Statistica Sinica **11**(2001), 499-514

TEST OF HYPOTHESES BASED ON THE WEIGHTED LIKELIHOOD METHODOLOGY

Claudio Agostinelli and Marianthi Markatou

University of Padua and Columbia University

Abstract: Weighted versions of the likelihood ratio, Wald, score and disparity tests are proposed for parametric inference. If the parametric model is correct, the weighted likelihood tests are asymptotically equivalent to the corresponding likelihood based tests, while the disparity test has asymptotically the same distribution as that of $\sum_{i=1}^{p} \lambda_i Z_i^2$, where Z_i are standard normal random variables and λ_i are eigenvalues of an appropriate matrix. The tests have high level and power breakdown points and they perform well in finite samples. A simulation study and a data example illustrate the performance of the tests in the presence of symmetric and asymmetric contamination.

Key words and phrases: Asymptotic distribution, breakdown function, likelihood ratio test, robustness, score test, Wald test, weighted likelihood.

1. Introduction

The investigation of robustness of testing procedures dates back to 1931, when E. S. Pearson discovered the nonrobustness of the test of variances. However, most research effort has focused on robust estimation and less attention has been given to tests.

The aim of robust testing is twofold. The level of the test should be stable under small, arbitrary departures from the null hypothesis (*robustness of validity*), and the test should retain good power under small, arbitrary departures from specified alternatives (*robustness of efficiency*). Many authors have proposed tests with desirable robustness properties. Those include Huber (1965, 1981), Schrader and Hettmansperger (1980), Ronchetti (1982), Markatou and Hettmansperger (1990), Markatou and He (1994), Heritier and Ronchetti (1994), He (1991) and Simpson (1989), to mention just a few. For a review, see Markatou, Stahel and Ronchetti (1991).

In this article, we present robust versions of the Wald, score, and likelihood ratio-type tests that are based on the weighted likelihood estimators (WLEE) proposed by Markatou, Basu and Lindsay (1997, 1998). Section 2 gives a short review of the weighted likelihood methodology. Section 3 presents the test statistics under study and looks at disparity test analogous to the one studied by

Simpson (1989) and Lindsay (1994). Section 4 contains results on the asymptotic and stability properties of the proposed tests. It is shown that, when the model is correctly specified, the weighted likelihood ratio-type, the Wald-type and the score-type tests have, under the null hypothesis, a central chi-square distribution. The disparity-based test, has asymptotically the same distribution as that of the quadratic form $\sum_{i=1}^{p} \lambda_i Z_i^2$, where Z_i are $N(0,1)$ random variables and λ_i are eigenvalues of an appropriate matrix. This is in constrast with results obtained by Simpson (1989) and Lindsay (1994).

The new tests have a high breakdown point in terms of level and power. Markatou and He (1994) showed that the breakdown points of the Wald test that uses one-step high breakdown point estimates are determined by the breakdown points of the parameter estimate and the associated variance-covariance matrix. They also proved that the score-type and likelihood ratio-type tests exhibit high level breakdown but not high power breakdown. In contrast with those results, our tests have high level and power breakdown points as long as the hypothesized models belong to the exponential family. Section 5 presents an example and simulations to illustrate the performance of the tests.

2. Background

Let $X_1, X_2, \ldots, X_n$ be a random sample with density function $m_\theta(x)$ corresponding to a continuous probability measure $M_\theta(x)$. Let $u(x;\theta) = \frac{\partial}{\partial\theta} \log m_\theta(x)$ be the score function. Under regularity conditions the maximum likelihood estimator of θ is a solution of the score equation $\sum_{i=1}^{n} u(x_i;\theta) = 0$.

Let $\hat{F}_n$ be the empirical distribution function. Given any point x_i in the sample space, a weight function $w(x_i; M_\theta, \hat{F}_n)$ is constructed. The parameter estimates, obtained as solutions of the set of estimating equations

$$\frac{1}{n} \sum_{i=1}^{n} w(x_i; M_\theta, \hat{F}_n) u(x_i;\theta) = 0, \tag{2.1}$$

are called the weighted likelihood estimators (Markatou et al. (1998)). The weight function is defined as $w(x; M_\theta, \hat{F}_n) = \min\left\{1, \frac{[A(\delta_\theta(x))+1]^+}{\delta_\theta(x)+1}\right\}$, where $[\cdot]^+$ denotes positive part. For an extensive discussion and motivation, see Markatou et al. (1997, 1998).

The quantity $\delta_\theta(x)$ that enters in the construction of the weights is called the Pearson residual and is defined as $\delta_\theta(x) = f^*(x)/m_\theta^*(x) - 1$, where $f^*(x) = \int k(x;t,h)\ d\hat{F}_n(t)$ is a kernel density estimator and $m_\theta^*(x) = \int k(x;t,h)\ dM_\theta(t)$ is the smoothed model density. The Pearson residual expresses the agreement between the data and the assumed probability model. In what follows we suppress the dependence of δ on the parameter θ. The function $A(\cdot)$ in the definition of the weight is a residual adjustment function (RAF, Lindsay (1994)). It operates on Pearson residuals as the Huber ψ-function operates on ordinary residuals. When $A(\delta) = \delta$ the weight equals 1, and this corresponds to maximum likelihood. The choice $A(\delta) = 2\{(\delta+1)^{1/2} - 1\}$ corresponds to Hellinger distance. Generally, the weight functions we work with use an $A(\cdot)$ that corresponds to a minimum disparity problem. Since the weights are functions of δ the notation $w(\delta)$ is used.

The estimating equations (2.1) do not necessarily have a unique solution. Let $\theta_1, \theta_2, \ldots, \theta_p$ denote the roots of (2.1) and let $\rho(f^*, m^*_\theta)$ be the *parallel disparity measure* (Markatou et al. (1998)), that is, the disparity constructed using the RAF of the weight formulation. The parallel disparity is defined as $\rho\left(f^*, m^*_\theta\right) = \int G\left(\delta(x)\right) m^*_\theta dx$, where G is a strictly convex, thrice differentiable function. When $G(\delta) = 2\delta^2/(\delta+2)$ then $\rho\left(f^*, m^*_\theta\right)$ corresponds to a chi-squared distance. Notice also that the RAF $A(\delta)$ is defined as $A(\delta) = (\delta+1)G'(\delta) - G(\delta)$, prime denoting differentiation.

