Robust Inference

Bruce E. Hansen[*]
University of Wisconsin[†]

This draft: September 2014
Preliminary. Do not cite.

Abstract

This paper examines inference in sieve nonparametric regression allowing for asymptotic bias. We show how to construct asymptotically valid confidence intervals which are robust in the sense that they account for finite sample estimation bias. The theory is based on a non-central approximation to the asymptotic distribution of the t-statistic, where the non-centrality parameter captures the asymptotic bias. This non-centrality parameter can be estimated, though not consistently. A plug-in critical value, based on the non-central normal distribution with the estimated non-centrality parameter, works better than the classical critical value, but does not provide uniform valid coverage. By test statistic inversion, we can construct an asymptotically valid confidence interval for the non-centrality parameter. Using the upper bound of this confidence interval, we can construct a conservative critical value for the parameter of interest. Furthermore, by suitable choice of level of this first-stage confidence interval, we can construct confidence intervals for the parameter of interest which have uniformly correct coverage.

[*] Research supported by the National Science Foundation.

[†] Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706.

1 Introduction

To be written

2 Series Regression

Consider a sample of iid observations (y_i, z_i), $i = 1, ..., n$ where $z_i \in \mathbb{R}^d$. Define the conditional mean $g(z) = \mathbb{E}(y_i \mid z_i = z)$ and the regression error $e_i = y_i - g(z_i)$.

We examine the estimation of $g(z)$ by series or sieve regression. For some increasing sequence $K = K(n)$ let $x_K(z)$ be a set of $K \times 1$ basis functions. For example, a power series sets $x_K(z) = \left(1, z, ..., z^{K-1}\right)$. Construct the regressors $x_{Ki} = x_K(z_i)$.

A series regression approximates the conditional mean $g(z)$ by a linear projection of y_i on x_{Ki}. We can write this approximating model as

$$y_i = x_{Ki}'\beta_K + e_{Ki} \tag{1}$$

where the coefficient is defined by linear projection

$$\beta_K = \left(\mathbb{E}\left(x_{Ki}x_{Ki}'\right)\right)^{-1}\mathbb{E}\left(x_{Ki}y_i\right). \tag{2}$$

The K^{th} series approximation to $g(z)$ is the linear function $g_K(z) = x_K(z)'\beta_m$. The corresponding approximation error is $r_K(z) = g(z) - x_K(z)'\beta_K$.

A real-valued parameter of interest θ may be a the regression function $g(z)$ at a fixed point z or some other linear function of g. Linear functions includes derivatives and integrals over g. In this case we can write the parameter as

$$\theta = a(g) \tag{3}$$

where the function a is linear. Linearity inplies that if we plug in the series approximation $g_K(z)$ into (3), then we obtain the approximating (or pseudo-true) parameter value

$$\theta_K = a(g_K) = a_K'\beta_K \tag{4}$$

for some $K \times 1$ vector $a_K \neq 0$. For example, if the parameter of interest is the regression function $g(z)$, then $a_K = x_K(z)$.

The standard estimator of (2) is least-squares of y_i on x_{Ki}:

$$\widehat{\beta}_K = \left(\sum_{i=1}^n x_{Ki}x_{Ki}'\right)^{-1}\sum_{i=1}^n x_{Ki}y_i.$$

The corresponding estimator of $g(z)$ is

$$\widehat{g}_K(z) = x_K(z)'\widehat{\beta}_K$$

and that of θ is

$$\widehat{\theta}_K = a\left(\widehat{g}_K\right) = a_K'\widehat{\beta}_K.$$

The least-squares residuals are $\widehat{e}_{Ki} = y_i - x_{Ki}'\widehat{\beta}_K$ and standardized residuals $\overline{e}_{Ki} = \widehat{e}_{Ki}/\left(1 - h_{Ki}\right)^{1/2}$ where $h_{Ki} = x_{Ki}'\left(\sum_{i=1}^n x_{Ki}x_{Ki}'\right)^{-1} x_{Ki}$.

To form a standard error for $\widehat{\theta}_K$ define

$$\begin{aligned}
\widehat{Q}_K &= \frac{1}{n}\sum_{i=1}^n x_{Ki}x_{Ki}' \\
\widehat{S}_K &= \frac{1}{n}\sum_{i=1}^n x_{Ki}x_{Ki}'\overline{e}_{Ki}^2 \\
\widehat{V}_K &= \widehat{Q}_K^{-1}\widehat{S}_K\widehat{Q}_K^{-1} \qquad (5)
\end{aligned}$$

and then the standard error is

$$s(\widehat{\theta}_K) = \sqrt{a_K'\widehat{V}_K a_K/n}. \qquad (6)$$

These are estimates of

$$\begin{aligned}
Q_K &= \mathbb{E}\left(x_{Ki}x_{Ki}'\right) \\
S_K &= \mathbb{E}\left(x_{Ki}x_{Ki}'e_{Ki}^2\right) \\
V_K &= Q_K^{-1}S_KQ_K^{-1}
\end{aligned}$$

For inference on θ, the absolute value of the t statistic is

$$T_n(\theta) = \left|\frac{\widehat{\theta}_K - \theta}{s(\widehat{\theta}_K)}\right|. \qquad (7)$$

To test the hypothesis

$$H_0 : \theta = \theta_0$$

a classical level α asymptotic test is to reject H_0 if $T_n(\theta_0) \geq q_{1-\alpha}$ where q_η is the η quantile of the distribution of $|N(0,1)|$ e.g. q_η solves $\Phi(q_\eta) - \Phi(-q_\eta) = \eta$ where $\Phi(u)$ is the standard normal distribution function. Furthermore, a classical level α asymptotic confidence interval for θ is

$$\begin{aligned}
C_0 &= \{\theta : T_n(\theta) \leq q_\alpha\} \qquad (8) \\
&= \left\{\widehat{\theta}_K \pm s(\widehat{\theta}_K)q_\alpha\right\}.
\end{aligned}$$

3 Asymptotic Distribution

By linearity and (4), $g(z) = g_K(z) + r_K(z)$ and $\theta = \theta_K + a(r_K)$. It follows that for each fixed K, θ_K is the pseudo-true parameter value (the value being estimated by $\widehat{\theta}_K$ for fixed K) and the finite-K estimation bias is $a(r_K)$. Under standard regularity conditions, the estimator $\widehat{\theta}_K$ has the asymptotic distribution

$$\frac{\sqrt{n}\left(\widehat{\theta}_K - \theta + a\left(r_K\right)\right)}{\left(a_K' V_K a_K\right)^{1/2}} \longrightarrow_d N(0,1). \tag{9}$$

To characterize this distribution more precisely we add some additional assumptions designed to simplify the above representation.

Assumption 1 *For some constants $\phi > 0$, $\gamma > 0$, and $\tau_1 > 0$*

1. $\lim_{K\to\infty} K^{-\phi} a_K' V_K a_K = D > 0$
2. $\lim_{K\to\infty} K^{\gamma} a(r_K) = A \neq 0$
3. $\theta = \theta_0 + \overline{\delta} n^{-\gamma/(\phi+2\gamma)}$
4. $K = \tau_1 n^{1/(\phi+2\gamma)}$

Assumption 1.1 states that the asymptotic variance in (9) is increasing in the number of regressors at some rate K^{ϕ}. This means that the parameter estimate $\widehat{\theta}_K$ is converging at a rate slower than $\sqrt{n}$. This is typical for parameters of interest such as the regression function $g(z)$ at a point. Chen and Liao (2014) follow Khan and Tamer (2010) and refer to such parameters as *irregular*. We prefer to label such parameters as *nonparametric*.

Assumption 1.2 states that the approximation error decreases at some rate $K^{-\gamma}$ and when scaled converges to a constant. This may be stronger than necessary, but allows a precise characterization of how the approximation error enters the asymptotic distribution.

Applying Assumptions 1.1 and 1.2 to the asymptotic distribution (9), we can calculate that the asymptotic mean-squared error of $\widehat{\theta}_K$ for θ is $DK^{\phi}/n + A^2 K^{-2\gamma}$. The K which minimizes this asymptotic MSE is $K_{opt} = \left(2\gamma A^2/\phi D\right)^{1/(\phi+2\gamma)} n^{1/(\phi+2\gamma)}$. This shows that under these asssuptions the MSE-optimal choice of K should increase at a rate proportional to $n^{1/(\phi+2\gamma)}$, resulting in an optimal MSE of order $n^{-2\gamma/(\phi+2\gamma)}$.

Assumption 1.3 states that the true parameter θ is local to the constant θ_0 with localizing parameter δ. The rate of convergence of θ to θ_0 is set equal to the square root of the optimal MSE.

Assumption 1.4 states that K grows at the MSE-optimal rate $n^{1/(\phi+2\gamma)}$.

Given Assumption 1, we can give a precise characterization of the asymptotic distribution of the t statistic (7) evaluated at the true value θ or the hypothesized value θ_0.

