Robust Hypothesis Tests for M-Estimators with Possibly Non-differentiable Estimating Functions†

Wei-Ming Lee
Department of Economics
National Chung Cheng University

Yu-Chin Hsu∗
Institute of Economics
Academia Sinica

Chung-Ming Kuan
Department of Finance
National Taiwan University

This version: October 5, 2014

† A previous version was circulated under the title "Robust Tests of Hypotheses in Models with M-Estimation."

* Corresponding author. E-mail: ychsu@econ.sinica.edu.tw. 128 Academia Road, Section 2, Nankang, Taipei, 115 Taiwan. Tel: 886-2-27822791 ext. 322. Fax: 886-2-2785-3946.

Acknowledgement: We thank Editor Michael Jansson, two anonymous referees, Anil Bera, Joon Park, Werner Ploberger, Jeffrey Racine, and the participants at the 2006 Far Eastern Meeting of the Econometric Society in Beijing for helpful suggestions and comments. All errors are our responsibility. The research support from the National Science Council of the Republic of China (NSC94-2415-H-194-009 for Lee) is also gratefully acknowledged.
Abstract

We propose a new robust hypothesis test for (possibly nonlinear) constraints on M-estimators with possibly non-differentiable estimating functions. The proposed test employs a random normalizing matrix computed from recursive M-estimators to eliminate the nuisance parameters arising from the asymptotic covariance matrix. It does not require consistent estimation of any nuisance parameters, in contrast with the conventional heteroskedasticity autocorrelation consistent (HAC)-type test and the KVB-type test of Kiefer, Vogelsang, and Bunzel (2000). Our test reduces to the KVB-type test in simple location models with OLS estimation, so the error in rejection probability of our test in a Gaussian location model is $O_p(T^{-1} \log T)$. We discuss robust testing in quantile regression, and censored regression models in details. In simulation studies, we find that our test has better size control and better finite sample power than the HAC-type and KVB-type tests.

JEL classification: C12, C22

Keywords: censored regression, generalized method of moments, robust hypothesis testing, KVB approach, M-estimator, quantile regression,
1 Introduction

Conventional hypothesis testing rests on consistent estimation of the asymptotic covariance matrix. In time series econometrics, the nonparametric kernel estimator originating from spectral estimation of Priestley (1981) is a leading example; also see Newey and West (1987), Andrews (1991), Newey and West (1994), and den Haan and Levin (1997) for econometric contributions. This estimator, which is also known as a heteroskedasticity-autocorrelation consistent (HAC) estimator, leads to asymptotic chi-squared tests that are robust to heteroskedasticity and serial correlations of unknown form, but the testing results can vary with the choices of the kernel function and its bandwidth.

In view of this, Kiefer, Vogelsang, and Bunzel (2000, KVB hereafter) propose to replace the HAC estimator with a random normalizing matrix to avoid the selection of the bandwidth in the nonparametric kernel estimation in linear regression models. This approach is extended to robust testing in nonlinear regression and generalized method of moments (GMM) models; see Bunzel, Kiefer, and Vogelsang (2001) and Vogelsang (2003) for more details. As for specification testing, Lobato (2001) develops a portmanteau test for serial correlations, and Kuan and Lee (2006) propose general M tests of moment conditions that are robust not only in the KVB sense but also to the presence of an estimation effect. For robust over-identifying restrictions (OIR) tests, please see Sun and Kim (2012) and Lee, Kuan, and Hsu (2014).

As we will see later, to test for (possibly nonlinear) constraints on the class of M-estimators of Huber (1967), a consistent estimator for the derivative of the expectation of the estimating function is needed. When the estimating function is differentiable with respect to the parameter vector, a consistent estimator for this is simply the sample average of the derivative of the estimating function. However, when the estimating function is not differentiable, the estimation is less straightforward; a leading example is the quantile regression (QR) estimator of Koenker and Basset (1978). Although an explicit form of the derivative of the expectation of the estimating function is available in this case, the conditional density of model errors is in the expression. Therefore, a consistent estimator for this matrix involves nonparametric kernel estimation of the conditional density. Therefore, user-chosen bandwidth is needed and the performance of the HAC-type and KVB-type tests can be sensitive to this choice. On the other hand, one may appeal to the bootstrap method to circumvent consistent estimation of any nuisance matrix, see, e.g., Buchinsky (1995) and Fitzenberger (1997), but the resulting tests are computationally demanding. Moreover, tests based on the moving
blocks bootstrap (MBB) as suggested by Fitzenberger (1997) can be sensitive to the
selection of block length and the number of bootstrap samples. Subsampling may also
be applied. It suffers from a problem similar to MBB, however.

In this paper, we propose a new robust hypothesis test for possibly nonlinear con-
straints on M-estimators with possibly non-differentiable estimating functions. The pro-
posed test employs a normalizing matrix computed from recursive M-estimators to elimi-
nate the nuisance parameters in the limit and hence does not require consistent estimation
of any nuisance parameters, in contrast with the HAC-type and KVB-type tests. This
feature makes the proposed test appealing because consistent estimators for such nuis-
ance parameters may not only be difficult to obtain but also sensitive to user-chosen
parameters which in turn lead to poor finite sample performance of the test. The null
limit of the proposed test is shown to be the same as that of Lobato (2001) and hence the
asymptotic critical values are already available. Moreover, we show that the proposed
test reduces to the KVB-type test in simple location models with OLS estimation so the
error in rejection probability of the proposed test in a Gaussian location model is thus
$O_p(T^{-1} \log T)$, in contrast with the conventional tests for which the error in rejection
probability is typically no better than $O_p(T^{-1/2})$. We consider robust testing in QR and
censored regression models in details. In simulation studies, we find that our test can
have better size control and better finite sample power than the HAC-type and KVB-type
tests.

Our method is also known as the self-normalization method in the statistics litera-
ture.\footnote{This line of research started after the first version of our paper which was presented in a Econometric
Society conference in 2006. We thank a referee for bring the self-normalization literature to our attention.} Note that, while Shao (2010) constructs confidence intervals for the parameters
that are functionals of the marginal or joint distribution of stationary time series (e.g.
mean or normalized spectral means), we consider hypothesis testing on parameters de-
defined in a more general class of econometric models that include these parameters as
special cases. Shao (2012) later constructs confidence intervals for the parameters in
stationary time series models based on the frequency domain maximum likelihood esti-
mator that can be applied to a large class of long/short memory time series models with
weakly dependent innovations; see also Zhou and Shao (2013) and Huang, Volgushev and
Shao (2014) for the inference for parameters in different context.

This paper proceeds as follows. In Section 2, we introduce M-estimation and related
asymptotic results. The proposed test as well as the HAC- and KVB-type tests are
then presented in Section 3. As examples, robust testing in QR, and censored regression
models are discussed in Section 4. Simulation results are reported in Section 5 and Section 6 concludes this paper. All proofs are deferred to the Appendix.

2 M-Estimation and Asymptotic Results

Let \( \{z_t\} \) be a sequence of \( k \times 1 \) random vectors and \( \theta \) be a \( p \times 1 \) unknown parameter vector with the true value \( \theta_o \). In the context of M-estimation, an M-estimator \( \hat{\theta}_T \) for \( \theta_o \) can be defined as the one satisfying the following estimating equation:

\[
\sum_{t=1}^{T} \phi(z_t; \hat{\theta}_T) = o_P(\sqrt{T}),
\]

where \( T \) is the sample size and \( \phi \) is a measurable and separable \( \mathbb{R}^p \)-valued estimating function with \( \mathbb{E}[\phi(z_t; \theta_o)] = 0 \). Clearly, the class of M-estimators includes many estimators as special cases. For example, consider the linear regression model:

\[
y_t = x_t' \theta + \epsilon_t(\theta).
\]

The OLS estimator \( \hat{\theta}_T \) satisfies \( \sum_{t=1}^{T} x_t (y_t - x_t' \hat{\theta}_T) = 0 \), the asymmetric least squares estimator \( \hat{\theta}_{\tau,T} \) of Newey and Powell (1987) and Kuan, Yeh, and Hsu (2009) for the \( \tau \)-th conditional expectile of \( y_t \) satisfies \( \sum_{t=1}^{T} x_t \cdot |\tau - I(y_t < x_t' \hat{\theta}_{\tau,T})| \cdot (y_t - x_t' \hat{\theta}_{\tau,T}) = 0 \), where \( I \) is the indicator function, and the QR estimator \( \hat{\theta}_{\tau,T} \) for the \( \tau \)-th conditional quantile of \( y_t \) is such that

\[
\sum_{t=1}^{T} x_t \text{sgn}_\tau \left( y_t - x_t' \hat{\theta}_{\tau,T} \right) = o_P(\sqrt{T}),
\]

where \( \text{sgn}_\tau(z) = \tau I(z > 0) - (1 - \tau) I(z < 0) \); see Fitzenberger (1997) for more details. The symmetrically trimmed least squares estimator \( \hat{\theta}_T \) proposed by Powell (1986b) is also an M-estimator because, for truncated data, it satisfies

\[
\sum_{t=1}^{T} x_t I(y_t < 2x_t' \hat{\theta}_T)(y_t - x_t' \hat{\theta}_T) = o_P(\sqrt{T}).
\]