To guarantee uniqueness, define the weighted likelihood estimator θ_w as $\text{argmin}_{\theta_i\ i=1,\ldots,p}\rho(f^*, m^*_{\theta_i})$. Note that the weighted likelihood estimators are asymptotically normal with the inverse of the Fisher information as covariance matrix.

The computation of Pearson residuals requires selection of the smoothing parameter h. This is done by setting $h^2 = k\hat{\sigma}^2$, where k is a constant independent of the scale of the model. It provides the mean downweighting that occurs when the model is correct. In Section 5 we discuss an alternative way to obtain the parameter k in location-scale models and we use it in our examples and simulations.

3. The Tests

We present the test statistics under study. The objective here is to create test procedures that are asymptotically equivalent to the corresponding tests based on maximum likelihood when the model is correct, but are more robust in the sense of preserving size and power when the true density is contaminated. Let Θ be the parameter space, a subset of $\mathcal{R}^p$. We are interested in testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$, $\Theta_1 \subseteq \Theta/\Theta_0$.

We consider three classes of tests.

1. A Wald-type test statistic is a quadratic form

$$W(\theta) = (\theta - \hat{\theta})^t \{nI(\hat{\theta})\}(\theta - \hat{\theta}), \tag{3.2}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ in Θ, the superscript t denotes transpose, and the Fisher information is evaluated at $\hat{\theta}$.

The Wald test statistic we use is

$$W_w(\theta) = (\theta - \hat{\theta}_w)^t \Big\{ \sum_{i=1}^{n} w(\delta_{\hat{\theta}_w}(x_i)) I(\hat{\theta}_w) \Big\} (\theta - \hat{\theta}_w), \tag{3.3}$$

where $\hat{\theta}_w$ is the weighted likelihood estimator of θ in Θ.

A comparison between (3.2) and (3.3) shows that we have replaced n in (3.2) by the sum of the weights that the observations receive when the unrestricted weighted likelihood estimator is used, and the asymptotic Fisher information is evaluated at the unrestricted weighted likelihood estimator.

2. A score-type test function is

$$T(\theta) = S^t(\theta)(nI(\theta))^{-1} S(\theta), \tag{3.4}$$

where $S(\theta) = \sum_{i=1}^{n} u(x_i, \theta)$ is the score function evaluated at θ.

We use the score-type test function

$$T_w(\theta) = S_w^t(\theta) \Big[\sum_{i=1}^{n} w\left(\delta_{\hat{\theta}_w}(x_i)\right) I(\theta) \Big]^{-1} S_w(\theta), \tag{3.5}$$

where $S_w(\theta) = \sum_{i=1}^{n} w\left(\delta_{\hat{\theta}_w}(x_i)\right) u(x_i, \theta)$ and $\hat{\theta}_w$ is the unrestricted weighted likelihood estimate. A comparison between (3.4) and (3.5) shows that we replaced n by the sum of weights that the observations receive in the unrestricted model. The weights evaluated at the unrestricted estimate guarantee the consistency of the test.

3. A likelihood ratio-type test function is

$$\Lambda_w(\theta) = -2 \sum_{i=1}^{n} w\left(\delta_{\hat{\theta}_w}(x_i)\right) \left[l(x_i; \theta) - l(x_i; \hat{\theta}_w) \right], \tag{3.6}$$

where $l(x_i; \theta) = \log m_\theta(x_i)$ is the log-likelihood.

A disparity-based test analogous to the one proposed by Simpson (1989) can also be constructed. The test statistic is

$$D(\theta) = -2 \left\{ \rho(f^*, m^*_{\hat{\theta}_w}) - \rho(f^*, m^*_\theta) \right\}, \tag{3.7}$$

where ρ is the parallel disparity. If ρ is the log-likelihood the disparity test is analogous to the likelihood ratio except that the convoluted model is used instead of the original model $m_\theta(x)$.

Notice that when the data are from a $N(\mu, \sigma^2)$ with σ^2 known, the Wald, score and likelihood ratio-type tests have exactly the same form.

To define the tests under a composite null hypothesis and/or in the presence of nuisance parameters, we evaluate them at

$$\hat{\theta}_0 = \mathrm{argmax}_{\theta\in\Theta_0} \prod_{i=1}^{n} m_\theta(x_i)^{w(\delta_{\hat{\theta}_w}(x_i))}, \tag{3.8}$$

which is weighted likelihood with fixed weights. This is similar to the classical approach to a composite null hypothesis problem (see Cox and Hinkley (1974) among others). The computation can be done by numerical optimization under constraints.

4. Asymptotic and Robustness Properties

In this section we discuss the asymptotic and robustness properties of the different test statistics. Here are the necessary assumptions.

A1. The weight function $w(\delta)$ is a nonnegative, bounded, and differentiable function.

A2. The weight function $w(\delta)$ has $w'(\delta)(\delta+1)$ bounded.

A3. The kernel $k(x;t,h)$ is a function of bounded variation and is bounded for all x by a constant $M(h)$ that may depend on h, but not on t or x.

A4. The Fisher Information matrix is positive definite.
Let $\widetilde{u}(x;\theta) = \frac{\partial}{\partial\theta}\log m^*_\theta(x)$ and recall that $u(x;\theta) = \frac{\partial}{\partial\theta}\log m_\theta(x)$, further let $\widetilde{u}'(x;\theta) = \frac{\partial}{\partial\theta}\widetilde{u}(x;\theta)$, $u'(x;\theta) = \frac{\partial}{\partial\theta}u(x;\theta)$ and $u''(x;\theta) = \frac{\partial}{\partial\theta}u'(x;\theta)$.