Theorem 1 *Under standard sieve regularity conditions and Assumption 1, as $n \to \infty$*

$$T_n(\theta) \longrightarrow_d T(\lambda) \sim |Z_1 + \lambda| \tag{10}$$

and

$$T_n(\theta_0) \longrightarrow_d T(\lambda + \delta) \sim |Z_1 + \lambda + \delta| \tag{11}$$

where $Z_1 \sim N(0,1)$ *and*

$$\lambda = \frac{-A}{\sqrt{D\tau_1^{\phi+2\gamma}}}$$
$$\delta = \frac{\overline{\delta}}{\sqrt{D\tau_1^{\phi}}}$$

The asymptotic distribution $T(\zeta)$ in (10) and (11) is the absolute value of a normal random variable with mean ζ and unit variance. Equivalently, $T(\zeta)^2 \sim \chi_1^2\left(\zeta^2\right)$, a non-central chi-square distribution with one degree of freedom and non-centrality parameter ζ^2. We will call the distribution of $T(\zeta)$ a *non-central normal* distribution for brevity.

The asymptotic distribution (10) is for the t-statistic evaluated at the true value of θ. In contrast to the parametric case, where the t statistic is asymptotically standard normal, the distribution in (10) is noncentral normal with noncentrality parameter λ which is an *asymptotic bias.* The asymptotic distribution (11) is the distribution of the t-statistic evaluated at the hypothesized value θ_0. This asymptotic distribution depends on the localizing parameter δ due to the deviation of the true value θ from θ_0. Under the null hypothesis they agree. The noncentrality in the distribution (11) is the sum of λ and δ, so both the asymptotic bias and the localizing parameter affect the distribution of the test statistic.

It is instructive to use Theorem 1 to investigate how the choice of the number of regressors K impacts the distribution of the statistic. Under Assumption 1.3 the choice of K is determined by the scale parameter τ_1. Larger τ_1 implies a larger K. From Theorem 1 we can see that the non-centrality parameters λ and δ are both decreasing in τ_1. Selecting a large K is often called *undersmoothing*, and is typically assumed in the nonparametrics sieve literature to eliminate the asymptotic bias term. Indeed, as τ_1 increases to infinity the bias term λ decreases to zero. However, δ also decreases to zero as τ_1 increases to infinity, so the effect of undersmoothing on the asymptotic distribution is to eliminate both noncentrality terms, thus eliminating both the bias and the power of the test against the alternatives specified in Assummption 1.3

4 Classical Confidence Interval

Let $F(x,\lambda)$, $f(x,\lambda)$, and $q_\eta(\lambda)$ denote the distribution function, density function and quantile function of the non-central normal distribution $T(\lambda)$. It may be convenient to observe that the distribution and density functions can be written as

$$\begin{aligned} F(x,\lambda) &= \Pr\left(T(\lambda) \le x\right) \\ &= \Phi\left(x - |\lambda|\right) - \Phi\left(-x - |\lambda|\right) \\ f(x,\lambda) &= \phi\left(x - |\lambda|\right) + \phi\left(x + |\lambda|\right). \end{aligned}$$

The quantile function $q_\eta(\lambda)$ is the solution to the equation $F(q_\eta(\lambda), \lambda) = \eta$. We don't have an explicit solution, but have the lower bound $q_\eta(\lambda) \geq \lambda + \Phi^{-1}(\eta)$ which approaches equality as λ increases.

Given Theorem 1 and this notation, it is simple to calculate the asymptotic coverage probability of the classical confidence interval (8).

Corollary 1 $\Pr(\theta \in C_0) \to F(q_\alpha, \lambda) \leq \alpha$, *with strict inequality when* $\lambda \neq 0$.

Corollary 1 shows that the classical confidence interval only has correct (pointwise in λ) coverage in the special case $\lambda = 0$. Otherwise it is guarenteed to *undercover* the parameter θ.

In Figure 1 we plot $F(c_{.95}, \lambda)$ as a function of λ. It shows that the coverage probability is a declining funtion of λ, and coverage probabilities are far from the nominal 0.95 even for modest values of λ. For example, if $\lambda = 1$ then the coverage probability is 0.83, and for $\lambda = 2$ the probability is 0.48. The coverage probability asymptotes to 0 as λ increases.

This result shows that the classical confidence interval is fragile to the undersmoothing assumption and is severely non-robust.

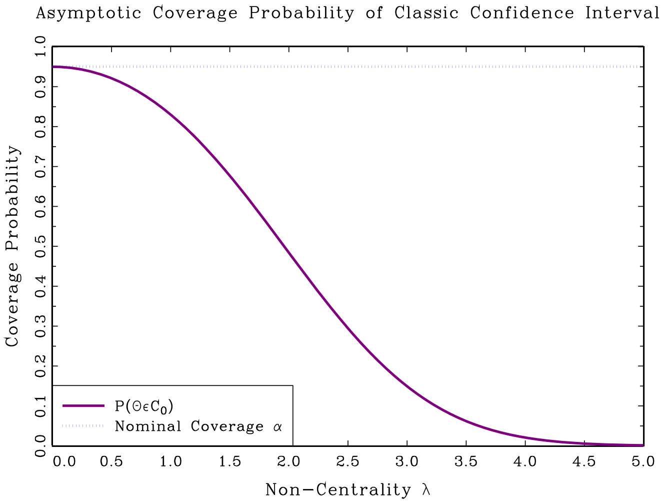

Figure 1: Asymptotic coverage probability of classic confidence interval

5 Classical Test Power

Coverage probability is equivalent to (one minus) the Type I error probability. Also of interest is test power, which in the context of confidence intervals affects their length. The classical t-test rejects H_0 if $T_n(\theta_0) \geq q_\alpha$. The probability of this event is the power of the test.

Corollary 2 $\Pr(T_n(\theta_0) \geq q_\alpha) \to 1 - F(q_\alpha, \lambda + \delta)$

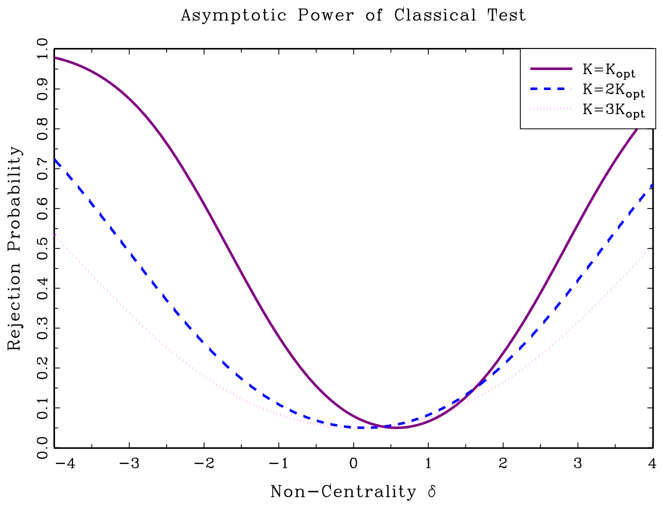

Figure 2: Asymptotic power of classical test

To illustrate, in Figure 2 we plot the asymptotic power of a nominal 5% test, setting $\phi = 1$, $\gamma = 2$, $D = 1$, and $A = 1$. We set $\tau_1 = \left(4A^2\right)^{1/3}$ as a baseline (as this is the choice which minimizes the asymptotic MSE), and then also plot the power with $\tau_1 = 2\left(4A^2\right)^{1/3}$ and $\tau_1 = 3\left(4A^2\right)^{1/3}$ to see the effect of increasing K. The power of the tests are plotted against $\overline{\delta}$. Because $A \neq 0$, there is asymptotic bias so the classical test overrejects at $\overline{\delta} = 0$ and has power equal to the nominal size at $\overline{\delta} = A$. The overrjection at $\overline{\delta} = 0$ is identical to the undercoverage reported in Figure 1. What is new about Figure 2 is the comparison across choices for K. A conventional recommendation in the sieve literature is to undersmooth, e.g. to assume that K is sufficiently large that the bias can be ignored. In the context of Figure 2 this is equivalent to selecting a large K, which implies the dashed or dotted power curves. Indeed, by examining these curves at $\overline{\delta} = 0$ we can see that they have meaningfully reduced size distortion. However, the test power is also reduced.

If λ were known we could use the noncentral critical values $q_\alpha(\lambda)$ and these would correct the size distortion shown in Figure 2. A test which rejects for $T_n(\theta_0) \geq q_\alpha(\lambda)$ would have the asymptotic power $1 - F(q_\alpha(\lambda), \lambda + \delta)$. To illustrate, Figure 3 plots these asymptotic power functions for the same context as in Figure 2. Thus the three curves represent the power of the tests using $K = K_{opt}$, $K = 2K_{opt}$ and $K = 4K_{opt}$, where K_{opt} is the choice of K which minimizes the asymptotic MSE. These curves have been constructed so to have correct asymptotic size, and thus pass through 0.05 at $\overline{\delta} = 0$. The three curves differ in terms of their power properties, and there is no single uniformly most powerful test. All three tests have low power for $\overline{\delta}$ near one (due to the asymptotic bias) and the test with $K = K_{opt}$ has the smallest power in this region. For values of $\overline{\delta}$ further from one the test with $K = K_{opt}$ has the highest power (and is considerably higher than the other tests).