See also Subsection 4.2 for the censored regression models. In what follows, \([c]\) denotes the integer part of the real number \( c \), \( \| \cdot \| \) denotes the sup norm for a vector, \( \xrightarrow{P} \) denotes convergence in probability, \( \xrightarrow{D} \) denotes convergence in distribution, \( =d \) denotes equality in distribution, and \( \Rightarrow \) denotes weak convergence of associated probability measures. We also let \( W_q \) be a vector of \( q \) independent, standard Wiener processes and \( B_q \) be the Brownian bridge with \( B_q(r) = W_q(r) - rW_q(1) \) for \( 0 \leq r \leq 1 \).
2.1 Asymptotic Normality of M-Estimators

Define

\[ m_{rT}(\theta) = \frac{1}{T} \sum_{t=1}^{rT} \phi(z_t; \theta), \]

where \(0 < r \leq 1\); for \(r = 1\), we simply write \(m_T(\theta)\), which is the sample average of \(\phi(z_t; \theta)\). To derive the limiting distribution of \(\hat{\theta}_T\), we impose the following “high-level” conditions that \(\hat{\theta}_T\) is \(\sqrt{T}\) consistent and \(\{\phi(z_t; \theta_0)\}\) obeys a multivariate functional central limit theorem (FCLT).

[A1] \( \sqrt{T}(\hat{\theta}_T - \theta_0) = O_p(1) \).

[A2] \( \sqrt{T}m_{rT}(\theta_0) \Rightarrow SW_p(r) \), where \(0 < r \leq 1\) and \(S\) is the matrix square root of \(V = \lim_{T \to \infty} \text{var}(T^{1/2}m_T(\theta_0))\); i.e., \(SS' = V\).

[A1] holds under various sets of primitive regularity conditions. These conditions typically require that some stochastic properties (such as memory, heterogeneity, and moment restrictions) of \(\{z_t\}\) and some smoothness and domination conditions on \(\phi\) so that \(\{\phi(z_t; \theta)\}\) obeys a weak uniform law of large numbers, \(\mathbb{E}[\phi(z_t; \theta)]\) is continuous in \(\theta\), and that \(\theta_0\) is identifiably unique. The discussion for consistency of M-estimation can be found, for example, in Huber (1981, Chapter 6), Tauchen (1985, pp. 422–424), and White (1994, pp. 33–35). [A2] is more than enough to establish the asymptotic normality of M-estimators but is required to derive the weak limit of recursive M-estimators. Note that the conditions that ensure multivariate FCLT are sufficient for [A2], e.g., Corollary 4.2 of Wooldridge and White (1988) or Theorem 7.30 of White (2001). By [A2], \( \sqrt{T}m_{rT}(\theta_0) \Rightarrow SW_p(1) \overset{d}{=} \mathcal{N}(0, V) \).

When \(\phi\) is twice continuously differentiable with respect to \(\theta\) and the corresponding second derivative is bounded in probability, the limiting distribution of \(\hat{\theta}_T\) can be easily derived using the first-order Taylor expansion. By expanding the estimating equation (1) about \(\theta_0\), we immediately have under [A1] that

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \phi(z_t; \hat{\theta}_T) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \phi(z_t; \theta_0) + \left[ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla_{\theta} \phi(z_t; \theta_0) \right] \sqrt{T}(\hat{\theta}_T - \theta_0) + o_p(1). \]
As long as a law of large numbers can be applied to the sample averages of \( \{ \nabla_{\theta} \phi(z_t; \theta_o) \} \), a Bahadur representation for \( \hat{\theta}_T \) can then be given by

\[
\sqrt{T}(\hat{\theta}_T - \theta_o) = -M_o^{-1}\sqrt{T}m_T(\theta_o) + o_{\mathbb{P}}(1),
\]

where \( M_o = \lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} \mathbb{E}[\nabla_{\theta} \phi(z_t; \theta_o)] \), a \( p \times p \) nonsingular matrix of constants. Under [A2], \( T^{1/2}(\hat{\theta}_T - \theta_o) \overset{D}{\to} \mathcal{N}(0, \Sigma_o) \), where \( \Sigma_o = M_o^{-1}V M_o^{-1} \).

As many M-estimators such as the LAD and QR estimators rely on non-differentiable \( \phi \), it is not capable of expanding (1) around \( \theta_o \) using the technique above, but a Bahadur representation as in (2) is still available for such M-estimators. To see this, define

\[
\lambda_{[rT]}(\theta) = \mathbb{E}_{[\theta]}[m_{[rT]}(\theta)] = \frac{1}{T} \sum_{t=1}^{[rT]} \mathbb{E}[\phi(z_t; \theta)],
\]

and \( M_{[rT]}(\theta) = (T/[rT]) \nabla_{\theta} \lambda_{[rT]}(\theta) \); for \( r = 1 \), we simply write \( \lambda_T(\theta) \) and \( M_T(\theta) \). We now impose the following two conditions; see also Weiss (1991) for primitive conditions.

[A3] \( \lambda_T(\theta) \) is twice continuously differentiable with a bounded second derivative and satisfies the following property:

\[
\sqrt{T} \lambda_T(\hat{\theta}_T) + \sqrt{T} m_T(\theta_o) = o_{\mathbb{P}}(1).
\]

[A4] \( M_{[rT]}(\theta_o) \to M_o = \lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} \nabla_{\theta} \mathbb{E}[\phi(z_t; \theta_o)] \) uniformly in \( r \in (0, 1] \) and \( M_o \) is non-singular.

Given these two conditions and by a first-order Taylor expansion of \( \sqrt{T}\lambda_T(\hat{\theta}_T) \) around \( \theta_o \), we obtain an expression as in (2) and hence the asymptotic normality for such M-estimators follows immediately from [A2].

### 2.2 Weak Convergence of Recursive M-Estimators

Let \( \hat{\theta}_j \) be the recursive M-estimator using the subsample of first \( j \) observations; i.e., \( \hat{\theta}_j \) is the one such that

\[
\sum_{t=1}^{j} \phi(z_t; \hat{\theta}_j) = o_{\mathbb{P}}\left(\sqrt{j}\right).
\]

To derive its weak limit, we modify [A1] and [A3] to:

[B1] \( \sqrt{[rT]}(\hat{\theta}_{[rT]} - \theta_o) = O_{\mathbb{P}}(1) \), uniformly in \( r \in (0, 1] \).
\[ \lambda_{[rT]}(\theta) \text{ is twice continuously differentiable with a bounded second derivative and satisfies} \]
\[ \sqrt{T} \lambda_{[rT]}(\hat{\theta}_{[rT]}) + \sqrt{T} m_{[rT]}(\theta_o) = o_\mathbb{P}(1), \]
ununiformly in \( r \in (0, 1] \).

These two conditions should not be abrupt as the recursive M-estimators and the full-sample M-estimator ought to behave similarly. It can be shown that
\[ \frac{[rT]}{\sqrt{T}} \left( \hat{\theta}_{[rT]} - \theta_o \right) = -M_o^{-1} \sqrt{T} m_{[rT]}(\theta_o) + o_\mathbb{P}(1), \]
which implies that under the condition [A2],
\[ \frac{[rT]}{\sqrt{T}} \left( \hat{\theta}_{[rT]} - \theta_o \right) \Rightarrow \Lambda W_p(r), \quad 0 < r \leq 1, \]
where \( \Lambda = -M_o^{-1} S \). By replacing \( \theta_o \) with the full-sample M-estimator \( \hat{\theta}_T \), we can apply the continuous mapping theorem to obtain that
\[ \frac{[rT]}{\sqrt{T}} \left( \hat{\theta}_{[rT]} - \hat{\theta}_T \right) = \frac{[rT]}{\sqrt{T}} \left( \hat{\theta}_{[rT]} - \theta_o \right) - \frac{[rT]}{T} \frac{T}{\sqrt{T}} \left( \hat{\theta}_T - \theta_o \right) \Rightarrow \Lambda B_p(r). \quad (3) \]
This result has been established in the literature on testing parameter constancy for linear mean and median regressions; see, e.g., Ploberger, Krämer, and Kontrus (1989), Kuan and Hornik (1995), and Chen and Kuan (2001). Here, we show that it remains valid for recursive M-estimators. This result is needed to construct robust tests for parameter restrictions.