A5. For every $\theta_0 \in \Theta$, there is a neighborhood $N(\theta_0)$ such that for $\theta \in N(\theta_0)$, the quantities $|\widetilde{u}(x;\theta)u'(x;\theta)|$, $\left|\tilde{u}^2(x;\theta)u(x;\theta)\right|$, $|\tilde{u}'(x;\theta)u(x;\theta)|$ and $|u''(x;\theta)|$ are bounded by $M_i(x)$, where $E_{\theta_0}[M_i(X)] < \infty$, $i = 1,2,3,4$.

A6. $E_{\theta_0}\left[\widetilde{u}^2(X;\theta)u^2(X;\theta)\right] < \infty$.

A7. a. $\int\left|\frac{\partial}{\partial\theta}m_\theta(x)/m^*_\theta(x)\right|\ dx < \infty$; b. $\int|\tilde{u}(x;\theta)u(x;\theta)|\left[m_\theta(x)/m^*_\theta(x)\right]\ dx < \infty$; c. $\int|u'(x;\theta)|\left[m_\theta(x)/m^*_\theta(x)\right]\ dx < \infty$.

Theorem 4.1. *Under the conditions* A1-A7 *and a correctly specified model, the Wald test, the score-type test and the likelihood ratio-type test are asymptotically distributed as central chi-squares under* $H_0 : \theta = \theta_0$*, with degrees of freedom* $dim(\Theta) - dim(\Theta_0)$*. Under local alternatives* $H_1 : \theta = \theta_0 + \Delta/\sqrt{n}$*, the asymptotic distribution is non-central chi-square with the same degrees of freedom and a non-centrality parameter* $0.5(\Delta/\sqrt{n})^t I(\theta)(\Delta/\sqrt{n})$.

To prove this theorem we need two auxiliary lemmas. Let us write the density estimator at X_1 as $f^*(X_1) = (1/n)\sum_{=1}^{n}(k(X_i;X_1,h) + k(X_1;X_1,h)) = a_n f^*_{<1>}(X_1) + b_n$, where $a_n = (n-1)/n$, $b_n = k(X_1;X_1,h)/n$ and $f^*_{<1>}(.)$ is the kernel density estimate computed without the first observation.

Lemma 4.1. *For* $p \in [0,2]$

$$\begin{aligned} E[Y_n^p(t)] &\leq E[|\frac{a_n f^*_{<1>}(t) + b_n - m^*_\theta(t)}{m^*_\theta(t)}|^p] n^{p/2} \\ &\leq \frac{[a_n \lambda(t) + (1/n)(k(X_1;X_1,h) - m^*_\theta(t))^2]^{p/2}}{[m^*_\theta(t)]^p}, \end{aligned}$$

where $\lambda(t) = \operatorname{Var}(k(X;t,h))$ *and* $Y_n(t) = n^{1/2}((\frac{a_n f^*_{<1>}(t)+b_n}{m^*_\theta(t)})^{1/2} - 1)^2$.

Proof. Begin by writing

$$\begin{aligned} &E\Big[\frac{a_n f^*_{<1>}(t) + b_n - m^*_\theta(t)}{m^*_\theta(t)}\Big]^2 \\ &= \Big(E\Big[\frac{a_n f^*_{<1>}(t) + b_n - m^*_\theta(t)}{m^*_\theta(t)}\Big]\Big)^2 + \operatorname{Var}\Big(\frac{a_n f^*_{<1>}(t) + b_n - m^*_\theta(t)}{m^*_\theta(t)}\Big), \end{aligned}$$

where,

$$E\Big[\frac{a_n f^*_{<1>}(t) + b_n - m^*_\theta(t)}{m^*_\theta(t)}\Big] = \frac{k(X_1,t,h) - m^*_\theta(t)}{n m^*_\theta(t)},$$

$$\operatorname{Var}\Big[\frac{a_n f^*_{<1>}(t) + b_n - m^*_\theta(t)}{m^*_\theta}\Big] = \frac{a_n^2 \lambda(t)}{(n-1) m^{*2}_\theta(t)}.$$

By a standard inequality,

$$E\Big[\frac{a_n f^*_{<1>}(t)+b_n-m^*_\theta(t)}{m^*_\theta(t)}\Big]^p \leq \frac{((1/n)^2[k(X,t,h)-m^*_\theta(t)]^2+(a_n^2/(n-1))\lambda(t))^{p/2}}{(m^*_\theta(t))^p}.$$

Multiply this last expression by $n^{p/2}$ to get the desired result.

Lemma 4.2. *For* $p \in (0,2)$*,* $\lim E(Y_n^p(t) = 0$*, as* n *converges to infinity.*

Proof. As n converges to infinity we get convergence in probability to 0 for each t. From Lemma 4.1, for each $q < 2$, the L_2 norm of the sequence is bounded.

Proof of the Theorem. Let $\Lambda(\theta) = -2\sum_{i=1}^n \Big[l(x_i;\theta) - l(x_i;\hat\theta)\Big]$ be the classical likelihood ratio test function. Then

$$\begin{aligned} &\frac{1}{n}|\Lambda_w(\theta) - \Lambda(\theta)| \\ &= \frac{2}{n}\Big|\sum_{i=1}^n w(\delta_{\hat\theta_w}(x_i))[l(x_i;\theta) - l(x_i;\hat\theta_w)] - \sum_{i=1}^n [l(x_i;\theta) - l(x_i;\hat\theta)]\Big| \\ &= \frac{2}{n}\Big|\sum_{i=1}^n [w(\delta_{\hat\theta_w}(x)) - 1](l(x_i;\theta) - l(x_i;\hat\theta)) + \sum_{i=1}^n w(\delta_{\hat\theta_w}(x_i))[l(x_i;\hat\theta) - l(x_i;\hat\theta_w)]\Big| \\ &\leq A_1 + A_2, \end{aligned}$$

where

$$A_1 = \frac{2}{n}\Big|\sum_{i=1}^{n}[w(\delta_{\hat{\theta}_w}(x_i) - 1](l(x_i;\theta) - l(x_i;\hat{\theta}))\Big|,$$

$$A_2 = \frac{2}{n}\sum_{i=1}^{n}|l(x_i;\hat{\theta}) - l(x_i;\hat{\theta}_w)|.$$

Hence by Lemma 4.1, Lemma 4.2, the root-n consistency of $\hat{\theta}$ and $\hat{\theta}_w$, and the fact that A_1, A_2 asymptotically converge in probability to zero, the weighted likelihood ratio test is asymptotically equivalent to the likelihood ratio test for $\theta = \theta_T$, the true parameter value.