The message from Figures 2 and 3 is that while undersmoothing can reduce size distortion, it cannot eliminate size distortion, and reduces the power of tests against distant alternatives. Our goal is to develop tests which can control for size distortion while retaining good power.

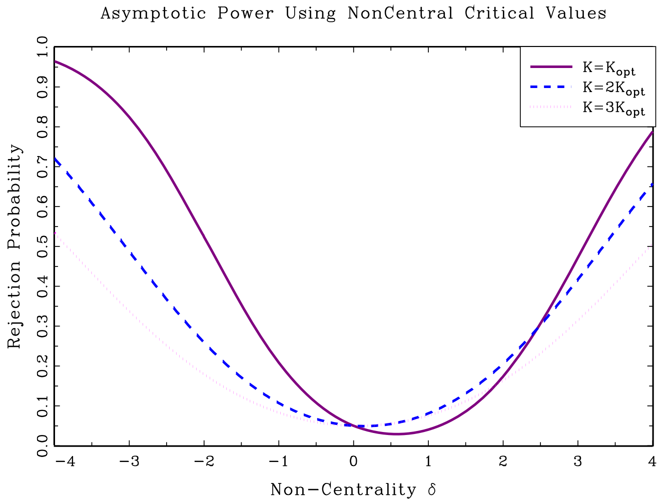

Figure 3: Asymptotic power using noncentral critical value

6 Illustration

How large is the undercoverage distortion in practice? To investigate, let us consider the following simple experiment. Consider the following the process

$$\begin{aligned} y_i &= g(z_i) + e_i \\ g(z) &= \frac{\sin(z)}{z^{2/3}} \\ z_i &\sim U[0,10] \\ e_i &\sim N(0,1). \end{aligned}$$

Suppose we have a sample of 100 observations from this process, and the goal is to estimate the conditional mean. Since the sample size is small, an investigator would probably estimate a relatively low-dimensional model. Suppose a quadratic regression is fit to the observations, and 95% classic confidence intervals formed. The fitted values and intervals are displayed in Figure 4, along with the true conditional mean

Given the strong nonlinearity in the true conditional mean, it is not surprising to observe that the estimated regression is a poor fit. What is more important from our point of view is that the confidence intervals are highly misleading. They fail to cover the true conditional mean, for most values of z. The reason is because they do not address estimation bias. Equivalently, they are constructed using the central normal distributional approximation, when the non-central normal is more appropriate.

Certainly, a wiser researcher might have estimated a more flexible specification. Suppose instead that a quadratic spline with one knot was fit to the same data. The result is displayed in Figure 5.

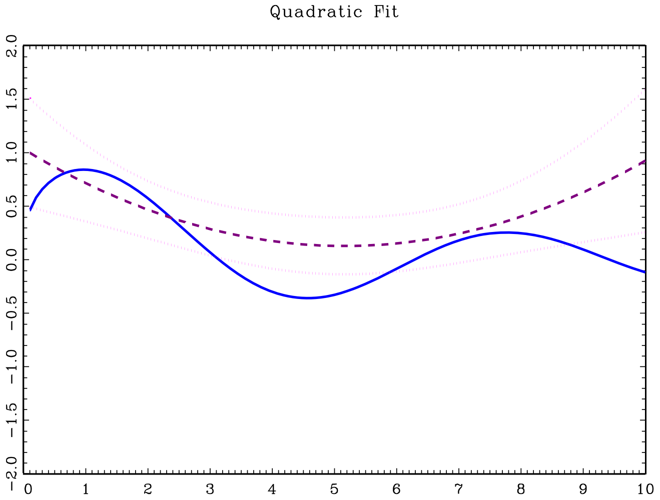

Figure 4: Quadratic fit

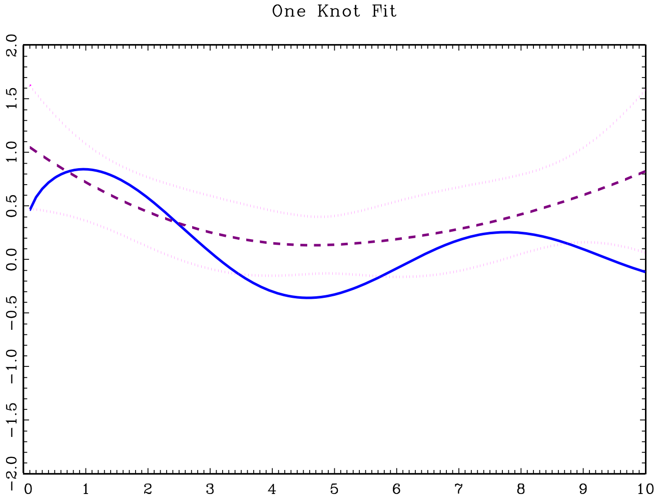

Figure 5: One knot fit

It is not clear if these confidence intervals are an important improvement. Now suppose the researcher fits a quadratic spline with 2 knots. The result is displayed in Figure 6.

The two-knot estimate works much better. The fit is better, and the confidence intervals contain the true regression for most values of z. Still, the intervals exhibit uncercoverage. The estimate with 3 knots has a similar characteristic.

We next investigated the performance in a simulation experiment. We generated 100,000 samples from the same model, and estimated the regression function using quadratic splines with N

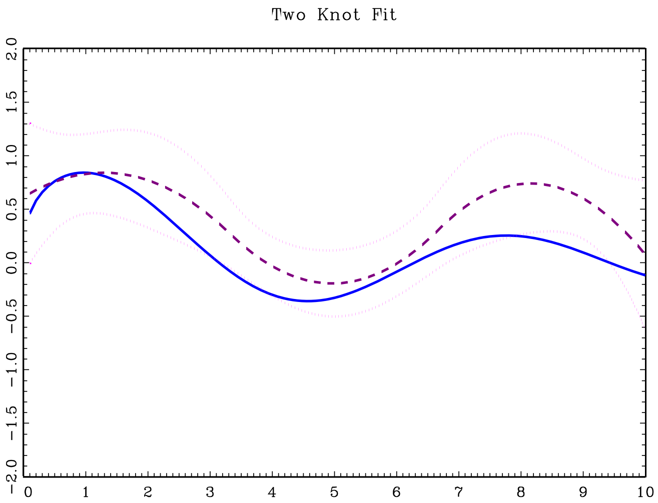

Figure 6: Two knot fit

equally spaced knots. By simulation, it was determined that $N = 2$ minimized the integrated mean squared error, so the remainder of the analysis fixes the number of knots at $N = 2$.

Next, we calculated the finite sample bias and variance of the spline estimator. Let $\widehat{g}_K(z)$ be the spline estimator, let $g_K(z) = \mathbb{E}(\widehat{g}_K(z))$ be its expectation, and $V_K(z) = \text{var}(\widehat{g}_K(z))$ its variance.

In Figure 7 we plot the absolute bias $|g(z) - g_K(z)|$ and standard deviation $V_K(z)^{1/2}$. The main point to see is that the two functions are of a similar order of magnitude. Consequently, it should not be surprising to find that inference which takes only the latter into account will be distorted.

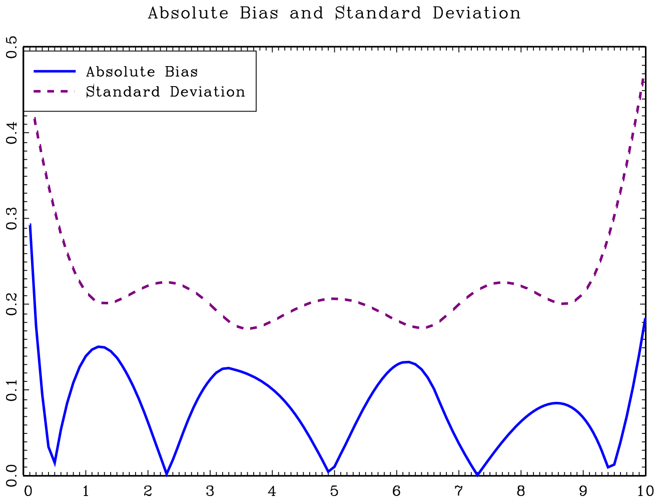

Figure 7: Absolute bias and standard deviation

In Figure 8 we plot the pointwise coverage probability of the 95% classical confidence interval C_0. The coverage probability oscillates between 0.85 and 0.94, and has undercoverage for all values of z. Comparing Figures 7 and 8, we can see that for values of z with larger estimation bias, the corresponding coverage probability is low.