### 3 Tests for General Hypotheses

The null hypothesis of interest consists of \( q \leq p \) possibly nonlinear restrictions on \( \theta_o \):
\[ H_o : \gamma(\theta_o) = 0, \quad (4) \]
where \( \gamma \) is a \( q \times 1 \) vector of continuously differentiable functions with \( R(\theta) = \nabla_\theta \gamma(\theta) \) that is of full row rank in a neighborhood of \( \theta_o \). Under the conditions above and by a delta method, one can show that \( \sqrt{T} (\gamma(\hat{\theta}_T) - \gamma(\theta_o)) \overset{D}{\longrightarrow} \mathcal{N}(0, \Omega_o) \), where \( \Omega_o = R(\theta_o) M_o^{-1} V M_o^{-1} R(\theta_o)' \).
3.1 The HAC-Type Test

To test the hypothesis defined in (4), the test statistic for a HAC-type test is defined as

\[ J_{\text{HAC}} = T^{-\frac{1}{2}} \hat{\Theta}_T \gamma(\hat{\theta}_T)' \left( \sum_{t=1}^{T} I(t) \right)^{-\frac{1}{2}} R(\hat{\theta}_T) \tilde{V}_{\text{HAC}} \tilde{M}_T^{-\frac{1}{2}} R(\hat{\theta}_T)', \]

where \( \hat{\Theta}_T \) is the HAC estimator for the asymptotic covariance matrix of \( \sqrt{T} \gamma(\hat{\theta}_T) \). The null distribution of \( J_{\text{HAC}} \) is a \( \chi^2(q) \) distribution. In this paper, a test based on test statistic \( J_{\text{HAC}} \) will be called the \( J_{\text{HAC}} \) test.

Although the \( J_{\text{HAC}} \) test is consistent for \( \Omega_o \) and \( \tilde{V}_{\text{HAC}} \) is a nonparametric kernel estimator of \( V \) given by

\[ \tilde{V}_{\text{HAC}} = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{T} \kappa \left( \frac{|i-j|}{\ell(T)} \right) \phi(z_i; \hat{\theta}_T) \phi(z_j; \hat{\theta}_T)', \]

where \( \kappa \) is a kernel function and \( \ell(T) \) denotes the truncation lag or bandwidth. As \( \sqrt{T} (\gamma(\hat{\theta}_T) - \gamma(\theta_o)) \overset{\text{D}}{\to} N(0, \Omega_o) \), where \( \Omega_o = R(\theta_o)M_o^{-1} VM_o^{-1} R(\theta_o)' \), and \( \hat{\Theta}_T \) is consistent for \( \Omega_o \). The null distribution of \( J_{\text{HAC}} \) is a \( \chi^2(q) \) distribution. In this paper, a test based on test statistic \( J_{\text{HAC}} \) will be called the \( J_{\text{HAC}} \) test.

Although the \( J_{\text{HAC}} \) test has a standard null limit and is robust to heteroskedasticity and serial correlations of unknown form, its finite sample performance will depend on choices of the kernel function and the truncation lag. To implement the \( J_{\text{HAC}} \) test, it also requires consistent estimator for \( M_o \). If \( \phi \) is continuously differentiable, then

\[ M_o = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \nabla_{\theta} \phi(z_t; \theta_o) \right], \]

and the sample average of \( \nabla_{\theta} \phi(z_t; \hat{\theta}_T) \) is a natural consistent estimator for \( M_o \). However, when \( \phi \) is not differentiable, the estimation is less straightforward. For example, in the QR regression models, the \( M_o \) involves the conditional density of the error terms. Although kernel-based estimators for \( M_o \) are available (cf. Weiss 1991), the resulting \( J_{\text{HAC}} \) test would suffer from the problems arising from nonparametric kernel estimation. For more details and more examples, please see Section 4. In addition, if the constraints are nonlinear, we will need to estimate \( R(\theta_o) \) by \( R(\hat{\theta}_T) \). Even it is straightforward to obtain \( R(\hat{\theta}_T) \), but the finite sample performance of the tests may be adversely affected by this estimator.

3.2 The KVB-Type Test

The main idea underlying the KVB approach is to employ a random normalizing matrix in place of the kernel-based covariance matrix estimator \( \tilde{V}_{\text{HAC}} \) to avoid the problems arising from nonparametric kernel estimation. Following this approach, we construct
\[ V_{KVB} \text{ as} \]

\[ V_{KVB} = \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{\sqrt{T}} \sum_{i=1}^{t} \phi(z_i; \hat{\theta}_T) \right) \left( \frac{1}{\sqrt{T}} \sum_{i=1}^{t} \phi(z_i; \hat{\theta}_T) \right)' \]

\[ = \frac{1}{T} \sum_{t=1}^{T} \left( \sqrt{T} m_t(\hat{\theta}_T) \right) \left( \sqrt{T} m_t(\hat{\theta}_T) \right)' . \]

Clearly, it reduces to that of Kiefer et al. (2000) when \( \phi(z_t; \theta) = x_t(y_t - x'_t \theta) \) with \( z_t = [y_t \ x'_t]' \). Given \( V_{KVB} \), a KVB-type test statistic is defined as

\[ J_{KVB} = T \gamma(\hat{\theta}_T)' \tilde{\Omega}_{KVB}^{-1} \gamma(\hat{\theta}_T), \]

where \( \tilde{\Omega}_{KVB} = R(\hat{\theta}_T) \tilde{\Sigma}_{KVB} R(\hat{\theta}_T)' \) with \( \tilde{\Sigma}_{KVB} = \tilde{M}_T^{-1} V_{KVB} (\tilde{M}_T^{-1})' \).

It is easy to derive the weak limit of \( \sqrt{T} m_{[rT]}(\hat{\theta}_T) \) (and \( J_{KVB} \)) when \( \phi \) is smooth (cf. Kuan and Lee 2006), and it is less straightforward when \( \phi \) is non-smooth. To allow for non-smooth \( \phi \), we impose the following condition, which is similar to the condition (v) of Theorem 7.2 in Newey and McFadden (1994).

\[ \text{[A5] Define} \]

\[ \zeta_{[rT]}(\eta, \theta) = \frac{\sum_{i=1}^{[rT]} \left( \phi(z_i; \eta) - \phi(z_i; \theta) - \frac{T}{rT} \lambda_{[rT]}(\eta) + \frac{T}{rT} \lambda_{[rT]}(\theta) \right)}{\sqrt{[rT]} + |rT| \| \lambda_{[rT]}(\eta) \|}. \]

Then there exists a \( d_o > 0 \) such that \( \sup_{\| \eta - \theta_o \| \leq d_o, r \in (0,1]} \zeta_{[rT]}(\eta, \theta_o) \overset{P}{\to} 0. \)

This condition should not be restrictive because sufficient conditions for [B3] are also sufficient for [A5]; see e.g., Huber (1967, p. 227) and Weiss (1991, p. 62).

\[ \text{Lemma 3.1 Suppose that conditions [A1], [A2], [B3], [A4], and [A5] hold. Then } V_{KVB} \Rightarrow SP_p S', \text{ where } P_p = \int_0^1 B_p(r) B_p(r)' \, dr \text{ and } S \text{ is the matrix square root of } V. \]

As shown in Lemma 3.1, \( V_{KVB} \) remains random in the limit so the null distribution of \( J_{KVB} \) will not be a \( \chi^2(q) \) distribution, but the following theorem shows that the \( J_{KVB} \) is asymptotically pivotal under the null and its weak limit is the same as that of Lobato (2001). Therefore, the corresponding critical values can be found in Lobato (2001, p. 1067).
Theorem 3.2 Suppose that conditions \([A1], [A2], [B3], [A4],\) and \([A5]\) hold. Then under the null,

\[ \mathcal{J}_{KV_B} \overset{D}{\to} W_q(1)'P_q^{-1}W_q(1). \]

Although the \(\mathcal{J}_{KV_B}\) test does not require consistent estimation of \(V\), it still requires consistent estimation of \(M_o\) and \(R(\theta_o)\). Especially, as we discuss before, it can be hard to obtain \(M_o\) and there might be additional smoothing parameter to pick when \(\phi\) is not differentiable. In view of this, in the next subsection, we propose a new way to construct robust tests without consistent estimation of \(M_o\) and \(R(\theta_o)\).