To study the asymptotic local power, let $\epsilon = \Delta/\sqrt{n}$ and $\theta = \theta_T + \epsilon$. Then $1/n\ |\Lambda_w(\theta) - \Lambda(\theta)| \leq B_1 + B_2 + B_3$, where

$$B_1 = \frac{2}{n}\Big|\sum_{i=1}^{n}[w(\delta_{\hat{\theta}_w}(x_i)) - 1](l(x_i;\theta) - l(x_i;\theta_T))\Big|,$$

$$B_2 = \frac{2}{n}\Big|\sum_{i=1}^{n}[l(x_i;\theta_T) - l(x_i;\hat{\theta})]\Big|,$$

$$B_3 = \frac{2}{n}\Big|\sum_{i=1}^{n}w(\delta_{\hat{\theta}_w}(x_i))[l(x_i;\theta_T) - l(x_i;\hat{\theta}_w)]\Big|.$$

Under A1-A7 and the above two lemmas, each of the above quantities converges in probability to zero as $n \to \infty$. Thus, the weighted likelihood-based test and the likelihood ratio test have asymptotically, under the correct model, the same local power.

Similarly we can prove the equivalence of the score and Wald type tests to those based on the maximum likelihood estimator. This completes the proof of the theorem.

For the disparity test statistic we have the following result.

Theorem 4.2. *Under the conditions* A1-A7 *and a correctly specified model, the asymptotic distribution of the disparity-based test is given by a weighted sum of independent chi-squares with one degree of freedom, with weights the eigenvalues of the matrix $I^*(\theta)I^{-1}(\theta)$, where $I^*(\theta)$ is Fisher information calculated using the smoothed model.*

Proof. For the derivation of the asymptotic results we rely on the quadratic approximation of the disparity test statistic. The proof is then similar to the one given in Markatou and He (1994).

Notice that the asymptotic distribution of the disparity based test is different from the one obtained by Simpson (1989). This is the effect of kernel smoothing and the fact that the weighted likelihood estimators are not solutions of an optimization problem based on the disparity ρ. The asymptotic distribution of the disparity test will be a chi-square only if $I^{-1}(\theta)$ is the c-inverse of $I^*(\theta)$, which in turn means that the bandwidth should converge to 0.

We now turn to the study of robustness properties of the tests. The robustness of an estimator can be studied via the concepts of influence function and breakdown point. Ronchetti (1982) extended the definition of the influence function to tests, but the influence functions of the weighted likelihood tests are the same as those of the corresponding maximum likelihood-based tests. Therefore we study the breakdown of the new procedures.

Roughly speaking, the breakdown point of a statistical functional is the smallest fraction of contamination in the data that can lead to arbitrarily extreme results. Accordingly one might say that a test statistic breaks down, for a given level of contamination, if its p-value can be driven to its maximum or minimum achievable value. For ease of presentation, our analysis focuses on the abstract functionals associated with the test statistics. These functionals are given as follows:

$$W_w(\theta, F) = (\theta - \theta_w(F))^t \Big\{ \int w(\delta_{\theta_w(F)}(x)) dF(x) \Big\} I\left(\theta_w(F)\right) \left(\theta - \theta_w(F)\right); \quad (4.9)$$

$$T_w(\theta, F) = \Big(\int w(\delta_{\theta_w(F)}(x)) u(x;\theta) dF(x) \Big)^t \Big[\int w(\delta_{\theta_w(F)}(x)) dF(x) I(\theta) \Big]^{-1} \times \Big(\int w(\delta_{\theta_w(F)}(x)) u(x;\theta) dF(x) \Big); \quad (4.10)$$

$$\Lambda_w(\theta, F) = -2 \int w(\delta_{\theta_w(F)}(x)) \{ l(x;\theta) - l(x;\theta_w(F)) \} dF(x). \quad (4.11)$$

Recall that $\theta_w(F)$ denotes the functional that introduces the weighted likelihood estimator.

In what follows we define breakdown functions for the three tests. These definitions are similar to the ones presented in Simpson (1989). For an extensive discussion of breakdown concepts in hypotheses testing, see He, Simpson and Portnoy (1990).

Let $t(\theta, F)$ be a generic functional with domain $\mathbf{F}$, the set of all proper distributions on $\mathcal{R}^d$. Let $t_{max} = \sup_{F \in \mathbf{F}} \inf_{\theta \in \Theta_0} t(\theta, F)$ and $t_{min} = \inf_{F \in \mathbf{F}} \inf_{\theta \in \Theta_0} t(\theta, F)$, and define the level breakdown function as

$$\varepsilon_0(M_{\theta_T}; t) = \inf \Big\{ \varepsilon : \sup_{G \in \mathbf{F}} \inf_{\theta \in \Theta_0} t\left(\theta, (1-\varepsilon) M_{\theta_T} + \varepsilon G\right) = t_{max} \ ; \theta_T \in \Theta_0 \Big\}$$

and the power breakdown function as

$$\varepsilon_1(M_{\theta_T}; t) = \inf \Big\{ \varepsilon : \inf_{G \in \mathbf{F}} \inf_{\theta \in \Theta_0} t\left(\theta, (1-\varepsilon) M_{\theta_T} + \varepsilon G\right) = t_{min} \ ; \theta_T \in \Theta_1 \Big\},$$

where G is the contaminating distribution, $\mathbf{F}$ is assumed convex, and θ_T is the "true" value of the parameter. Heuristically, ε_s is the smallest fraction of contamination that can drive the p-value to s and is of most interest under the hypotheses H_s $(s = 0, 1)$. If $\theta_T \in \Theta_1$ and $t_{min} = 0$, then $\varepsilon_1(M_{\theta_T}; t)$ is the smallest fraction of contamination on M_{θ_T} that can make the test inconsistent. We then define the power breakdown point, $\varepsilon_1^*(t) = \sup_{\theta_T \in \Theta_1} \varepsilon_1(M_{\theta_T}, t)$, and the level breakdown point $\varepsilon_0^*(t) = \sup_{\theta_T \in \Theta_0} \varepsilon_0(M_{\theta_T}, t)$.