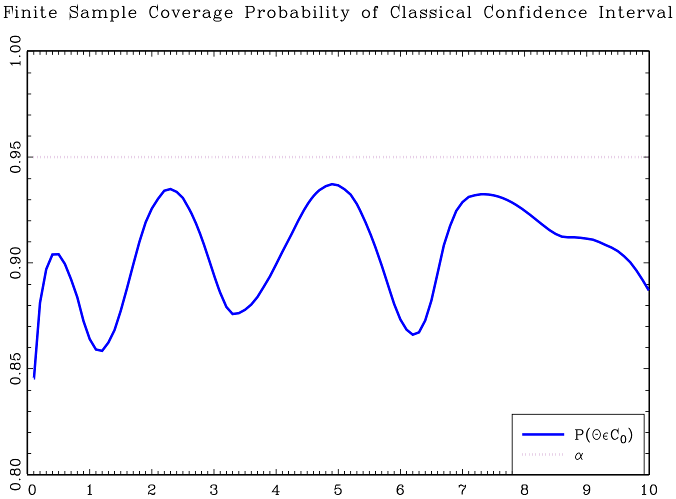

Figure 8: Finite sample coverage probability of classical confidence interval

7 Local Robustness

Following an idea of Hansen and Sargent (2007) we could consider inference which is locally robust to misspecification. Namely, it may be reasonable to assume that λ is non-zero but small. Hansen and Sargent (2007) recommend that we consider decision rules which are locally robust in the sense that they are minimax (minimize the worst-case risk) in misspecification regions which are local to the specified model. In our context this is equivalent to constructing confidence intervals which have correct asymptotic coverage for all values of λ in a local set $\{\lambda \leq c\}$ for some $c > 0$, perhaps $c = 1$ or $c = 2$. For a confidence interval of the form $\{\theta : T_n(\theta) \leq q\}$ for some critical value q, the uniform asymptotic coverage probability in the set $\{\lambda \leq c\}$ is

$$\inf_{\lambda \leq c} \lim_{n\to\infty} \Pr(T_n(\theta) \leq q) = F(q, c).$$

To set this equal to the nominal coverage probability α, it follows that we need to set the critical value as $q = q_\alpha(c)$, the level-α critical value from the non-central normal distribution with noncentrality parameter c. For example, $q_{.95}(1) = 2.65$ and $q_{.95}(2) = 3.64$.

We define the local robust confidence intervals

$$C_c = \left\{ \widehat{\theta}_K \pm s(\widehat{\theta}_K) q_\alpha(c) \right\}.$$

We find

$$\Pr(\theta \in C_c) \to F(q_\alpha(c), \lambda)$$

$$\inf_{\lambda \le c} \lim_{n\to\infty} \Pr(\theta \in C_c) = \alpha$$

Thus the confidence intervals C_c are locally robust to $\lambda \le c$.

To illustrate, in Figure 9 we display the asymptotic coverage probabilities of the classical confidence interval C_0 and the locally robust intervals C_1 and C_2. We can see that the intervals are quite conservative for small λ, cross the nominal coverage level α at c (as designed) but are globally non-robust.

Figure 9 suggests that the locally robust confidence intervals are not an attractive option.

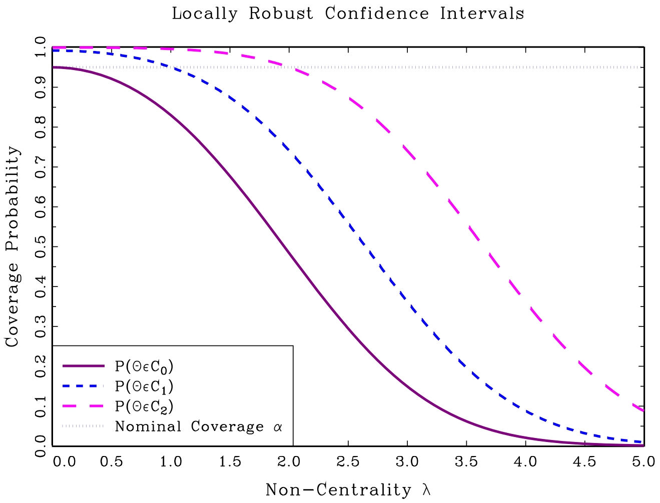

Figure 9: Locally robust confidence intervals

8 Estimation of NonCentrality Parameter

Corollary 1 shows that to construct a confidence interval with correct asymptotic coverage $1-\alpha$, we could use the critical value $q_\alpha(\lambda)$ from the non-central normal distribution. This is not feasible as λ is unknown.

We can, however, attempt to estimate λ. For some $L > K$, consider the estimate $\widehat{\theta}_L = a(\widehat{g}_L)$. Given some constant $\varepsilon > 0$ to be discussed below, an estimator of λ is

$$\widehat{\lambda}_n = \frac{\sqrt{n}\left|\widehat{\theta}_K - \widehat{\theta}_L\right|}{\sqrt{a_K' \widehat{V}_K a_K}} (1+\varepsilon). \tag{12}$$

Since $\widehat{\lambda}$ is random, and not consistent for λ, we need to assess its sampling distribution. We define the following

$$\begin{aligned}
Q_L &= \mathbb{E}\left(x_{Li}x_{Li}'\right) \\
S_L &= \mathbb{E}\left(x_{Ki}x_{Li}'e_{Li}^2\right) \\
V_L &= Q_L^{-1}S_LQ_L^{-1} \\
S_{KL} &= \mathbb{E}\left(x_{Ki}x_{Li}'e_{Ki}e_{Li}\right) \\
V_{KL} &= Q_K^{-1}S_{KL}Q_L^{-1}.
\end{aligned}$$

Assumption 2 *For some $\tau_2 > \tau_1$, $L = \tau_2 n^{1/(\phi+2\gamma)}$.*

Assumption 2 specifies that L is larger than K but increases with n at the same rate.

Theorem 2 *Under standard sieve regularity conditions and Assumptions 1 and 2, as $n \to \infty$*

$$\begin{pmatrix} T_n(\theta_0) \\ \widehat{\lambda}_n \end{pmatrix} \longrightarrow_d \begin{pmatrix} T(\lambda+\delta) \\ \xi \end{pmatrix} \sim \begin{pmatrix} |Z_1+\lambda+\delta| \\ v\left|Z_2+\dfrac{\lambda B}{v}\right| \end{pmatrix}$$

where

$$\begin{aligned}
B &= \left(1-R^{-\gamma}\right)(1+\varepsilon) \\
R &= \frac{\tau_2}{\tau_1} \\
v &= (1+\varepsilon)\sqrt{1+v_1-2v_2},
\end{aligned}$$

$$\begin{aligned}
v_1 &= R^{\phi/(\phi+2\gamma)} \\
v_2 &= \lim \frac{a_K'V_{KL}a_L}{a_K'V_Ka_K}
\end{aligned}$$

$$\rho = \frac{1-v_2}{\sqrt{1+v_1-2v_2}},$$

and

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Theorem 2 provides the joint asymptotic distribution of the t statistic and the noncentrality parameter estimate. It shows that $\widehat{\lambda}_n$ is asymptotically a scaled non-central normal distribution with non-centrality parameter $\lambda B/v$. It will be useful to note that the marginal density of ξ is $v^{-1}f\left(\xi/v, \lambda B/v\right)$. It is particularly important to observe that the result shows that the asymptotic distribution of $\widehat{\lambda}_n$ does not depend on the localizing parameter δ, thus is unaffected under the alternative hypothesis.

One can check that by the Cauchy-Schwarz inequality, $|v_2| \leq v_1^{1/2}$ and thus $1 + v_1 - 2v_2 \geq \left(v_1^{1/2} - 1\right)^2 > 0$ since $v_1 > 1$ under Assumption 2. Thus the parameters v and ρ are well defined.

For our subsequent calculations, independence of $T(\lambda)$ and ξ will be a great convenience, which occurs when $\rho = 0$. This occurs when the regressor vectors x_{Ki} and x_{Li} are nested and the error e_i is conditionally homoskedastic $\mathbb{E}\left(e_i^2 \mid z_i\right) = \sigma^2$. For then by the classic theory of nested efficient estimation (Hausman, 1978) the estimators $\widehat{\theta}_K$ and $\widehat{\theta}_L - \widehat{\theta}_K$ are asymptotically uncorrelated.

Furthermore, it will be convenient to assume that $B \geq 1$, which will allow us to bound the coverage probabilities in the next sections. For any ε this is achieved by selecting τ_2 sufficiently large. Suppose that $\phi = 1$ (as it typical for nonparametric parameters θ), $\gamma = 2$ (which is a mild convergence rate for the approximation error), and $R = 2$. Then setting $\varepsilon = 1/3$ is sufficient for $B \geq 1$.

We now impose these conditions.

Assumption 3 $\rho = 0$ *and* $B \geq 1$.