3.3 The Proposed Test

We propose to use the following normalizing matrix:

\[ \widehat{\Omega}_{RE} = \frac{1}{T} \sum_{j=p}^{T} \left( \frac{j}{\sqrt{T}} (\gamma(\hat{\theta}_j) - \gamma(\hat{\theta}_T)) \right) \left( \frac{j}{\sqrt{T}} (\gamma(\hat{\theta}_j) - \gamma(\hat{\theta}_T)) \right)' , \]

which is computed using only recursive estimators, and define the test statistic of our test as \(\mathcal{J}_{RE} = T\gamma(\hat{\theta}_T)'\widehat{\Omega}_{RE}^{-1}\gamma(\hat{\theta}_T)\). Note that we do not need to estimate \(M_o\) and \(R(\theta_o)\) to obtain \(\widehat{\Omega}_{RE}\) and this makes our test appealing. Under the assumptions above, we can show that

\[ \frac{[rT]}{\sqrt{T}} \left( \gamma(\hat{\theta}_{[rT]}) - \gamma(\theta_o) \right) = R(\theta_o) \frac{[rT]}{\sqrt{T}} (\hat{\theta}_{[rT]} - \theta_o) + o_P(1) \]

\[ \Rightarrow R(\theta_o)\Lambda W_p(r) \overset{d}{=} \Xi W_q(r), \]

where \(\Xi\) is the matrix square root of \(\Omega_o\) and \(\Omega_o\) is the asymptotic variance of \(\sqrt{T}(\hat{\gamma} - \gamma(\theta_0))\). It follows that

\[ \frac{[rT]}{\sqrt{T}} \left( \gamma(\hat{\theta}_{[rT]}) - \gamma(\theta_T) \right) \Rightarrow \Xi B_q(r). \]

Then (6) and (7) are sufficient to derive the null distribution of \(\mathcal{J}_{RE}\) which is summarized in the following theorem.

Theorem 3.3 Suppose that the conditions \([B1], [A2], [B3],\) and \([A4]\) hold. Then \(\widehat{\Omega}_{RE} \overset{D}{\to} \Xi P_q\Xi'\). Moreover,

\[ \mathcal{J}_{RE} \overset{D}{\to} W_q(1)'P_q^{-1}W_q(1), \]
under the null hypothesis that $\gamma(\theta_o) = 0$.

Remarks:

1. In (5) the summation starts with $j = p$ where $p$ is the number of unknown parameters. When the criterion function is not smooth, it may not be easy to compute the M estimators. In addition, if the sample size is small, the optimization might not be stable and the global minimization or maximization may not be easy to guarantee. Therefore, one might want to start with a larger subsample size to avoid these problems. In fact, the theory would still hold as long as the the summation starts with $b_T \geq p$ and the sequence $b_T$ satisfies $b_T/T \to 0$ in that if $b_T/T \to 0$, then

$$
\frac{1}{T} \sum_{j=b_T}^{T} \left( \frac{j}{\sqrt{T}} (\gamma(\hat{\theta}_j) - \gamma(\hat{\theta}_T)) \right) \left( \frac{j}{\sqrt{T}} (\gamma(\hat{\theta}_j) - \gamma(\hat{\theta}_T)) \right)' \Rightarrow \Xi P_q \Xi'
$$

as in Theorem 3.3.

2. To shed more insight on the difference and similarity between $J_{RE}$ and $J_{KVB}$, we consider the linear model: $y_t = x_t^\prime \theta + \epsilon_t(\theta)$ and the null hypothesis (4) with $\gamma(\theta_o) = R\theta_o - r$. Based on the OLS estimator $\hat{\theta}_T$, the KVB normalizing matrix can be expressed as

$$
\hat{\Omega}_{KVB} = R \left\{ \frac{1}{T} \sum_{j=1}^{T} \left[ \hat{M}_T^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{j} x_t \hat{e}_t \right) \right] \left[ \hat{M}_T^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{j} x_t \hat{e}_t \right) \right]' \right\} R',
$$

where $\hat{M}_T = -T^{-1} \sum_{t=1}^{T} x_t x_t'$ and $\hat{e}_t = y_t - x_t^\prime \hat{\theta}_T$. As for $\hat{\Omega}_{RE}$, we first show that for $j \geq p$,

$$
\hat{\theta}_j - \hat{\theta}_T = \left( \sum_{t=1}^{j} x_t x_t' \right)^{-1} \left[ \sum_{t=1}^{j} x_t y_t - \left( \sum_{t=1}^{j} x_t x_t' \right) \hat{\theta}_T \right] = \left( \sum_{t=1}^{j} x_t x_t' \right)^{-1} \sum_{t=1}^{j} x_t \hat{e}_t.
$$

Then the proposed normalizing matrix becomes

$$
\hat{\Omega}_{RE} = R \left\{ \frac{1}{T} \sum_{j=p}^{T} \left[ \frac{j}{\sqrt{T}} \left( \sum_{t=1}^{j} x_t x_t' \right)^{-1} \left( \sum_{t=1}^{j} x_t \hat{e}_t \right) \right] \left[ \frac{j}{\sqrt{T}} \left( \sum_{t=1}^{j} x_t x_t' \right)^{-1} \left( \sum_{t=1}^{j} x_t \hat{e}_t \right) \right]' \right\} R'
$$

$$
= R \left\{ \frac{1}{T} \sum_{j=p}^{T} \hat{M}_j^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{j} x_t \hat{e}_t \right) \right[ \hat{M}_j^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{j} x_t \hat{e}_t \right) \right]' \right\} R',
$$

\(^2\)We thank a referee for pointing this out.
where $\hat{M}_j = -j^{-1} \sum_{t=1}^{j} x_t x'_t$. The result above shows that $J_{RE}$ and $J_{KV B}$ employ different but similar normalizing matrices in the context of linear regression with OLS estimation. Of particular interest is that in a simple location model (i.e., $p = 1$ and $x_t = 1$ for all $t$), these two tests are algebraically equivalent because $\hat{\Omega}_{RE} = \hat{\Omega}_{KV B}$ in this case. Therefore, the error in rejection probability of the proposed $J_{RE}$ test is also $O_p \left( T^{-1} \log T \right)$ in a Gaussian location model as shown in Jansson (2004).

### 3.4 Asymptotic Local Power

In this section we compare the local powers of the $J_{RE}$, $J_{KV B}$, and $J_{HAC}$ tests under local alternatives defined as follows:

$$H_1 : \gamma(\theta_o) = \delta_o / \sqrt{T}, \quad (8)$$

where $\delta_o$ is a $q \times 1$ vector of nonzero constants. Under the local alternatives, the limiting distributions of these tests are summarized in the following theorem. Let $U_q = \left[ \Xi^{-1} \delta_o + W_q(1) \right] P_q^{-1} \left[ \Xi^{-1} \delta_o + W_q(1) \right]'$.

**Theorem 3.4** Suppose that the conditions [B1], [A2], [B3], [A4], and [A5] hold. Then under the local alternatives defined in (8), $J_{HAC} \xrightarrow{D} \chi^2(q; \delta_o' \Xi^{-1} \delta_o)$, $J_{KV B} \xrightarrow{D} U_q$, and $J_{RE} \xrightarrow{D} U_q$.

Given Theorem 3.4, we can derive the asymptotic local power for these tests now. Let $\chi^2_{q, \alpha}$ and $c_{q, \alpha}$ be the critical values at $\alpha$ significance level taken from $\chi^2(q)$ and $W_q(1)' P_q^{-1} W_q(1)$, respectively. Also let $\text{ALP}_{q, \alpha}^{HAC}$, $\text{ALP}_{q, \alpha}^{KV B}$, and $\text{ALP}_{q, \alpha}^{RE}$ be the asymptotic local powers of the $J_{HAC}$, $J_{KV B}$, and $J_{RE}$ tests, respectively.

**Corollary 3.5** Suppose that the conditions [B1], [A2], [B3], [A4], and [A5] hold. Then under the local alternatives, $\text{ALP}_{q, \alpha}^{HAC} = \mathbb{P}(\chi^2(q; \delta_o' \Xi^{-1} \delta_o) > \chi^2_{q, \alpha})$ and $\text{ALP}_{q, \alpha}^{KV B} = \text{ALP}_{q, \alpha}^{RE} = \mathbb{P}(U_q > c_{q, \alpha})$.

It is obvious that $J_{KV B}$ and $J_{RE}$ have the same asymptotic local power because their limiting distribution are identical. The local power curves of these tests are the same as those in Figure 1 of Kuan and Lee (2006), so we omit them here. As we can see from Figure 1 of Kuan and Lee (2006), $J_{HAC}$ has better local power than the other two tests. However, $J_{HAC}$ may not outperform the other two tests in finite samples because it performance would depend on user-chosen parameters as well.
4 Examples

In this section, we illustrate the application of the proposed $J_{RE}$ test in QR, and censored regression models. QR is a leading example for a non-differentiable $\phi$. In the second example, whether the corresponding $\phi$ is differentiable or not and whether consistent estimation of $M_o$ is easy or not depend on the estimation method used.