Theorem 4.3. *Assume* A1-A7 *hold and that the models belong to the exponential family. Then the level and power breakdown points of W_w, T_w and Λ_w tests are the same as the breakdown point ε^* of the weighted likelihood estimator.*

Proof. Clearly $W_{min} = 0$ and $W_{max} = \infty$. Power breakdown occurs if the test statistic can be driven to W_{min} when $\theta_\tau \in \Theta_1$. Because the breakdown point of $\hat{\theta}_w$ is ε^* and if $\|\theta - \theta_T\|$ is bigger than the maximum bias of θ_w then $\|\theta - \hat{\theta}_w\|$ is away from 0 if $\varepsilon < \varepsilon^*$. Moreover, if $\varepsilon < \varepsilon^*$, $\sum_{i=1}^n w\left(\delta_{\hat{\theta}_w}(x_i)\right)$ stays away from 0 and the power breakdown is ε^*.

Level breakdown occurs if the test statistics can be driven to its largest value W_{max} when $\theta_T \in \Theta_0$. By the triangle inequality we have $\|\theta - \theta_T\| < \infty \Leftrightarrow \|\theta - \hat{\theta}_w(F)\| < \infty$ for all $\varepsilon \leq \varepsilon^*$. Hence, since conditions A1 and A4 hold, W_{max} is reached if and only if the test based on the "true" value θ_T also reached this value (for every $\varepsilon \leq \varepsilon^*$). This show that the level breakdown point $\varepsilon_0(W_w)$ is ε^*.

To address the breakdown analysis of the likelihood ratio and the score tests we first rewrite them in a different form. Using a Taylor expansion of $l(x;\theta) = \log m_\theta(x)$ in a neighborhood of $\hat{\theta}_w$, we get

$$\begin{aligned}
&\Lambda_w(\theta, F)\\
&= -2\int \Big[w(\delta_{\theta_w(F)}(x))\Big(u(x;\theta_w(F))(\theta-\theta_w(F)) + \frac{1}{2}\frac{\partial}{\partial\theta}u(x;\theta)\Big|_{\theta=\theta^*}(\theta-\theta_w(F))^2\Big)\Big]\\
&= -\int \Big[w(x;\theta_w(F),F)\frac{\partial}{\partial\theta}u(x;\theta)\Big|_{\theta=\theta^*}\Big](\theta-\theta_w(F))^2,
\end{aligned}$$

where θ^* is between θ and $\hat{\theta}_w$. Now from A1 and A4, and the continuity of the Fisher information with respect to θ, we deduce high level and power breakdown for models that belong to the exponential family. Similarly, we can prove high level and power breakdown for the score test.

5. Examples and Simulations

In this section we present a simulation of the performance of the test procedures in terms of their level and power. We also present an example to compare the new procedures with known procedures.

Data were generated from the models $(1-\varepsilon)N(0,1)+\varepsilon N(0,25)$ and $(1-\varepsilon)N(0,1)+\varepsilon N(8,1)$, with $\varepsilon = 0\%$, 5%, 10%, 20%, 30%, 40%, 50%. We used the IMSL subroutine DRNNOR to generate the normal variates. The nominal model is assumed to be $N(\mu,\sigma^2)$. In our simulations we use sample sizes of 20 and 80.

To study the level of the tests we use $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. To study the power of the tests we use $H_0 : \mu = 0.5$ vs $H_1 : \mu \neq 0.5$, and we report the times the null hypothesis is rejected at a fixed level of significance. The variance σ^2 is treated as unknown and is estimated from the data. The weighted likelihood estimator is used; for the case where σ needs to be estimated under H_0, we use (3.8). The nominal levels of significance used were 0.1, 0.05, 0.01.

We now present an alternative to the way Markatou et al. (1998) select the parameter k that enters in the bandwidth construction. If we have a model M_θ, we can calculate an asymptotic residual function arising from contamination at the point y, in proportion ε, by calculating

$$\delta(x) = \frac{(1-\varepsilon)m_\theta^*(x)+\varepsilon k(x,y;h)}{m_\theta^*(x)} - 1.$$

In this case, $w(\delta(y))$ becomes the asymptotic weight at the outlier y.

To illustrate, assume the nominal model is $N(0,\sigma^2)$ and let $F_\varepsilon(x) = (1-\varepsilon)N(0,\sigma^2)+\varepsilon\Delta_y(x)$ be the true model. If we use a normal kernel with variance $k\sigma^2$, then the asymptotic Pearson residual at the targeted outlier value y is

$$\delta(y) = \varepsilon\Big(\frac{1+k}{k}\Big)^{1/2}\exp\Big(\frac{y^2}{2\sigma^2(1+k)}\Big) - 1. \qquad (5.12)$$

Now if the outlier is expressed on the standard deviation scale, $y = c\sigma$ where c is a constant, $\delta(y) = \varepsilon[(1+k)/k]^{1/2}\exp(c^2/2(1+k)) - 1$ independent of σ^2. From this formula, we can determine the downweighting applied to any fraction ε and outlier value y.

Thus, to calculate the weighted likelihood estimates we used a normal kernel with variance $h^2 = k\sigma^2$, $k = 0.003$. This corresponds to a weight of 0.2, which is obtained by defining an observation as an outlier when it is 3 or more standard deviations away from the mean, and the contamination is fixed at 20%.

In our simulations, we used weights based on the Hellinger distance RAF. The number of Monte Carlo replications is 5000 and the maximum standard error of the Monte Carlo trials is 0.00707. To obtain the initial values needed to start the algorithm that computes the weighted likelihood estimator, we used 100 bootstrap samples of size 2 to obtain the sample mean and sample variance.