9 Plug-In Confidence Intervals

Given $\widehat{\lambda}_n$ from (12) we can use the estimate $q_\alpha(\widehat{\lambda}_n)$ as a critical value, where $q_\alpha(\lambda)$ is the critical value function of the noncentral normal distribution. This yields a plug-in confidence interval

$$C_{Plug-In} = \left\{\theta : T_n(\theta) \leq q_\alpha(\widehat{\lambda}_n)\right\}.$$

More generally, for any non-negative and non-decreasing function $q(\lambda)$ of λ we could use the critical value $q(\widehat{\lambda}_n)$. This leads to the class of confidence intervals

$$C = \left\{\theta : T_n(\theta) \leq q(\widehat{\lambda}_n)\right\}. \tag{13}$$

It turns out that under Assumption 3 the coverage probability of the confidence interval class (13) can be represented as an integral problem. By independence of $T(\lambda) \sim |Z_1 + \lambda|$ and $\xi \sim v\,|Z_2 + \lambda B/v|$,

$$\begin{aligned}
\Pr(\theta \in C) &= \Pr(T_n(\theta) \leq q(\widehat{\lambda}_n)) \\
&\longrightarrow \Pr(T(\lambda) \leq q(\xi)) \\
&= \mathbb{E}F(q(\xi), \lambda) \\
&= \frac{1}{v}\int_0^\infty F(q(\xi), \lambda) f\left(\xi/v, \lambda B/v\right) d\xi \\
&\geq \frac{1}{v}\int_0^\infty F(q(\xi), \lambda) f\left(\xi/v, \lambda/v\right) d\xi \qquad (14) \\
&= P(\lambda)
\end{aligned}$$

say. The inequality holds since the assumption $B \geq 1$ means that the density $f\left(\xi/v, \lambda B/v\right)$ first-order stochastically dominates $f\left(\xi/v, \lambda/v\right)$ and the integral is over a non-decreasing function.

In some cases, for example if $q(\xi)$ is linear in ξ, it may be possible to calculate the integral (14) using series methods (Fayed and Atiya, 2014). In general, however, this appears to be not possible, so instead we use numerical integration. We approximate (14) by the approximation

$$\frac{1}{v}\int_0^\infty F(q_\alpha(\xi),\lambda) f\left(\xi/v, \lambda/v\right) d\xi \simeq \left(\frac{4v + \dfrac{\lambda}{v}}{5000v}\right) \sum_{i=1}^{5000} F(q_\alpha(\xi_i),\lambda) f\left(\xi_i/v, \lambda/v\right)$$

$$\xi_i = i\left(\frac{4v + \dfrac{\lambda}{v}}{5000v}\right)$$

This is a numerical integral using 5000 gridpoints uniformly over the interval $[0, 4v + \dfrac{\lambda}{v}]$, which is used since the density $f\left(\xi/v, \lambda/v\right)$ has minimal support outside this interval.

We report in Figure 10 the asymptotic coverage probability of the Plug-In interval as a function of λ/v for $v = 1/2$, 1, and 2.

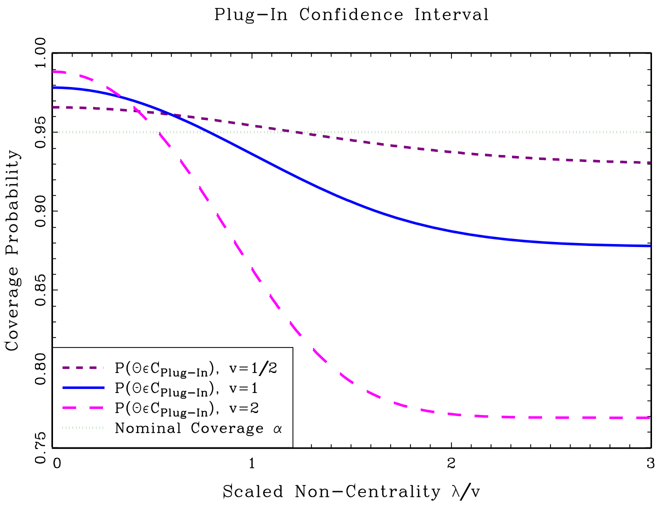

Figure 10: Plug-in confidence interval

The calculations reported in Figure 10 are quite interesting and a dramatic contrast to Figure 9. The plug-in coverage probabilities are much less sensitive to λ than the classical and locally robust coverage probabilities. They are locally robust (their coverage probabilities exceed the nominal level for small λ), and asymptote to a positive coverage probability as $\lambda \to \infty$. The latter implies that the interval $C_{Plug-In}$ has better global robustness than the locally robust intervals C_c.

10 Conservative Confidence Intervals

The plug-in confidence intervals are promising, but we may be able to do better.

We now describe two properties of the coverage probability $P(\lambda)$ for general critical value functions $q(\lambda)$.

Theorem 3 *If* $q(\lambda) - \lambda \longrightarrow \kappa$ *as* $\lambda \to \infty$ *then* $P(\lambda) \longrightarrow \Phi\left(\frac{\kappa}{\sqrt{1+v^2}}\right)$ *as* $\lambda \to \infty$.

Theorem 4 *If* $q'(\lambda) \leq 1$ *for all* λ, *and* $v \geq 1$, *then* $P'(\lambda) \leq 1$ *for all* λ.

Theorem 3 is fairly remarkable. It shows that the essential condition for the asymptotic coverage probability to be bounded away from 0 is for $q(\lambda)$ to be an asymptotically (as $\lambda \to \infty$) bounded above a linear function of λ. This property is satisfied by the plug-in critical value function $q_\alpha(\lambda)$ since $q_\alpha(\lambda) - \lambda \longrightarrow \Phi^{-1}(\alpha)$. This explains why the coverage probabilities in Figure 10 appear to converge for large λ.

Theorem 4 shows that a sufficient condition for the asymptotic coverage probability to be monotonically decreasing is that the function $q(\lambda)$ has a slope less than one. This condition is also satisfied by the plug-in critical value function $q_\alpha(\lambda)$ as we can calculate that $q'_\alpha(\lambda) = (\phi(q+\lambda) - \phi(q-\lambda)) / (\phi(q+\lambda) + \phi(q-\lambda)) \leq 0$. This explains why the coverage probabilities in Figure 10 are monotonically decreasing with λ.

A confidence interval (13) constructed with a critical value function $q(\lambda)$ which satisfies the conditions of Theorems 3 and 4 (including the plug-in interval) will have the property that $P(\lambda)$ is bounded above its asymptote $\Phi\left(\kappa/\sqrt{1+v^2}\right)$. We can select κ to set this asympote equal to the nominal coverage level, hence creating a confidence interval C with conservative coverage.

Consider the following class of critical value functions.

Condition 1 $q(\lambda)$ *satisfies*

1. $q(\lambda) - \lambda \geq \kappa$ *as* $\lambda \to \infty$.
2. $\kappa = \sqrt{1+v^2}\Phi^{-1}(\alpha)$
3. $q'(\lambda) \leq 1$.

Together, they imply that we can construct a robust confidence interval..

Corollary 3 *If* $q(\lambda)$ *satisfies Condition 1, and* $v \geq 1$, *then*

$$\liminf_{n\to\infty} \Pr(\theta \in C) \geq \alpha$$

Corollary 2 shows that confidence intervals constructed using a critical value function satisfying Condition 1 are asymptotically uniformly valid – their asymptotic coverage probability uniformly exceeds the nominal level. For example, the plug-in interval $q_{\alpha^*}(\widehat{\lambda})$ with $\alpha^* = \Phi\left(\sqrt{1+v^2}\Phi^{-1}(\alpha)\right)$ satisfies this condition.

However, we can do better. Consider the class of critical value functions $q(\lambda)$ which satisfy Condition 1. Any such critical value function would yield a uniformly valid confidence interval C. The interval in this class with the smallest distortion (the coverage probability bound $P(\lambda)$ closest to the nominal level α) is achieved by the critical value function $q(\lambda)$ which is the smallest in the class. This is

$$q(\lambda) = \kappa + \lambda. \tag{15}$$

Indeed, for an alternative critical value function $q^*(\lambda)$ to have smaller distortion than $q(\lambda)$, it would have to satisfy $q^*(\lambda) < q(\lambda)$ for some λ. But this is impossible without violating Condition 1. It follows that (15) is the critical value function in the class satisfying Condition 1 with the smallest distortion.

Given the critical value $\widehat{\lambda} + \kappa$, we can write the confidence interval as

$$C_\kappa = \left\{\widehat{\theta}_K \pm s(\widehat{\theta}_K)(\widehat{\lambda}_n + \kappa)\right\}.$$

It may be helpful to observe that if $\alpha = 0.95$ and $v = 1$, then $\kappa = \sqrt{2}\Phi^{-1}(.95) = 2.33$. In this case, the critical value (15) is $\widehat{\lambda} + 2.33$. If $\widehat{\lambda}$ is small, this critical value is only slightly more conservative than the classical critical value 1.96.

In Figure 11 we plot the coverage probability of C_κ as a function of λ for for $v = 1/2$, 1, and 2. As predicted by the theory, the coverage probabilities are uniformly above the nominal level $\alpha = 0.95$, are monotonically decreasing as λ increases, and asympote to α. [Note: The theory only predicts monotonic decline for $v \geq 1$, so the observed monotonic decline for $v = 1/2$ is a pleasant surprise.] The coverage probability displays much better performance than the locally robust confidence intervals. The performance is excellent for small v (with minimal distortion) but the coverage is conservative for large v.