4.1 QR Models

In the context of QR, the $\tau$-th conditional quantile of $y_t$ given a vector of explanatory variables $x_t$ is typically specified as

$$\inf\{y : F(y \mid x_t) \geq \tau\} = q_\tau(x_t; \theta),$$

where $F$ is the conditional distribution function of $y_t$ given $x_t$, $\theta_\tau \in \Theta_\tau$ is a $p \times 1$ vector of unknown parameters with the true value $\theta_{\tau,o}$, and $q_\tau$ is a real-valued function that is continuously differentiable in the neighborhood of $\theta_{\tau,o}$. Similar to the mean regression model, the QR model can also be written as

$$y_t = q_\tau(x_t; \theta) + e_t(\tau; \theta), \quad (9)$$

where $e_t(\tau; \theta)$ is the error term with the $\tau$-th conditional quantile being zero under the true value $\theta_{\tau,o}$, a condition ensuring correct specification for the model (9). Note that the model (9) with $\tau = 0.5$ is also known as a median regression model.

To estimate the unknown parameter vector $\theta_{\tau,o}$, Koenker and Bassett (1978) propose the QR estimator $\hat{\theta}_{\tau,T}$ that minimizes the following criterion function:

$$Q_\tau(\theta) = \frac{1}{T} \sum_{t=1}^{T} \rho_\tau(e_t(\tau; \theta)), $$

where $\rho_\tau(z) = z[\tau - I(z < 0)]$ is known as a check function. An analytic form for $\hat{\theta}_{\tau,T}$ is generally not available; nonetheless, many algorithms for obtaining $\hat{\theta}_{\tau,T}$ have been proposed in the literature; see e.g. Koenker and d’Orey (1987) and Koenker and Park (1996). Note also that the LAD estimator corresponds to the QR estimator with $\tau = 0.5$.

Note that $Q_\tau$ is not differentiable, so the limiting distribution of $\hat{\theta}_{\tau,T}$ is relatively difficult to derive. Nonetheless, under suitable conditions, the asymptotic normality result...
remains valid for the QR estimator (or LAD estimator); see Powell (1986a), Weiss (1991), and Fitzenberger (1997), among many others. Consider the linear quantile regression; i.e., \( q_t(x_t; \theta) = x_t' \theta + \epsilon_t \). The linear QR estimator \( \hat{\theta}_{\tau,T} \) is an M-estimator satisfying \( \sum_{t=1}^T x_t, \text{sgn}_t(y_t - x_t' \hat{\theta}_{\tau,T}) = o_p(T^{1/2}) \). For more general nonlinear models, we can follow Weiss (1991) and Fitzenberger (1997), and show that \( \sum_{t=1}^T \phi(y_t, x_t; \hat{\theta}_\tau) = o_p(T^{1/2}) \), where \( \phi(y_t, x_t; \theta) = \nabla_\theta q_t(x_t; \theta) \text{sgn}_t(y_t - q_t(x_t; \theta)) \). See also Koenker (2005, p. 124). Therefore, \( \hat{\theta}_{\tau,T} \) is an M-estimator with a non-differentiable \( \phi \).

To implement HAC-type and KVB-type tests in QR models, it is required to estimate \( M_o \) consistently and \( M_o \) is given as

\[
M_o = - \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}_t \left[ f_{\epsilon|x}(0) \nabla_\theta q_t(x_t; \theta_{\tau,o}) \nabla_\theta q_t(x_t; \theta_{\tau,o})' \right],
\]

where \( f_{\epsilon|x} \) is the conditional density of \( \epsilon_t = y_t - q_t(x_t; \theta_{\tau,o}) \). \( f_{\epsilon|x} \) can be estimated by a nonparametric kernel method, but the test performance will depend on the choices of kernel and bandwidth. To circumvent consistent estimation of \( M_o \), one may appeal to MBB or subsampling, but the testing results can be sensitive to the selection of block length instead. Therefore, the proposed \( J_{RE} \) tests seems to be practically useful for testing in QR models because it avoids consistent estimation of \( M_o \) and is thus free from user-chosen parameters.

### 4.2 Censored Regression Models

Let \( y^*_t \) be the dependent variable generated according to \( y^*_t = x_t' \theta_o + \epsilon_t \). If we only observe data points \( (y_t, x_t) \) with \( y_t = \max \{0, y^*_t \} \), then \( y_t = x_t' \theta + \epsilon_t(\theta) \) forms a censored regression model and the OLS estimator by regression \( y_t \) on \( x_t \) is not consistent for \( \theta_o \); see Amemiya (1984) for an early review. Most studies rely on maximum likelihood (ML) estimation. Let \( \epsilon_t|x_t \sim N(0, \sigma_o^2) \) and \( F^N_{\epsilon|x}(\cdot; \sigma_o^2) \) and \( f^N_{\epsilon|x}(\cdot; \sigma_o^2) \) be the corresponding conditional distribution and density functions. Under the assumption of independence, the log likelihood function can then be written as \( \mathcal{L}(\theta^*) = \sum_{t=1}^T g(z_t; \theta^*) \), where \( \theta^* = [\theta', \sigma^2]' \), \( z_t = [y_t, x_t']' \), and

\[
g(z_t; \theta^*) = [1 - I(y_t > 0)] \ln \left[ 1 - F^N_{\epsilon|x}(x_t' \theta; \sigma^2) \right] + I(y_t > 0) \ln f^N_{\epsilon|x}(y_t - x_t' \theta; \sigma^2).
\]

The Gaussian ML estimator \( \hat{\theta}^*_T \) is an M-estimator that solves \( \sum_{t=1}^T \nabla_{\theta^*} g(z_t; \hat{\theta}^*_T) = 0 \).

The Gaussian ML estimator is not robust to conditional heteroskedasticity and nonnormality, however. Powell (1984) propose censored LAD estimator instead which sat-
isfies (1) with non-differentiable \( \phi(z_t; \theta) = I(x_t' \theta > 0)[1/2 - I(y_t < x_t' \theta)]x_t \). Although such estimator is robust to conditional heteroskedasticity and non-normality, hypothesis tests based on this estimator are not easy to implement because

\[
M_o = - \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ I(x_t' \theta_o > 0) f_{\varepsilon|x}(0) x_t x_t' \right],
\]

where \( f_{\varepsilon|x} \) is the true conditional density of \( \varepsilon_t \). Therefore, the censored LAD-based test would suffer from the same problems as in QR models.

When the conditional distribution of \( y_t^* \) is symmetric about \( x_t' \theta_o \), the symmetrically trimmed least squares estimator introduced in Powell (1986b) is also consistent for \( \theta_o \) and robust to conditional heteroskedasticity and non-normality. Moreover, tests based on this estimator is easier to implement. To see this, from equation (2.9) in Powell (1986b), this estimator is an M-estimator with \( \phi(z_t; \theta) = I(x_t' \theta > 0)(\min\{y_t, 2x_t' \theta\} - x_t' \theta)x_t \). Although \( \phi \) is non-differentiable, the corresponding \( M_o \) can be expressed as

\[
M_o = - \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ I(-x_t' \theta_o < \varepsilon_t < x_t' \theta_o) x_t x_t' \right],
\]

for which consistent estimation is straightforward.

The censored regression model can be applied not only to cross sectional data but also to time series data. For time series data, the assumption of serial independence may not be appropriate, but Robinson (1982) show that the Gaussian ML estimator remains consistent and asymptotically normal with a complicated asymptotic covariance matrix. Therefore, the test developed in this paper can be useful too.

5 Monte Carlo Simulations

In this section, the finite sample performance of the proposed \( J_{RE} \) test is evaluated via Monte Carlo simulations. We consider three sample sizes: \( T = 50, 100, \) and \( 500 \), and two different nominal sizes: 5% and 10%. The number of replications is 5,000 for size simulations and 1,000 for power simulations. As the results for different nominal sizes are qualitatively similar, we report only the results for 5% nominal size.