In the case of symmetric contamination, the weighted likelihood equations have only one root. In the case of asymmetric contamination, the weighted likelihood equations have multiple roots. To obtain the estimator for use in our test statistics we evaluate the disparity at the different roots, and select the root that minimizes it. To compute the value of the test statistic on each simulated sample we need approximately 10 seconds of real time.

For comparison we also calculated the Wald-type test of Markatou and He (1994). This is a test based on a one-step estimator with high breakdown initial estimate. The one-step estimator used is that of Simpson, Ruppert and Carroll (1992) with Huber ψ-function (tuning constant 1.345). The initial high breakdown point estimate we use is the LTS (Rousseew (1984)). In the tables and figures we call this test HBT. Moreover, we report the classical Wald-type test based on maximum likelihood (W).

All programs were written in FORTRAN and Splus and all calculations were carried out on a SUN workstation at the Department of Statistics, Columbia University.

Here are the results for the symmetric contamination case. When the sample size is 20, the performance of the Wald-type test based on the weighted likelihood estimator, (W_w), is very similar to the performance of the HBT test. The likelihood-type test based on the weighted likelihood estimator, (Λ_w), and the score test based on the weighted likelihood estimator, (T_w), perform similarly, with T_w exhibiting the best performance in terms of level for high amounts of contamination. The Wald test based on the maximum likelihood estimator, W, performs very well in terms of level for all contamination percentages, but rapidly loses power. This result is in accord with the observations made by Tsou and Royall (1995). When the sample size increases, all three tests perform well in terms of level, and have improved power over W.

Table 1 presents the results for sample size 80. Notice that the HBT test performs well in terms of level for up to 15% contamination, and then breaks down. The poor performance of the HBT is due to the fact that it requires larger samples than the ones used here to hold its level and to exhibit good power. Moreover, its power performance improves when the alternative is not too close to the null hypothesis. The test on W again performs well in terms of level.

Table 2 presents the results when contamination is asymmetric. The superiority of the weighted likelihood-based tests is clear. Notice that the HBT and the W test perform similarly. When the sample size is 20, for contamination up to 30%, the Λ_w test performs best in terms of level. For higher percentages of contamination, the levels of all tests are inflated considerably. This behaviour is not observed when the sample size is 80.

Table 1. Level and power for the Wald tests based on the WLE, the MLE and the deviance test. The data come from $(1-\varepsilon)N(0,1)+\varepsilon N(0,25)$ and the hypothesized model is $N(\mu,\sigma^2)$. The sample size is 80.

		$H_0:\mu=0$ vs $H_1:\mu\neq 0$			$H_0:\mu=0.5$ vs $H_1:\mu\neq 0.5$		
Cont.	Test	10.0%	5.0%	1.0%	10.0%	5.0%	1.0%
$\varepsilon=0\%$	W_w	12.080	6.500	1.860	99.600	99.220	97.360
	Λ_w	11.640	6.120	1.740	99.580	99.220	97.000
	T_w	11.340	5.640	1.520	99.580	99.160	96.640
	W	10.100	5.120	1.120	99.620	99.300	97.380
	HBT	10.000	5.040	1.160	99.460	99.100	96.140
$\varepsilon=5\%$	W_w	11.80	6.600	1.860	99.44	98.880	95.860
	Λ_w	11.600	6.160	1.700	99.400	98.840	95.320
	T_w	11.160	5.800	1.540	99.360	98.760	94.620
	W	9.560	4.920	0.880	91.18	85.520	69.360
	HBT	11.20	5.500	1.480	99.16	98.460	93.800
$\varepsilon=10\%$	W_w	11.980	6.940	1.800	99.180	98.280	93.440
	Λ_w	11.640	6.580	1.640	99.160	98.100	92.580
	T_w	11.260	6.280	1.380	99.140	97.960	91.520
	W	10.500	4.600	0.680	78.760	69.440	48.280
	HBT	11.700	6.000	1.440	98.740	97.580	90.820
$\varepsilon=20\%$	W_w	12.480	6.760	1.820	97.320	94.940	83.960
	Λ_w	12.040	6.340	1.580	97.280	94.660	82.420
	T_w	11.660	6.040	1.320	97.220	94.340	80.760
	W	10.460	4.980	0.860	60.080	48.520	26.820
	HBT	13.340	7.580	2.100	96.220	92.840	81.460
$\varepsilon=30\%$	W_w	12.637	6.981	1.953	89.210	83.392	62.239
	Λ_w	11.982	6.201	1.460	89.338	82.937	64.933
	T_w	11.582	5.681	1.200	89.018	82.376	62.432
	W	10.210	5.074	0.994	52.198	31.102	19.628
	HBT	16.280	9.620	2.980	90.260	85.280	71.010
$\varepsilon=40\%$	W_w	12.983	7.393	2.399	70.731	60.716	40.130
	Λ_w	12.222	6.732	1.743	70.387	58.928	38.229
	T_w	11.821	6.211	1.403	69.746	58.926	35.784
	W	10.078	5.189	1.042	40.413	29.092	12.442
	HBT	19.120	12.20	4.520	82.280	75.780	58.900
$\varepsilon=50\%$	W_w	12.431	7.057	1.858	49.910	38.861	20.538
	Λ_w	11.854	6.348	1.402	48.899	37.405	18.803
	T_w	11.434	5.987	1.181	48.258	36.504	17.361
	W	10.272	5.226	0.901	34.622	24.569	9.932
	HBT	22.940	15.26	6.220	73.420	65.560	48.300

Table 2. Level and power for the Wald tests based on the WLE, the MLE and the deviance test. The data come from $(1-\varepsilon)N(0,1)+\varepsilon N(8,1)$ and the hypothesized model is $N(\mu,\sigma^2)$. The sample size is 80.