Returning to our simulation experiment, in Figure 12 we plot the coverage probability of the confidence intervals C_0 and C_κ pointwise as a function of the regressor z. The coverage probability of C_K is much better than C_0. For most values of z it has coverage over 0.95.

11 Confidence Interval for Non-Centrality Parameter

We next explore an alternative Bonferroni approach to construct a robust confidence interval, where we first form a confidence interval for λ, and then use this to form one for θ.

From Theorem 2, we see that the asymptotic distribution of $\widehat{\lambda}$ depends on λ and v. We can estimate v, and then construct a confidence interval for λ by test statistic inversion.

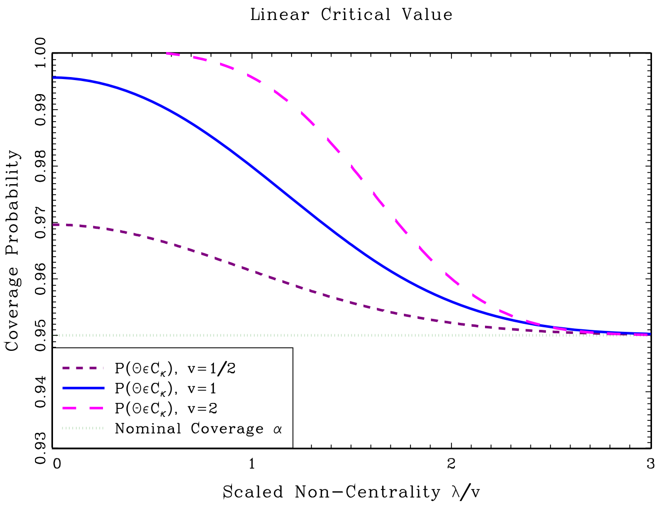

Figure 11: Linear critical value

First, to estimate v, we use

$$\widehat{v}_n = (1+\varepsilon)\sqrt{\frac{a_L'\widehat{V}_L a_L}{a_K'\widehat{V}_K a_K} - 1}$$

which is valid under our maintained assumption $\rho = 0$. Under the convergence rates of Assumption 1 and 2, $\widehat{v}_n$ is consistent for v. Hence we can treat v as if it is known.

Second, we define the λ-inverse function $\psi_\tau(x)$ of the non-central chi-square distribution. For any x for which $F(x,0) \geq \tau$, we define $\psi_\tau(x)$ as the solution to the equation

$$F(x, \psi_\tau(x)) = \tau, \tag{16}$$

(or equivalently, solves $x = q_\tau\left(\psi_\tau(x)\right)$) and for values of x such that $F(x,0) < \tau$ (equivalently, if $x \leq q_\tau(0)$) we define $\psi_\tau(x) = 0$. Notice that $\psi_\tau(x)$ is continuous and weakly increasing in x.

Computational side note: We are unaware of a hard code for $\psi_\tau(x)$. Consequently we compute the function by numerical solution to (16).

Our τ confidence interval for λ is $\Lambda = [0, \overline{\lambda}_\tau]$, where $\overline{\lambda}_\tau = \widehat{v}_n\psi_{1-\tau}\left(\widehat{\lambda}_n/\widehat{v}_n\right)$. We can show that Λ is an asymptotic level τ confidence interval for λ.

Theorem 5 $\liminf\limits_{n\to\infty} \Pr(\lambda \in \Lambda) \geq \tau$.

Theorem 5 shows that $\Lambda = [0, \overline{\lambda}_\tau]$ can be used as an asymptotically valid confidence interval for λ. Thus while λ cannot be consistently estimated, we can still correctly assess its likely values. The upper endpoint $\overline{\lambda}_\tau$ of the interval is the "highest" likely value of λ, at confidence level τ.

Proof of Theorem 5: Since $B \geq 1$, $q_\tau\left(\lambda\right)$ is increasing in λ, the definition $\overline{\lambda}_\tau/\widehat{v} = \psi_{1-\tau}\left(\widehat{\lambda}_n/\widehat{v}_n\right)$,

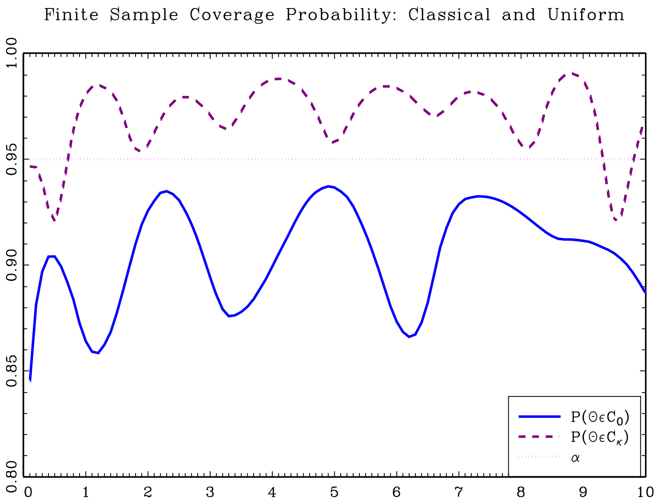

Figure 12: Finite sample coverage probability: Classical and uniform

the fact $q_{1-\tau}\left(\psi_{1-\tau}(x)\right) \geq x$, Theorem 2, and the fact that the marginal density of ξ/v is $f\left(\xi, \lambda B/v\right)$,

$$\begin{aligned}
\Pr(\lambda \in \Lambda) &= \Pr(\lambda \leq \overline{\lambda}_\tau) \\
&\geq \Pr\left(B\lambda \leq \overline{\lambda}_\tau\right) \\
&= \Pr\left(q_{1-\tau}\left(\frac{B\lambda}{\widehat{v}_n}\right) \leq q_{1-\tau}\left(\frac{\overline{\lambda}_\tau}{\widehat{v}_n}\right)\right) \\
&= \Pr\left(q_{1-\tau}\left(\frac{B\lambda}{\widehat{v}_n}\right) \leq q_{1-\tau}\left(\psi_{1-\tau}\left(\frac{\widehat{\lambda}_n}{\widehat{v}_n}\right)\right)\right) \\
&\geq \Pr\left(q_{1-\tau}\left(\frac{B\lambda}{\widehat{v}_n}\right) \leq \frac{\widehat{\lambda}_n}{\widehat{v}_n}\right) \\
&\longrightarrow \Pr\left(q_{1-\tau}\left(\frac{B\lambda}{v}\right) \leq \frac{\xi}{v}\right) \\
&= 1 - F\left(q_{1-\tau}\left(\frac{B\lambda}{v}\right), \frac{B\lambda}{v}\right) \\
&= \tau
\end{aligned}$$

as claimed.

12 Two-Stage Robust Interval

Given the confidence interval $\Lambda = [0, \overline{\lambda}_\tau]$ for λ, we can construct a Bonferroni-style one for θ. Taking the upper endpoint $\overline{\lambda}_\tau$ from Λ, we construct the critical value $q_\alpha(\overline{\lambda}_\tau)$ and set

$$C_B = \left\{\widehat{\theta}_K \pm s(\widehat{\theta}_K) q_\alpha(\overline{\lambda}_\tau)\right\}. \tag{17}$$

If we set $\tau = \alpha$, then by the Bonferroni method the interval C_B will necessarily have asymptotic uniform coverage level at least $2\alpha - 1$. In practice, such an interval will be overly conservative. We therefore consider setting τ using an alternative rule which will yield an interval with better properties.

One possibility is to set τ so that $\lim_{\lambda\to\infty} \liminf_{n\to\infty} \Pr(\theta \in C_B) \geq \alpha$. We can do this by observing that since $\overline{\lambda}_\tau = \widehat{v}_n \psi_{1-\tau}\left(\widehat{\lambda}_n/\widehat{v}_n\right)$, the critical value $q_\alpha(\overline{\lambda}_\tau)$ takes the form $q(\widehat{\lambda}_n)$ with $q(\lambda) = q_\alpha\left(\widehat{v}_n \psi_{1-\tau}\left(\lambda/\widehat{v}_n\right)\right)$. We know that as $\lambda \to \infty$, $q_\alpha(\lambda) \sim \Phi^{-1}(\alpha) + \lambda$. Similarly, $\psi_{1-\tau}(\lambda) \sim \lambda - \Phi^{-1}(1-\tau)$. Thus $q_\alpha\left(\widehat{v}_n \psi_{1-\tau}\left(\lambda/\widehat{v}_n\right)\right) \sim \lambda + \Phi^{-1}(\alpha) - \widehat{v}_n \Phi^{-1}(1-\tau)$. Theorem 3 implies that

$$\lim_{\lambda\to\infty} \liminf_{n\to\infty} \Pr(\theta \in C_B) \geq \Phi\left(\frac{\Phi^{-1}(\alpha) - v\Phi^{-1}(1-\tau)}{\sqrt{1+v^2}}\right).$$

Setting the right-hand side equal to α and solving for τ, we find

$$\tau = 1 - \Phi\left(-\frac{\left(\sqrt{1+v^2}-1\right)\Phi^{-1}(\alpha)}{v}\right). \tag{18}$$

For example, if $v = 1$ and $\alpha = 0.95$, then $\tau = 0.752$. In summary, given α and $\widehat{v}$ we calculate τ in (18), set $\overline{\lambda}_\tau = \widehat{v}\psi_{1-\tau}\left(\widehat{\lambda}_n/\widehat{v}\right)$, then $q_\alpha(\overline{\lambda}_\tau)$ and C_B as in (17). This is a confidence interval of the form (13) with $q(\lambda) = q_\alpha(\widehat{v}_n \psi_{1-\tau}\left(\widehat{\lambda}_n/\widehat{v}_n\right))$

In Figure 13 we plot the asymptotic coverage probability of C_B as a function of λ for for $v = 1/2$, 1, and 2. We see that for $v = 1/2$ the coverage appears to be uniformly above the nominal level, but not for $v = 1$ and $v = 2$. In all cases, the distortion is relatively small.