As in Kiefer et al. (2000), we consider the linear regression specification: \( y_t = x_t' \theta + \varepsilon_t(\theta) \), where \( x_t \) is a \( 5 \times 1 \) vector of regressors with the first element equal to 1 for all \( t \) and other elements are mutually independent AR(1) processes. The last four elements are
generated in the same way as the AR(1)-HOMO errors introduced below. In the same simulation design, the correlation coefficients and the distributions of the regressors are the same as those of the error terms. $\theta$ is an unknown parameter vector with the true value $\theta_o$, and $e_t(\theta)$ is an error term. The null hypothesis is given by

$$H_o : \theta_{2,o} = \cdots = \theta_{q+1,o} = 0,$$

where $\theta_{i,o}$ is the $i$-th element of $\theta_o$ and $q$ denotes the number of restrictions on $\theta_o$. We consider $q = 1, 2, 3$ for size simulations and $q = 1$ for power simulations. To test these hypotheses, we apply the OLS and LAD methods for estimating $\theta_o$. It is well known that the OLS estimator enjoys optimality under (i.i.d.) normal errors, but the LAD estimator may be a better one for leptokurtic errors.

In size simulations, we set $\theta_o = 0$ and $e_t = e_t(\theta_o) = \sigma_t \eta_t$, where $\text{var}(\eta_t) = 1$ and $\sigma_t^2$ represents the conditional variance of $e_t$. The data generating processes (DGPs) for $e_t$ are AR(1)-HOMO and AR(1)-HET. AR(1) indicates that $\eta_t$ is governed by the AR(1) model: $\eta_t = \rho \eta_{t-1} + c u_t$, where $\rho$ is set to be either 0.5 or 0.8, $u_t$ is a white noise with unit variance, and $c = (1 - \rho^2)^{1/2}$, which is a scaling factor such that $\text{var}(\eta_t) = 1$. While HOMO stands for conditional homoskedasticity of $e_t$ (we set $\sigma_t^2 = 1$ for all $t$), HET denotes that $e_t$ is conditionally heteroskedastic with $\sigma_t = \min\{3, \max\{0.21, |x_{i2}|\}\}$ as considered in Fitzenberger (1997, p. 255). We generate $u_t$ from i.i.d. $\mathcal{N}(0, 1)$ and standardized Student’s $t(4)$. This enables us to examine if LAD-based tests are more appropriate for leptokurtic data. As for the regressors, they follow the AR(1) model specified as that for $\eta_t$.

For comparison, we simulate the HAC-type and KVB-type tests. For the former, we consider the Bartlett kernel covariance matrix estimator for which the truncation lag is determined by the nonparametric method of Newey and West (1994) with the weighting vector $w = [0 \ 1 \ 1 \ 1 \ 1]'$ and two preliminary truncation lags: $\ell(T) = [a(T/100)^{2/9}]$, where $a = 4$ and 12. Unlike the $J_{RE}$ test, these two tests require consistent estimation of $M_o$. Thus, we employ the estimator $\tilde{M}_T = -T^{-1} \sum_{t=1}^{T} x_t x'_t$ for OLS and the following kernel-based estimator for LAD:

$$\tilde{M}_T = -\frac{2}{Tb(T)} \sum_{t=1}^{T} \kappa\left(\hat{e}_t/b(T)\right)x_t x'_t,$$

where $\kappa$ is the Gaussian kernel, $b(T)$ is the bandwidth that vanishes in the limit, and $\hat{e}_t$ are LAD residuals. Although optimal selection of bandwidth has been extensively
discussed in the literature on density estimation, it is not clear how the value of \( b(T) \) should be selected such that \( \hat{M}_T \) enjoys optimality in some sense. We thus examine the effect of selecting \( b(T) \) by employing the two bandwidths: (i) \( b(T) = 0.9AT^{-1/5} \) and (ii) \( b(T) = 0.9AT^{-1/3} \), where \( A = \min\{\hat{\sigma}_T, R/1.34\} \) with \( \hat{\sigma}_T \) and \( R \) being, respectively, the sample standard deviation and interquartile range of the LAD residuals. The first one is suggested by Silverman (1986, p. 48) to obtain an optimal rate for density estimation, but the second one goes to zero at the same rate as in Koenker (2005, p. 81). Given these choices, the OLS-based tests read \( J_{NW,a}^{HAC} \) and \( J_{KV,B} \), and the LAD-based tests are denoted as \( J_{NW,a,j}^{HAC} \) and \( J_{KV,B}^j \), where \( j = (i) \) and \( (ii) \).

The empirical sizes for the OLS-based tests are reported in Tables 1 and 2. Clearly, these tests are all over-sized in small samples (e.g., \( T = 50 \)) and the distortions deteriorate when \( q \) or \( \rho \) gets larger. We also observe that leptokurtosis has little effect on the size performance, but heteroskedasticity does result in more size distortions especially for leptokurtic data and smaller \( q \). Among these tests, the \( J_{NW,a}^{HAC} \) test has largest size distortions, regardless of the values of \( a \). In particular, its size distortions are much larger for \( a = 12 \) and remains even when \( T = 500 \). The other tests are clearly less over-sized and a quite encouraging result is that the proposed \( J_{RE} \) test dominates the \( J_{KV,B} \) test in terms of finite sample size. It is also found that when data become more persistent (i.e., \( \rho \) gets larger), the size distortion of the former increases slightly, yet the size distortion of the latter increases dramatically.

The empirical sizes for LAD are summarized in Tables 3 and 4. Generally, the \( J_{RE} \) test is still the best one and the HAC-type test is the worst one. As compared with the preceding tables, we observe that the OLS-based and LAD-based tests have similar patterns regarding size performance. It is also found that the LAD-based \( J_{RE} \) test has empirical sizes closer to the nominal size 5% but not necessary for the other tests. These results suggest that, as far as accurate finite sample size is concerned, the LAD-based \( J_{RE} \) test is a preferred one for testing in linear regressions.

For power simulations, we consider \( T = 50 \) and 100 and the null hypothesis \( H_0 : \beta_{2,o} = 0 \) against the alternatives for which \( \beta_{2,o} \in (0, 2] \). The DGPs considered are AR(1)-HOMO and AR(1)-HET with \( \rho = 0.5 \) and two error terms: \( \mathcal{N}(0, 1) \) and (standardized) Student’s \( t \) errors. As shown in the preceding results, the \( J_{RE} \) test is slightly over-sized for these DGPs, but other tests have substantial size distortions, especially when \( T = 50 \) and inappropriate user-chosen parameters are used. To provide a proper power comparison, we thus simulate the size-adjusted powers. It should be, however, stressed that, while size adjustment enables us to compare the power performance of tests with different finite
sample sizes, it is generally infeasible in practical applications.

The power curves for the OLS-based and LAD-based tests are plotted in Figures 1 and 2, with $\beta_{2,o}$ on the horizontal axis. Clearly, their powers grow with $\beta_{2,o}$ and $T$, but are adversely affected by leptokurtosis or heteroskedasticity. Comparing the OLS-based $J_{\text{RE}}$ and $J_{\text{KVB}}$ tests, we find that although the latter delivers slightly higher power when $T = 50$, they perform quite similarly when $T = 100$, which shows that their power differences disappear very quickly. In the LAD case, the KVB-type test is no longer free from user-chosen parameters and it is of interest to see that $J_{\text{RE}}$ performs similarly to $J_{\text{KVB}}^{(i)}$ and outperforms $J_{\text{KVB}}^{(ii)}$ in both samples. Note that $J_{\text{RE}}$ may be even more powerful than $J_{\text{KVB}}^{(i)}$ in a larger sample, in contrast with the OLS case. As compared $J_{\text{RE}}$ with the HAC-type test, $J_{\text{RE}}$ does suffer from power loss in the OLS case, but it still performs similarly to $J_{\text{HAC}}^{\text{NW,12}}$ when $\beta_{2,o}$ is small. For LAD, the HAC-type test depends on more user-chosen parameters and it is clear that $J_{\text{RE}}$ performs better than $J_{\text{HAC}}^{\text{NW,12,}(ii)}$ in a smaller sample. These simulation results together suggest that the proposed $J_{\text{RE}}$ test is practically useful because it dominates the other tests in terms of finite sample size and can enjoy power advantage when the other tests are computed using inappropriate user-chosen parameters. Finally, we also find by comparing Figures 1 and 2 that although the LAD-based $J_{\text{RE}}$ test is dominated by the OLS-based $J_{\text{RE}}$ test for the AR(1)-HOMO with normal errors, the former may perform better when data are leptokurtic (either resulted from heteroskedasticity or leptokurtic errors).