		$H_0:\mu=0$ vs $H_1:\mu\neq 0$			$H_0:\mu=0.5$ vs $H_1:\mu\neq 0.5$		
Cont.	Test	10.0%	5.0%	1.0%	10.0%	5.0%	1.0%
$\varepsilon=5\%$	W_w	12.100	6.960	1.860	99.500	98.980	96.320
	Λ_w	11.800	6.620	1.660	99.500	98.920	95.860
	T_w	11.420	6.280	1.460	99.480	98.900	95.080
	W	62.200	36.460	5.780	0.780	0.100	0.000
	HBT	18.620	11.260	3.080	96.100	92.420	78.720
$\varepsilon=10\%$	W_w	12.020	6.900	1.780	99.420	98.760	95.140
	Λ_w	11.660	6.560	1.600	99.380	98.720	94.580
	T_w	11.340	6.260	1.380	99.380	98.600	93.780
	W	99.880	97.860	68.800	5.480	1.080	0.020
	HBT	43.700	31.460	12.420	75.680	63.580	38.160
$\varepsilon=20\%$	W_w	13.540	8.060	3.100	98.800	97.000	91.280
	Λ_w	13.060	7.694	2.703	98.777	96.824	90.018
	T_w	12.705	7.299	2.466	98.659	96.646	88.538
	W	100.000	100.000	100.000	100.000	99.980	88.380
	HBT	94.780	89.440	68.600	69.000	51.360	19.740
$\varepsilon=30\%$	W_w	12.380	6.918	1.650	98.18	96.208	88.046
	Λ_w	10.927	5.642	0.965	97.945	95.701	85.717
	T_w	10.424	5.159	0.755	97.840	95.323	83.620
	W	100.00	100.00	100.00	100.00	100.00	99.60
	HBT	99.980	99.940	99.160	4.860	2.02	0.340
$\varepsilon=40\%$	W_w	11.26	6.028	1.542	96.25	92.67	79.211
	Λ_w	11.781	6.195	1.604	95.520	91.925	76.825
	T_w	11.338	5.642	0.940	95.299	91.427	73.119
	W	100.00	100.00	100.00	100.00	100.00	100.00
	HBT	94.78	89.44	68.60	4.88	2.04	0.360
$\varepsilon=50\%$	W_w	9.889	5.154	1.163	92.287	86.041	67.409
	Λ_w	8.621	3.448	0.431	90.948	83.190	62.500
	T_w	8.190	3.448	0.862	90.517	81.897	57.759
	W	100.00	100.00	100.00	100.00	100.00	100.00
	HBT	100.00	100.00	100.00	100.00	100.00	100.00

Figures 1, 2 and 3 show the performance of the various tests in terms of level and power under asymmetric contamination. The data consist of 100 points generated from $(1-\varepsilon)N(0,1)+\varepsilon N(6,0.5)$. Figure 1 shows the performance, in terms of level, as ε changes. The null hypothesis is $\mu=0$. In this case we report also a Wald-type test of Heritier and Ronchetti (1994) based on the Huber ψ-function. Since its behaviour is similar to the HBT based on one-step estimator, we do not report it in the following figure. Observe that, as ε increases, all tests but the WLEE-based become significant. When $\varepsilon = 0.1$, notice from Figure 2 that the W_w test becomes significant sooner than the remaining tests as the null hypothesis changes from $\mu = 0$ to $\mu = 1.5$. This indicates the superior performance of the WLEE-based test; this performance is also seen in Figure 3 in terms of power.

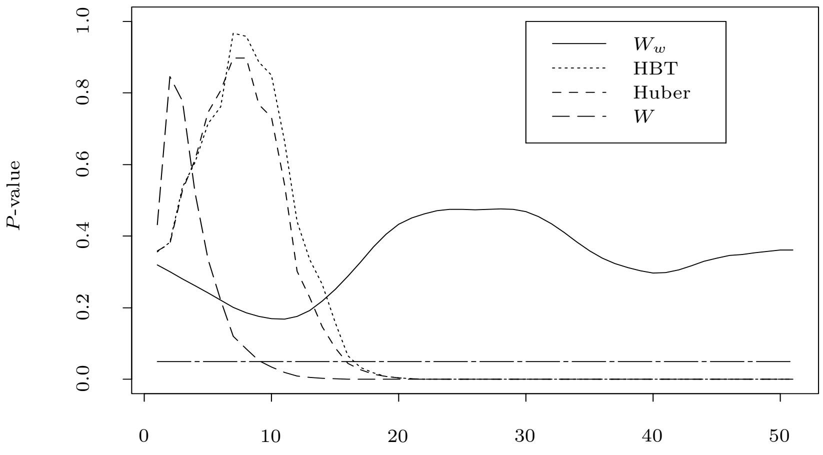

Figure 1. Level breakdown analysis of the W_w test based on weighted likelihood, HBT based on a one-step estimate (starting point LTS), Wald-type test based on a Huber M-estimator, and W based on likelihood. Data are $(1-\varepsilon)N(0,1) + \varepsilon N(6,0.5)$. The sample size is 100.

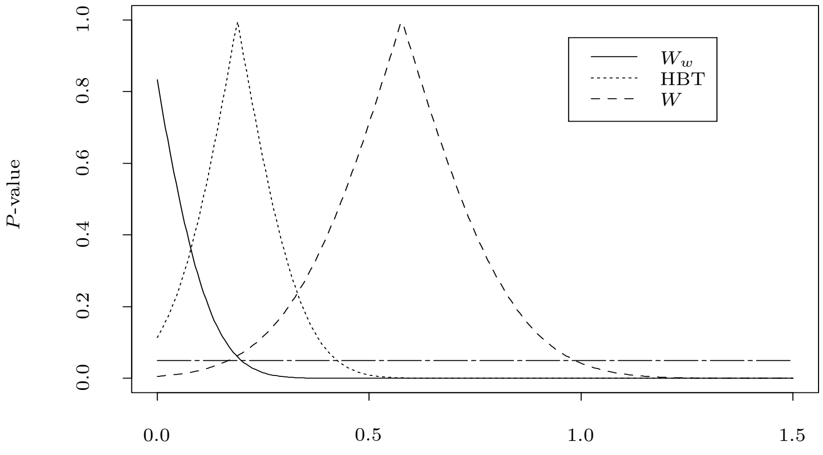

Figure 2. Levels of the W_w test based on weighted likelihood, HBT based on a one-step estimate (starting value LTS) and W based on likelihood. Data are $0.9N(0,1) + 0.1N(6,0.5)$. The sample size is 100.