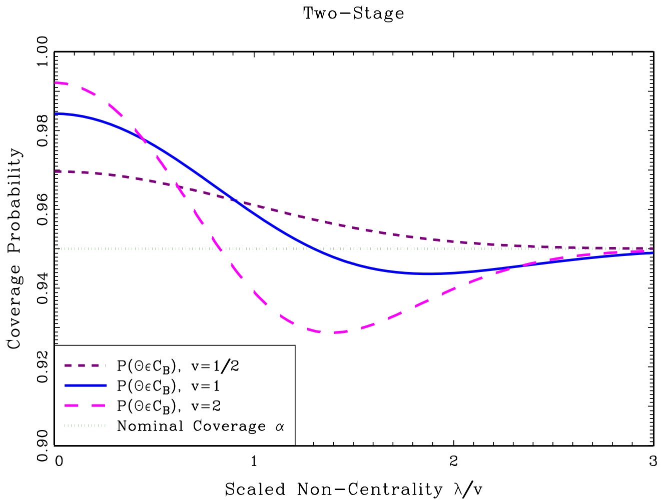

Figure 13: Two-stage

To correct the size distortion we suggest increasing the first-stage critical value τ so that the asymptotic coverage probability uniformly exceeds α. Specifically, we suggest setting $\tau = \tau(\alpha, v)$ so that

$$\inf_{\lambda} \frac{1}{v} \int_0^\infty F(q_\alpha \left(v\psi_{1-\tau} \left(\xi/v\right)\right), \lambda) f\left(\xi/v, \lambda/v\right) d\xi = \alpha$$

There is no closed form for this integral, so this needs to be done numerically. We do this as follows for $\alpha = 0.95$. We evaluated the integral as described before, for a grid of 1000 values of λ in $[0, 10]$. The coverage probability is monotonically increasing in τ, so finding the value of τ which sets the uniform coverage equal to α is straightforward. We did this for 20 values of $1/v$ ranging in the set $[.1, .2, ..., 2.0]$, or v ranging in [0.5, 10]. We set the grid on $1/v$ as it appeared the optimal value of τ is roughly linear in $1/v$. We then fit a quartic function and found the following excellent approximation (again for $\alpha = 0.95$)

$$\tau(v) = 0.948 + 0.026\frac{1}{v} - 0.393\frac{1}{v^2} - 0.247\frac{1}{v^3} - 0.047\frac{1}{v^4} \tag{19}$$

All our following calculations use this rule.

To summarize, the method is as follows. Given $\widehat{v}$, calculate τ in (19), set $\overline{\lambda}_\tau = \widehat{v}_n \psi_{1-\tau}\left(\widehat{\lambda}_n/\widehat{v}_n\right)$, then $q_\alpha(\overline{\lambda}_\tau)$ and C_B as in (17).

By construction, this interval should have (approximately) uniform coverage exceeding α.

In Figures 14, 15, and 16 we plot the asymptotic coverage probabilities of the uniform linear rule and this uniform Bonferroni rule for $v = 1/2$, 1, and 2. As predicted by the theory, both confidence intervals have approximate uniform coverage exceeding 0.95. For $v = 1/2$, the two methods are nearly identical, with minimal distortion. For $v = 1$ there is some difference between the methods. For $\lambda/v < 2$ the Bonferroni rule has less distortion, but for large λ/v the rankings are reversed. The contrast is greater for $v = 2$, where the linear rule has particularly high distortion for small λ/v while the Bonferroni rule yields much less distortion. While neither method dominates the other, we believe that the uniform Bonferroni method has lower overall distortion.

In Figure 17 we plot the asymptotic power of the robust test of H_0. This plot uses the same framework as for Figures 2 and 3, specifically we plot the asymptotic power of nominal 5% tests, setting $\phi = 1$, $\gamma = 2$, $D = 1$, $A = 1$, $\tau_1 = \left(4A^2\right)^{1/3}$ and $R = 2$. We plot the power (rejection probability) as a function of the localizing parameter $\overline{\delta}$. We compare three tests. First, the uniform linear rule with $\varepsilon = 1/3$. (In this experiment the uniform linear and Bonferroni rules are near equivalent so only the first is plotted). Second, the classical test without size correction. Third, the classical test with K equal twice the MSE optimal; we call this the undersmoothing test. The robust test is the only of the three tests with theoretically correct size control, though the undersmoothing test has minimal distortion with Type I error near 0.05. Since it is difficult to compare power when there is size distortion, in Figure 18 we plot the asymptotic power of the same tests, but using the infeasible correct critical values for the classical test and the undersmoothing test. Thus for these tests their Type I error is exactly 0.05 at the origin, while the robust test is conservative. Figure 18 suggests that the robust test is implementing a size-correction for the classical test, as the two power curves are very similar. The robust and undersmoothing tests have different power curves, with the robust test having higher power for negative values of $\overline{\delta}$ and conversely for positive values, though for the latter the two power curves are similar for rejection rates near one-half.

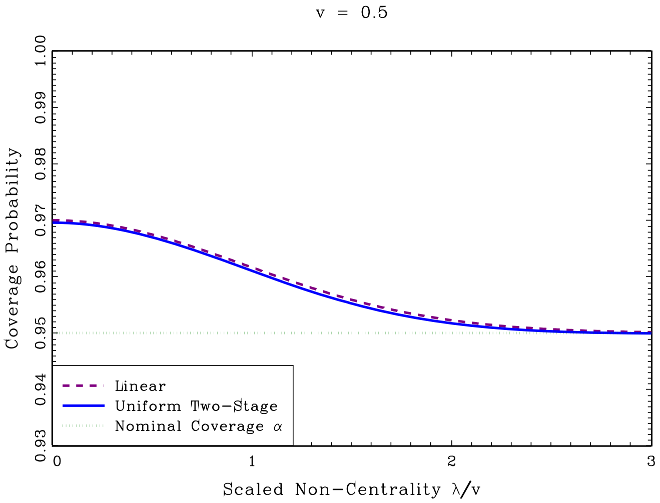

Figure 14: $v = 0.5$

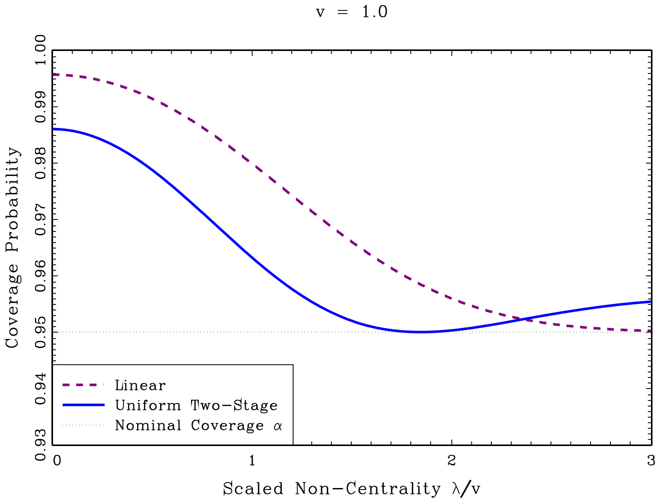

Figure 15: $v = 1.0$

Returning to our simulation experiment, in Figure 18 we plot the coverage probability of the confidence intervals C_κ, C_{B1} and C_{B2} pointwise as a function of the regressor z. We see that the C_{B1} and C_{B2} intervals are quite similar, and have reduced distortion relative to $C\kappa$