6 Conclusions

New robust hypothesis tests for nonlinear constraints on M-estimators with possibly non-differentiable estimating functions are proposed in the paper. The proposed approach may serve as a good alternative to hypothesis testing because it does not require consistent estimation of any nuisance parameters in the asymptotic covariance matrix and hence circumvents the problems arising from such consistent estimation, a sharp contrast with the HAC-type and KVB-type tests. Our simulations also suggest that the proposed test is practically useful because it performs better than the HAC-type and KVB-type tests in terms of finite sample size and has power advantage when the latter tests are computed with inappropriate user-chosen parameters.
Appendix

Proof of Lemma 3.1: Let $\xi_{o,[rT]}$ and $\xi_{1,[rT]}$ be defined as

$$
\xi_{o,[rT]} = \frac{1}{\sqrt{T}} \sum_{t=1}^{[rT]} \left[ \phi(z_t; \hat{\theta}_T) - \phi(z_t; \theta_o) - \frac{T}{[rT]} \lambda_{[rT]}(\hat{\theta}_T) + \frac{T}{[rT]} \lambda_{[rT]}(\theta_o) \right],
$$

$$
\xi_{1,[rT]} = \frac{1}{\sqrt{T}} \sum_{t=1}^{[rT]} \left[ \phi(z_t; \theta_o) - \frac{T}{[rT]} \lambda_{[rT]}(\theta_o) + \frac{T}{[rT]} \lambda_{[rT]}(\hat{\theta}_T) \right].
$$

Then it is easy to see that $T^{1/2} \mathbf{m}_{[rT]}(\hat{\theta}_T) = \xi_{o,[rT]} + \xi_{1,[rT]}$ with $\xi_{1,[rT]} = T^{1/2} \mathbf{m}_{[rT]}(\theta_o) + T^{1/2} \lambda_{[rT]}(\hat{\theta}_T)$. We first consider the term $\xi_{1,[rT]}$. By the first-order Taylor expansion about $\theta_o$, we have under [A1] and [A4] that

$$
\sqrt{T} \lambda_{[rT]}(\theta_o) = \left[ \frac{[rT]}{T} \mathbf{M}_{[rT]}(\theta_o) \right] \sqrt{T} (\hat{\theta}_T - \theta_o) + o_P(1)
$$

where the last equality follows from the Bahadur representation: $T^{1/2} (\hat{\theta}_T - \theta_o) = -\mathbf{M}_o^{-1} T^{1/2} \mathbf{m}_T(\theta_o) + o_P(1)$ (as shown in Section 2.1). As a result,

$$
\xi_{1,[rT]} = \sqrt{T} \mathbf{m}_{[rT]}(\theta_o) - r \sqrt{T} \mathbf{m}_T(\theta_o) + o_P(1).
$$

As for $\xi_{o,[rT]}$, its sup norm can be expressed as

$$
\left\| \xi_{o,[rT]} \right\| = \left( \sqrt{\frac{[rT]}{T}} + \frac{[rT]}{T} \left\| \sqrt{T} \lambda_{[rT]}(\hat{\theta}_T) \right\| \right) \zeta_{[rT]}(\hat{\theta}_T, \theta_o).
$$

Under [A1] and [A5], $\zeta_{[rT]}(\hat{\theta}_T, \theta_o) = o_P(1)$. Moreover, $[rT]/T \to r$ and, as shown in equation (11), $T^{1/2} \lambda_{[rT]}(\hat{\theta}_T) = O_P(1)$. These results together imply $\xi_{o,[rT]} = o_P(1)$. As such

$$
\sqrt{T} \mathbf{m}_{[rT]}(\hat{\theta}_T) = \sqrt{T} \mathbf{m}_{[rT]}(\theta_o) - r \sqrt{T} \mathbf{m}_T(\theta_o) + o_P(1).
$$

Given condition [A2] and applying the continuous mapping theorem, we immediately obtain $T^{1/2} \mathbf{m}_{[rT]}(\hat{\theta}_T) \Rightarrow \mathbf{SB}_P(r)$. Applying again the continuous mapping theorem yields the weak limit of $\hat{V}_{KV_B}. \quad \square$
Proof of Theorem 3.2: As $T^{1/2}m_{[T]}(\theta_T)$ ⇒ $SB_p(r)$, we immediately have
\[
R(\hat{\theta}_T)\tilde{M}_T^{-1}\sqrt{T}m_{[T]}(\hat{\theta}_T) \Rightarrow \frac{R(\theta_o)M_o^{-1}SB_p(r) = -R(\theta_o)\Lambda B_p(r) \overset{d}{=} -\Xi B_q(r)}{d}
\]
where $\Lambda = -M_o^{-1}S$ and $\Xi$ is a nonsingular $q \times q$ matrix square root of $\Omega_o$. It then follows from Lemma 3.1 that $\hat{\Omega}_{KVB}$ has the following weak limit:
\[
\hat{\Omega}_{KVB} = \frac{R(\hat{\theta}_T)\tilde{M}_T^{-1}\tilde{V}_{KVB}\tilde{M}_T^{-1}}{d} R(\hat{\theta}_T)^\prime
\]
\[
\Rightarrow R(\theta_o)M_o^{-1}SP_pS^\prime M_o^{-1}R(\theta_o)^\prime
\]
\[
= \int_0^1 \left[-R(\theta_o)\Lambda B_p(r)\right] \left[-R(\theta_o)\Lambda B_p(r)^\prime\right] dr
\]
\[
= \Xi P_q \Xi^\prime.
\]
Given $T^{1/2}(\theta_T - \theta_o) = -M_o^{-1}\sqrt{T}m_T(\theta_o) + o_p(1)$ and applying the delta method,
\[
\sqrt{T} \left[\gamma(\hat{\theta}_T) - \gamma(\theta_o)\right] = -R(\theta_o)M_o^{-1}\sqrt{T}m_T(\theta_o) + o_p(1).
\]
We thus have under the null that
\[
\sqrt{T} \gamma(\hat{\theta}_T) \overset{D}{\rightarrow} -R(\theta_o)M_o^{-1}SW_p(1) = R(\theta_o)\Lambda W_p(1) \overset{d}{=} \Xi W_q(1),
\]
where $\Xi$ is defined as above. Applying the continuous mapping theorem, we can obtain under the null that
\[
J_{KVB} \Rightarrow W_q(1)^\prime \Xi [\Xi P_q \Xi^\prime]^{-1} \Xi W_q(1)
\]
\[
= W_q(1)^\prime \Xi^\prime [\Xi^\prime]^{-1} P_q^{-1} \Xi^{-1} \Xi W_q(1)
\]
\[
= W_q(1)^\prime P_q^{-1} W_q(1). \qed
\]

Proof of Theorem 3.3: Equations (6) and (7) are sufficient to show Theorem 3.3. \qed

Proof of Theorem 3.4: Under the local alternative (8), we have
\[
\sqrt{T} \gamma(\hat{\theta}_T) = \delta_o - R(\theta_o)M_o^{-1}\sqrt{T}m_T(\theta_o) + o_p(1)
\]
\[
\overset{D}{\rightarrow} \delta_o + \Xi W_q(1);
\]
see also the proof of Theorem 3.2. With this result and that $\hat{\Omega}_{HAC} \overset{p}{\Rightarrow} \omega_o$, $\hat{\Omega}_{KVB} \Rightarrow \Xi P_q \Xi^\prime$, and $\hat{\Omega}_{RE} \Rightarrow \Xi P_q \Xi^\prime$ regardless of the values of $\delta_o$, the weak limits for the three
tests immediately follow. □

**Proof of Corollary 3.5:** The results follows directly from Theorem 3.4.
References


Table 1: Empirical sizes of the robust hypothesis tests (OLS, $\rho = 0.5$).

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0,1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$q$ $T = 50$ 100 500</td>
<td>50 100 500</td>
</tr>
<tr>
<td>$J_{RE}$</td>
<td>1 6.3 5.3 5.0</td>
<td>6.3 8.2 9.2</td>
</tr>
<tr>
<td></td>
<td>2 7.8 6.3 5.3</td>
<td>7.8 8.2 9.2</td>
</tr>
<tr>
<td></td>
<td>3 10.4 7.7 5.8</td>
<td>9.7 7.8 7.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0,1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$J_{KVB}$ $q$ $T = 50$ 100 500</td>
<td>50 100 500</td>
</tr>
<tr>
<td></td>
<td>1 9.9 7.3 5.6</td>
<td>9.7 7.9 5.5</td>
</tr>
<tr>
<td></td>
<td>2 11.3 8.2 5.8</td>
<td>8.2 8.7 6.1</td>
</tr>
<tr>
<td></td>
<td>3 14.5 9.5 6.2</td>
<td>14.3 10.5 6.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0,1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$J_{NW}^{HAC}$ $q$ $T = 50$ 100 500</td>
<td>50 100 500</td>
</tr>
<tr>
<td></td>
<td>1 16.2 12.2 8.3</td>
<td>8.3 7.9 7.0</td>
</tr>
<tr>
<td></td>
<td>2 23.7 17.9 9.9</td>
<td>9.9 8.7 7.0</td>
</tr>
<tr>
<td></td>
<td>3 33.4 22.9 10.5</td>
<td>10.5 9.6 8.7</td>
</tr>
</tbody>
</table>

Note: The entries are rejection frequencies in percentage; the nominal size is 5%.

