}{f}f^{1-q})]\Big]^{-1} \\ &= \mathbb{E}_h[\frac{f'f'^T}{f^2}(f^{1-q})^2]\Big[\mathbb{E}_h[\frac{qf'f'^T}{f^2}f^{1-q}] - \mathbb{E}_h[\frac{f''_\theta}{f}f^{1-q}]\Big]^{-1} \end{aligned}$$

Taking inverse of both sides of the equation, we have

$$\begin{aligned} \underbrace{[AB^{-1}]^{-1}}_{K(\epsilon,q)} &= \Big[\mathbb{E}_h[\frac{qf'f'^T}{f^2}f^{1-q}] - \mathbb{E}_h[\frac{f''_\theta}{f}f^{1-q}]\Big]\mathbb{E}_h[\frac{f'f'^T}{f^2}(f^{1-q})^2]^{-1} \\ &= \underbrace{\mathbb{E}_h[\frac{qf'f'^T}{f^2}f^{1-q}]\mathbb{E}_h[\frac{f'f'^T}{f^2}(f^{1-q})^2]^{-1}}_{M(\epsilon,q)} \underbrace{-\mathbb{E}_h[\frac{f''_\theta}{f}f^{1-q}]\mathbb{E}_h[\frac{f'f'^T}{f^2}(f^{1-q})^2]^{-1}}_{N(\epsilon,q)} \qquad (1) \end{aligned}$$

We define $K(\epsilon, q) = [AB^{-1}]^{-1}$, $M(\epsilon, q) = \mathbb{E}_h[\frac{f'f'^T}{f^2}(f^{1-q})^2]\mathbb{E}_h[\frac{qf'f'^T}{f^2}f^{1-q}]^{-1}$ and $N(\epsilon, q) = -\mathbb{E}_h[\frac{f''_\theta}{f}f^{1-q}]\mathbb{E}_h[\frac{f'f'^T}{f^2}(f^{1-q})^2]^{-1}$.

When $q = 1$, the eigenvalues of M, $\lambda_j(M)$, is 1. $A(\epsilon = 0, q = 1) = B(\epsilon = 0, q = 1)$ are both positive definite. Since A and B are continuous in ϵ, therefore, there exists an interval $\epsilon \in [0, E]$ such are A and B have all positive eigenvalues. From the assumption A2, we also know $\mathbb{E}_g[f''_\theta(X, \theta)/f(X, \theta)]$ has all positive eigenvalues. Therefore, $N(\epsilon, q = 1)$ has all negative eigenvalues. For all the eigenvalues of K, $\lambda_j(K)$, when $\epsilon > 0$ and $q = 1$, we have

$$\lambda_j(K) = \lambda_j(M + N) \leq \lambda_1(M) + \lambda_1(N) < 1$$

Therefore, for the eigenvalues of AB^{-1}, $\lambda_j(AB^{-1}) = 1/\lambda_j(K) > 1$. □

7.7 Proof of Theorem 7

Theorem 7: Under assumptions A1, A2 and A3, for $\epsilon > 0$ there exists a q such that, we have $|\lambda_j(A(\epsilon, q)B(\epsilon, q)^{-1}) - 1| < |\lambda_j(A(\epsilon, 1)B(\epsilon, 1)^{-1}) - 1|$ for $j = 1, ..., r$.

Proof. Since $\psi_q(x,\theta) = \frac{f'}{f} f^{1-q}$ is a bounded function in x, therefore each element of A is bounded and the trace of A, $|\text{trace}(A)| = |\sum_{j=1}^{r} \lambda_j(A)|$, is also bounded. We know that all eigenvalues of A are positive, which implies each eigenvalue of A is also bounded, $|\lambda_j(A)| < C_1$. Meanwhile, $|\frac{\partial}{\partial\theta}(\frac{f'}{f} f^{1-q})|$ is bounded. Since $\frac{\partial}{\partial\theta}(\frac{f'}{f} f^{1-q})$ is a continuous function in q, therefore by setting $q < 1$, the eigenvalues of B is also bounded, $C_2 < |\lambda_j(B)| < C_3$. For the eigenvalues of AB^{-1}, $\lambda_j(AB^{-1})$, we have $\lambda_j(AB^{-1}) < \lambda_1(A)/\lambda_r(B)$, which is also bounded. On the other hand, when $q = 1$, $\frac{f'}{f} f^{1-q} = \frac{f'}{f}$ is clearly unbounded for different models (e.g. normal distribution), which implies the eigenvalues of $\lambda_j(AB^{-1}) > \lambda_r(A)/\lambda_1(B)$ is also unbounded. Therefore, we can find a contamination component g, such that $\lambda_j(AB^{-1}(\epsilon, q)) < \lambda_j(AB^{-1}(\epsilon, 1))$ for $\epsilon > 0$ and $q < 1$. Since $\lambda_j(AB^{-1}(\epsilon, 1)) > 1$ and $\lambda_j(AB^{-1}(\epsilon, q))$ is continuous in q, therefore, $|\lambda_j(A(\epsilon, q)B(\epsilon, q)^{-1}) - 1| < |\lambda_j(A(\epsilon, 1)B(\epsilon, 1)^{-1}) - 1|$ for $j = 1, ..., r$. □

7.8 Proof of Theorem 8

We will prove the simple null hypothesis case. First, we rewrite the LqLR test statistic as

$$D_q = \frac{1}{2}(\hat{\theta}_q - \theta_0)^T nB(\theta^*)(\hat{\theta}_q - \theta_0)$$

where θ^* is between $\hat{\theta}_q$ and θ_0. Since we know $B(\theta)$ is continuous in θ, we conclude that LqLR and $\hat{\theta}_q$ have the same breakdown points.

References

Agostinelli, C. and Markatou, M. (2001). Test of hypotheses based on the weighted likelihood methodology. *Statistica Sinica*, 11:499–514.

Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27(325-336):186–190.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *Annals of Statistics*, 5(3):445–463.

Bickel, P. J. and Doksum, K. A. (2007). *Mathematical Statistics Basic Ideas and Selected Topics Volume I.* Pearson Prentice Hall, second edition.

Cushny, A. R. and Peebles, A. R. (1905). The action of optical isomers ii. hyoscines. *Journal of Physiology*, 32(5-6):501–510.

Ferrari, D. and Yang, Y. (2010). Maximum Lq-likelihood estimation. *Annals of Statistics*, 38:753–783.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, first edition.

He, X. (1991). A local breakdown property of robust tests in linear regression. *Journal of Multivariate Analysis*, 38:294–305.

He, X., Simpson, D. G., and Portnoy, S. L. (1990). Breakdown robustness of tests. *Journal of the American Statistical Association*, 85(410):446–452.

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, 89(427):897–904.

Huber, P. J. (1965). A robust version of the probability ratio test. *Annals of Mathematical Statistics*, 36(6):1753–1758.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics.* Wiley, second edition.

Lehmann, E. L. and D'Abrera, H. (2006). *Nonparametrics: Statistical Methods Based on Ranks*. Springer, first edition.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60.

Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93(442):740–750.

Ronchetti, E. M. (1979). Robustheitseigenschaften von tests. *Diploma thesis, ETH, Zurich.*

Ronchetti, E. M. (1982a). Robust alternatives the f test for the linear model. *Probability and Statistical Inference, eds. W. Grossman, C. Pflug, and W. Wertz, Dortrecht: Reidel*, pages 329–342.

Ronchetti, E. M. (1982b). Robust testing in linear models: the infinitesimal approach. *Ph.D. Thesis, ETH, Zurich.*

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing*. Wiley, first edition.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50 (1):1–25.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–86.