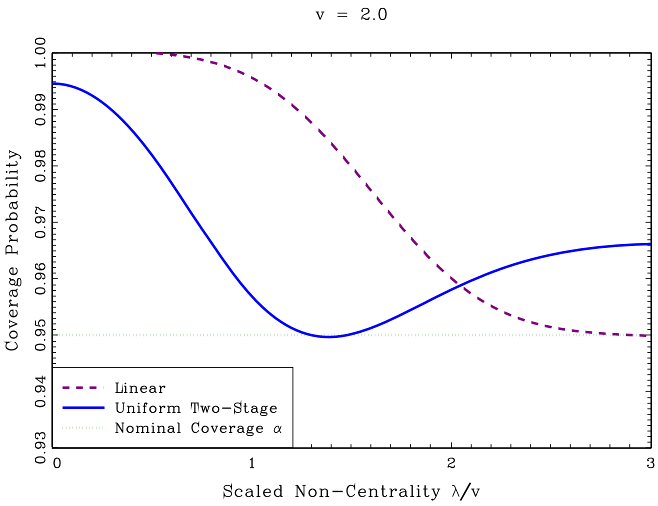

Figure 16: $v = 2.0$

13 Illustration

We illustrate the method by returning to the single sample discussed in Section 4. In Figures 20, 21, and 22 we display the regression estimate, classical confidence intervals, and robust confidence bands. Figure 20 is for the estimates using a quadratic regression, Figure 21 for the estimates using a quadratic spline with one knot, and Figure 22 for the estimates using a quadratic spline with two knots. In each figures, the solid line is the true regression function $g(z)$, the short dashes is the spline estimate $\widehat{g}_K(z)$, the dotted lines are the classic 95% confidence bands $\widehat{g}_K(z) \pm 1.96s(\widehat{g}_K)$, and the long dashes are the C_B confidence bands $\widehat{g}_K(z) \pm q_\alpha(\overline{\lambda}_\tau)s(\widehat{g}_K)$ with τ set by (19). The robust confidence bands are constructed pointwise in z, with $\widehat{\lambda}_n$ and $\widehat{v}_n$ calculated separately for each z. The noncentrality estimate is calculated using a spline with five equally spaced knots.

You can see in Figures 20 and 21 that while the classical confidence intervals fail to cover the true regression function, the robust confidence intervals do so. The robust intervals are much wider, to account for the estimation bias.

Figure 22 is a bit different. In this case the robust and classical intervals are quite similar, though where they deviate the robust intervals are more successful at covering the true regression line. In all cases, the robust intervals adapt to the misspecification.

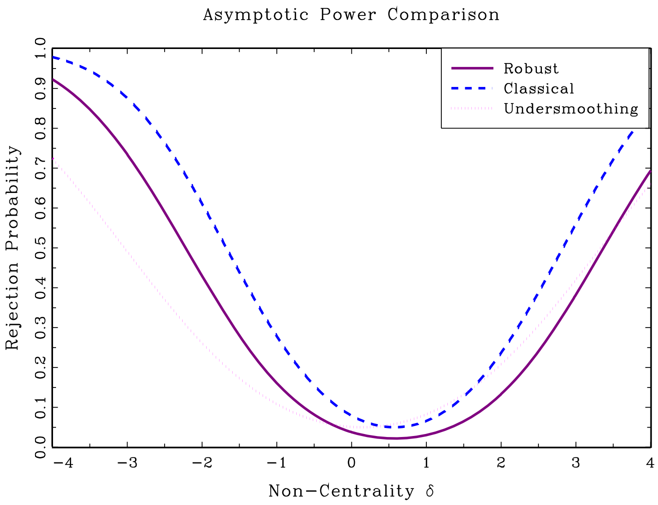

Figure 17: Asymptotic power comparisons

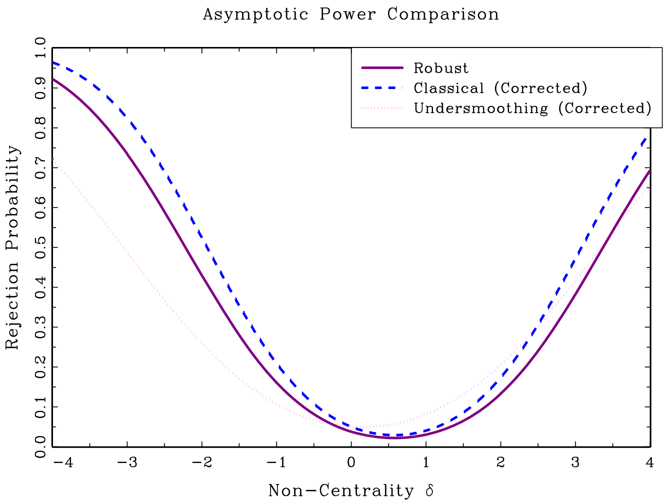

Figure 18: Size-Corrected Power

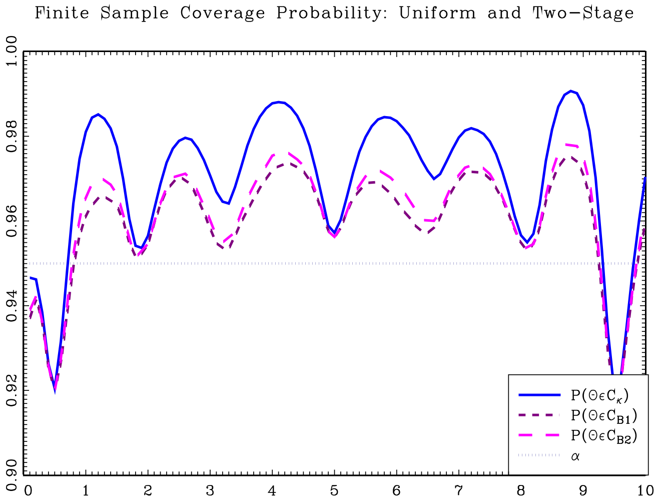

Figure 19: Finite sample coverage probability: Uniform and two-stage

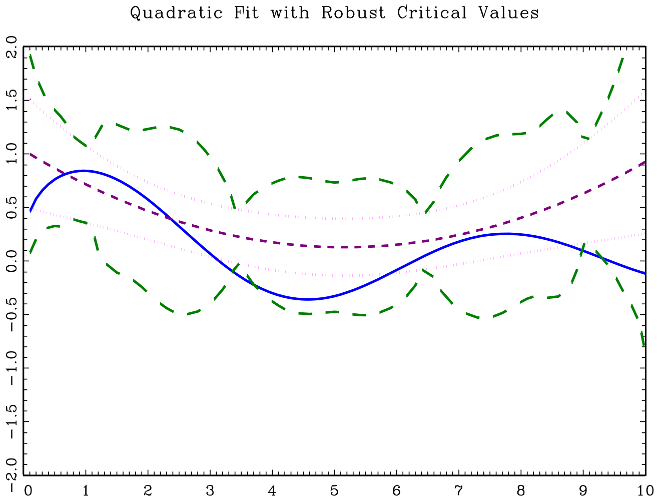

Figure 20: Quadatic fit with robust critical values

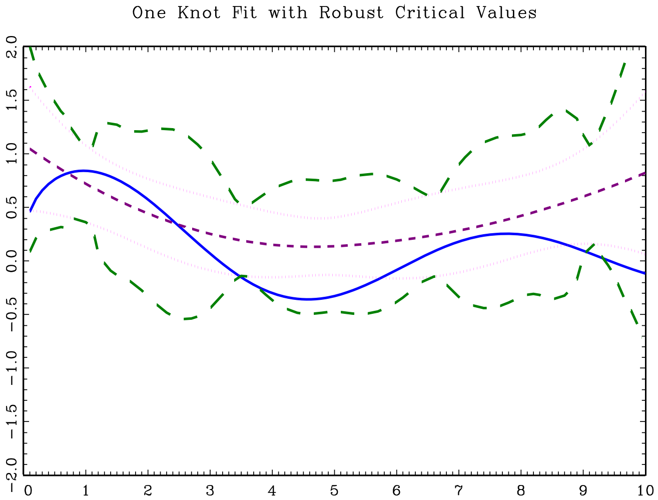

Figure 21: One knot fit with robust critical values

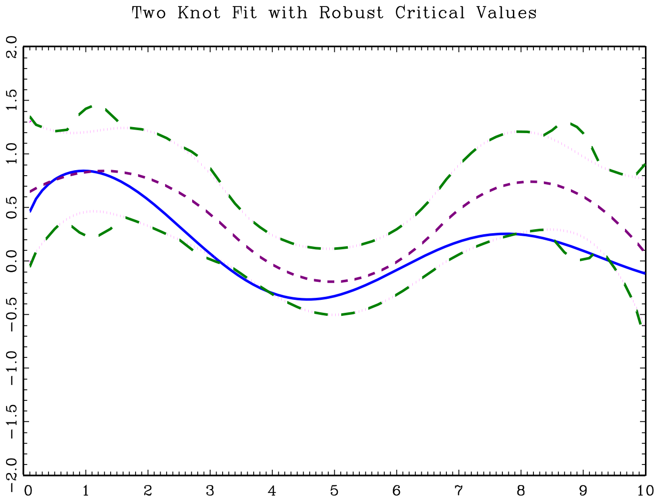

Figure 22: Two knot fit with robust critical values