Table 2: Empirical sizes of the robust hypothesis tests (OLS, $\rho = 0.8$).

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0,1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$q$ $T = 50$ 100 500</td>
<td>50 100 500</td>
</tr>
<tr>
<td>$J_{RE}$</td>
<td>1 7.5 5.9 5.6</td>
<td>7.1 6.3 5.3</td>
</tr>
<tr>
<td></td>
<td>2 10.2 8.7 5.5</td>
<td>9.6 8.3 5.7</td>
</tr>
<tr>
<td></td>
<td>3 14.6 11.4 7.9</td>
<td>13.5 11.7 7.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0,1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$J_{KVB}$ $q$ $T = 50$ 100 500</td>
<td>50 100 500</td>
</tr>
<tr>
<td></td>
<td>1 17.8 12.2 6.9</td>
<td>17.4 12.4 7.1</td>
</tr>
<tr>
<td></td>
<td>2 23.1 16.4 7.6</td>
<td>22.9 16.7 8.1</td>
</tr>
<tr>
<td></td>
<td>3 32.3 20.6 9.4</td>
<td>30.6 20.8 9.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0,1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$J_{NW}^{HAC}$ $q$ $T = 50$ 100 500</td>
<td>50 100 500</td>
</tr>
<tr>
<td></td>
<td>1 28.9 23.2 10.7</td>
<td>30.5 23.5 10.5</td>
</tr>
<tr>
<td></td>
<td>2 45.4 35.1 13.6</td>
<td>46.9 34.6 14.8</td>
</tr>
<tr>
<td></td>
<td>3 60.7 46.0 18.3</td>
<td>60.4 44.3 18.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0,1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$J_{NW}^{HAC}$ $q$ $T = 50$ 100 500</td>
<td>50 100 500</td>
</tr>
<tr>
<td></td>
<td>1 34.1 24.6 11.8</td>
<td>35.7 24.9 11.5</td>
</tr>
<tr>
<td></td>
<td>2 53.7 38.6 15.6</td>
<td>55.7 39.2 16.6</td>
</tr>
<tr>
<td></td>
<td>3 70.8 51.6 20.9</td>
<td>69.9 51.3 20.6</td>
</tr>
</tbody>
</table>

Note: The entries are rejection frequencies in percentage; the nominal size is 5%.
Table 3: Empirical sizes of the robust hypothesis tests (LAD, $\rho = 0.5$).

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{N}(0, 1)$</th>
<th>$t(4)$</th>
<th>$\mathcal{N}(0, 1)$</th>
<th>$t(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$q$ = 50</td>
<td>100</td>
<td>500</td>
<td>$q$ = 50</td>
</tr>
<tr>
<td>$J_{RE}$</td>
<td>1</td>
<td>4.4</td>
<td>3.5</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5.0</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7.1</td>
<td>5.4</td>
<td>4.2</td>
</tr>
<tr>
<td>$J_{KV.B}^{(i)}$</td>
<td>1</td>
<td>6.6</td>
<td>5.5</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.7</td>
<td>7.0</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11.2</td>
<td>8.2</td>
<td>5.9</td>
</tr>
<tr>
<td>$J_{KV.B}^{(ii)}$</td>
<td>1</td>
<td>9.3</td>
<td>7.4</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12.8</td>
<td>9.9</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>19.4</td>
<td>13.5</td>
<td>8.5</td>
</tr>
<tr>
<td>$J_{NW;4,(i)}$</td>
<td>1</td>
<td>13.6</td>
<td>9.8</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>20.0</td>
<td>14.0</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>28.8</td>
<td>20.3</td>
<td>9.7</td>
</tr>
<tr>
<td>$J_{NW;4,(ii)}$</td>
<td>1</td>
<td>17.5</td>
<td>13.2</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28.1</td>
<td>19.7</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>41.7</td>
<td>30.3</td>
<td>14.8</td>
</tr>
<tr>
<td>$J_{NW;12,(i)}$</td>
<td>1</td>
<td>23.7</td>
<td>17.2</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>39.1</td>
<td>27.8</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>55.8</td>
<td>41.6</td>
<td>16.6</td>
</tr>
</tbody>
</table>

Note: The entries are rejection frequencies in percentage; the nominal size is 5%.
Table 4: Empirical sizes of the robust hypothesis tests (LAD, $\rho = 0.8$).

<table>
<thead>
<tr>
<th></th>
<th>AR(1)-HOMO</th>
<th>AR(1)-HET</th>
<th>AR(1)-HET</th>
<th>AR(1)-HET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N(0, 1)$</td>
<td>$t(4)$</td>
<td>$N(0, 1)$</td>
<td>$t(4)$</td>
</tr>
<tr>
<td>$q$</td>
<td>$T = 50$</td>
<td>100</td>
<td>500</td>
<td>50</td>
</tr>
<tr>
<td>$J_{RE}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.3</td>
<td>4.9</td>
<td>4.2</td>
<td>4.6</td>
</tr>
<tr>
<td>2</td>
<td>7.1</td>
<td>6.3</td>
<td>5.0</td>
<td>6.8</td>
</tr>
<tr>
<td>3</td>
<td>9.2</td>
<td>7.6</td>
<td>6.0</td>
<td>9.1</td>
</tr>
<tr>
<td>$J_{KV,B}^{(i)}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>12.3</td>
<td>8.9</td>
<td>5.7</td>
<td>12.0</td>
</tr>
<tr>
<td>2</td>
<td>18.3</td>
<td>12.8</td>
<td>7.2</td>
<td>18.1</td>
</tr>
<tr>
<td>3</td>
<td>25.8</td>
<td>17.7</td>
<td>8.4</td>
<td>26.1</td>
</tr>
<tr>
<td>$J_{KV,B}^{(ii)}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15.7</td>
<td>11.5</td>
<td>6.1</td>
<td>15.2</td>
</tr>
<tr>
<td>2</td>
<td>24.3</td>
<td>16.7</td>
<td>9.1</td>
<td>25.2</td>
</tr>
<tr>
<td>3</td>
<td>36.1</td>
<td>26.0</td>
<td>11.0</td>
<td>36.2</td>
</tr>
<tr>
<td>$J_{NW,4}{i}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>24.5</td>
<td>18.0</td>
<td>9.1</td>
<td>23.6</td>
</tr>
<tr>
<td>2</td>
<td>37.2</td>
<td>28.2</td>
<td>12.6</td>
<td>38.4</td>
</tr>
<tr>
<td>3</td>
<td>51.5</td>
<td>39.6</td>
<td>17.1</td>
<td>51.9</td>
</tr>
<tr>
<td>$J_{NW,4}{ii}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>28.8</td>
<td>21.1</td>
<td>10.7</td>
<td>28.0</td>
</tr>
<tr>
<td>2</td>
<td>44.6</td>
<td>34.4</td>
<td>16.9</td>
<td>46.6</td>
</tr>
<tr>
<td>3</td>
<td>62.4</td>
<td>50.5</td>
<td>22.7</td>
<td>62.7</td>
</tr>
<tr>
<td>$J_{NW,12}{i}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>33.1</td>
<td>23.0</td>
<td>11.0</td>
<td>32.7</td>
</tr>
<tr>
<td>2</td>
<td>52.6</td>
<td>38.1</td>
<td>18.0</td>
<td>54.1</td>
</tr>
<tr>
<td>3</td>
<td>70.3</td>
<td>55.6</td>
<td>24.6</td>
<td>71.7</td>
</tr>
</tbody>
</table>

Note: The entries are rejection frequencies in percentage; the nominal size is 5%. 28
Figure 1: The size-adjusted powers of the OLS-based tests: AR(1)-HOMO with (a) normal errors and $T = 50$, (b) normal errors and $T = 100$, (c) $t(4)$ errors and $T = 50$, (d) $t(4)$ errors and $T = 100$; AR(1)-HET with normal errors, (e) $T = 50$ and (f) $T = 100$. 
Figure 2: The size-adjusted powers of the LAD-based tests: AR(1)-HOMO with (a) normal errors and $T = 50$, (b) normal errors and $T = 100$, (c) $t(4)$ errors and $T = 50$, (d) $t(4)$ errors and $T = 100$; AR(1)-HET with normal errors, (e) $T = 50$ and (f) $T = 100$. 