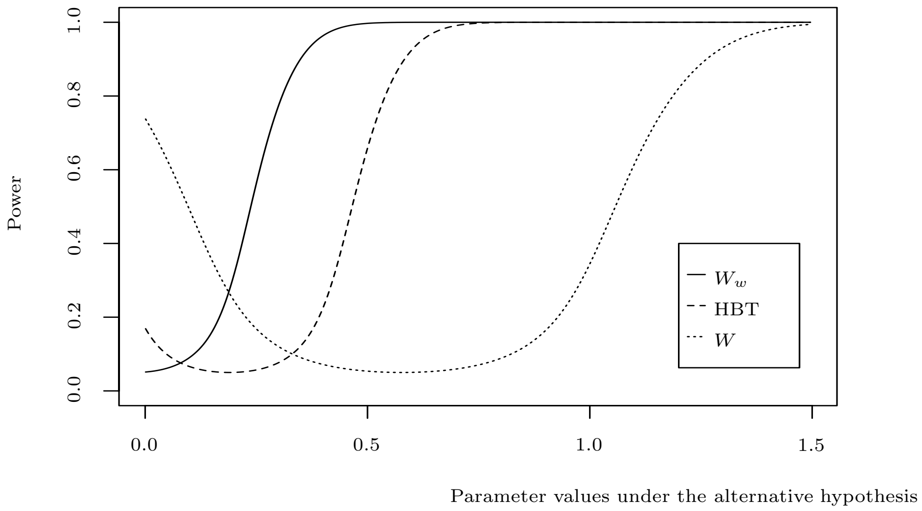

Figure 3. Power of the W_w test based on weighted likelihood, HBT based on a one-step (starting value LTS) and W based on likelihood. Data are $0.9N(0,1) + 0.1N(6,0.5)$. The sample size is 100.

To illustrate the applicability of the methodology, we use data on times between successive failures of air conditioning equipment in thirteen Boeing 720 aircrafts. The data set for plane number 7909 is taken from Proschan (1963), and has also been analyzed by Lawless (1982) and Keating, Glaser and Ketchum (1990). Lawless, using nonparametric methods, concluded there is a lack of evidence against the null hypothesis of exponentiality. Keating et al. (1990) model the data as a gamma distribution with shape parameter 1 and reject the hypothesis $H_0 : \theta = 1$ in favor of $H_1 : \theta > 1$. We used our methods to carry out the same test of exponentiality. To calculate the WLEE we used the folded normal density as a kernel with variance $h^2 = \hat{\sigma}^2$. The p-value obtained is approximately 0 and we reject H_0. Keating et al. (1990) reject the null hypothesis at the 5% level but not at the 1% level. A closer look at the data reveals outliers. Our Wald test rejects H_0 at both 5% and 1% levels.

Acknowledgements

The authors would like to thank an associate editor and two referees for their constructive comments. The work of M. Markatou was supported by NSF grant DMS-9973569.

References

Agostinelli, C. (1998). Inferenza statistica robusta basata sulla funzione di verosimiglianza pesata: alcuni sviluppi. Ph.D. Thesis, University of Padua, Padova.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman & Hall, London.

He, X. (1991). A local breakdown property of robust tests in linear regression. *J. Multivariate Anal.* **38**, 294-305.

He, X., Simpson, D. G. and Portnoy, S. L. (1990). Breakdown of robustness of tests. *J. Amer. Statist. Assoc.* **85**, 446-452.

Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36**, 1753-1758.

Huber, P. J. (1981). *Robust Statistics.* John Wiley, New York.

Heritier, S. and Ronchetti, E. (1994). Robust bounded influence tests in general parametric models. *J. Amer. Statist. Assoc.* **89**, 897-904.

Keating, J. P., Glaser, R. E. and Ketchum, N. S. (1990). Testing hypotheses about the shape parameter of a gamma distribution. *Technometrics* **32**, 67-82.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1018-1114.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* New York, John Wiley.

Markatou, M. and Hettmansperger, T. P. (1990). Robust bounded influence tests in linear models. *J. Amer. Statist. Assoc.* **85**, 187-190.

Markatou, M., Stahel, W. and Ronchetti, E. (1991). Robust M-type testing procedures for linear models. In *Directions in Robust Statistics and Diagnostics* (Part I) (Edited by W. Stahel and S. Weisberg), 201-220. Springer Verlag, New York.

Markatou, M. and He, X. (1994). Bounded influence and high breakdown point testing procedures in linear models. *J. Amer. Statist. Assoc.* **89**, 543-549.

Markatou, M., Basu, A. and Lindsay, B. G. (1997). Weighted likelihood estimating equations: the discrete case with applications to logistic regression. *J. Statist. Plann. Inference* **57**, 215-232.

Markatou, M., Basu, A. and Lindsay, B. G. (1998). Weighted likelihood estimating equations with a bootstrap root search. *J. Amer. Statist. Assoc.* **93**, 740-750.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika Ser. A* **23**, 114-133.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics* **5**, 375-383.

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

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79**, 871-880.

Schrader, R. M. and Hettamansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika* **67**, 93-101.

Simpson, D. G. (1989). Hellinger deviate tests: efficiency, breakdown points, and examples. *J. Amer. Statist. Assoc.* **84**, 107-113.

Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *J. Amer. Statist. Assoc.* **87**, 439-450.

Tsou, T. S. and Royall, R. M. (1995). Robust likelihoods. *J. Amer. Statist. Assoc.* **90**, 316-320.

Department of Statistics, University of Padua, 35121 Padua, Italy.
E-mail: claudio@stat.unipd.it

Department of Statistics, Columbia University, New York, N.Y. 10027, U.S.A.
E-mail: markat@stat.columbia.edu

(Received April 1999; accepted August 2000)