Marginal Treatment Effects with Many Instruments

Matias D. Cattaneo, University of Michigan

Michael Jansson, UC-Berkeley and CREATES

Xinwei Ma, University of Michigan

July 2016

Draft prepared for NBER conference is here:

http://www.umich.edu/~cattaneo/papers/Cattaneo-Jansson-Ma_2016_MTE.pdf

Overview

What this paper does:

1. Study estimation and inference for MTE and functionals thereof.
2. Develop new large-sample point estimation and distributional approximations allowing for many (included) instruments entering the probability of selection.
3. Results encompass standard ones, so we provide strict asymptotic improvements.
4. New estimation/inference methods are fully automatic (jackknife + wild bootstrap).
5. Results extend to a large class of two-step M-estimators with high-dimensional first step (e.g., propensity score weighting, generated regressors, etc.).

What this paper does not do:

1. Discuss identification or interpretation of MTE, and functional thereof, with "few" or "many" instruments.
   - Our model/estimand is taken as given based on results in the literature.
2. Discuss model selection and/or ultra-high-dimensional methods for selecting in or out instruments.
   - Our methods are high-dimensional in the sense that the number of included instruments may be "large" (relative to the sample size).

Setup and Estimation in Practice

- Random sample of units $i = 1, 2, \ldots, n$ from large population:

$$y_i = t_i y_i(1) + (1 - t_i) y_i(0), \qquad t_i = \mathbb{1}[p_i \geq v_i], \qquad v_i | \mathbf{x}_i \sim \text{Uniform}[0, 1].$$

- Parameter of interest:

$$\tau_{\texttt{mte}} = \tau_{\texttt{mte}}(a|\mathbf{x}_i) = \mathbb{E}[y_i(1) - y_i(0)|v_i = a, \mathbf{x}_i] = \frac{\partial}{\partial a}\mathbb{E}[y_i|p_i = a, \mathbf{x}_i]$$

  - Identification assumption: $(p_i, \mathbf{z}_i') \perp\!\!\!\perp (y_i(1), y_i(0), v_i) \mid \mathbf{x}_i$
  - Practical assumption: $\mathbb{E}[y_i|p_i = a, \mathbf{x}_i] = m(\mathbf{x}_i, a, \boldsymbol{\theta}_0)$

- Estimation approach:

  1. Estimate propensity score using possibly "many" instruments $\mathbf{z}_i \in \mathbb{R}^k$:

$$\hat{p}_{n,i} = \mathbf{z}_i'\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (t_i - \mathbf{z}_i'\boldsymbol{\beta})^2,$$

  2. Estimate MTE:

$$\hat{\tau}_{\texttt{mte}} = \hat{\tau}_{\texttt{mte}}(a|\mathbf{x}) = \frac{\partial}{\partial a} m(\mathbf{x}, a, \hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i, \hat{p}_{n,i}, \boldsymbol{\theta}))^2.$$

Overview of Results

1. **When k is "large" relative to n**, first-order bias in distributional approximation:

$$\sqrt{n}(\hat{\tau}_{\texttt{mte}} - \tau_{\texttt{mte}}) = \frac{k}{\sqrt{n}} B + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_{\mathbb{P}}(1) \to_d \mathcal{N}(\rho B, V)$$

   where

$$\frac{k}{\sqrt{n}} \to \rho \in [0, \infty)$$

2. **Jackknife** bias and variance estimation works, even when $\rho > 0$:

$$\hat{B}_{\texttt{Jack}} - B = o_{\mathbb{P}}(1) \qquad \textit{and} \qquad \hat{V}_{\texttt{Jack}} - V = o_{\mathbb{P}}(1)$$

3. **Wild bootstrap** works after jackknife estimation, even when $\rho > 0$:

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}^* \left[\frac{\sqrt{n} \left(\hat{\tau}^*_{\texttt{mte}} - \hat{\tau}_{\texttt{mte}} - \frac{k}{n} (\hat{B}^*_{\texttt{Jack}} - \hat{B}_{\texttt{Jack}}) \right)}{\sqrt{\hat{V}^*_{\texttt{Jack}}}} \leq t \right] - \mathbb{P} \left[\frac{\sqrt{n} \left(\hat{\tau}_{\texttt{mte}} - \tau_{\texttt{mte}} - \frac{k}{n} \hat{B}_{\texttt{Jack}} \right)}{\sqrt{\hat{V}_{\texttt{Jack}}}} \leq t \right] \right| = o_{\mathbb{P}}(1)$$

4. Results extend to functionals of $\hat{\tau}_{\texttt{mte}} = \hat{\tau}_{\texttt{mte}}(a|\mathbf{x})$, such as those used for extrapolation of treatment effects or to construct other treatment effects.

Motivation and Discussion: Simulations

- Standard $(1-\alpha)\%$ confidence intervals for $\tau_{\texttt{mte}}$:

$$\left[\ \hat{\tau}_{\texttt{mte}} - \Phi^{-1}_{1-\alpha/2} \cdot \sqrt{\frac{\hat{V}}{n}} \quad , \quad \hat{\tau}_{\texttt{mte}} - \Phi^{-1}_{\alpha/2} \cdot \sqrt{\frac{\hat{V}}{n}}\ \right]$$

  where:

  - Φ^{-1}_{α} is standard normal quantile (e.g., $\Phi^{-1}_{1-\alpha/2} = \Phi^{-1}_{\alpha/2} = 1.65$ if $\alpha = 0.90$).
  - $\hat{V}/n$ is some standard error estimator... Because V is complex and involves may unknown quantities, automatic methods such as $\hat{V}_{\texttt{Jack}}$ are preferred (if valid).

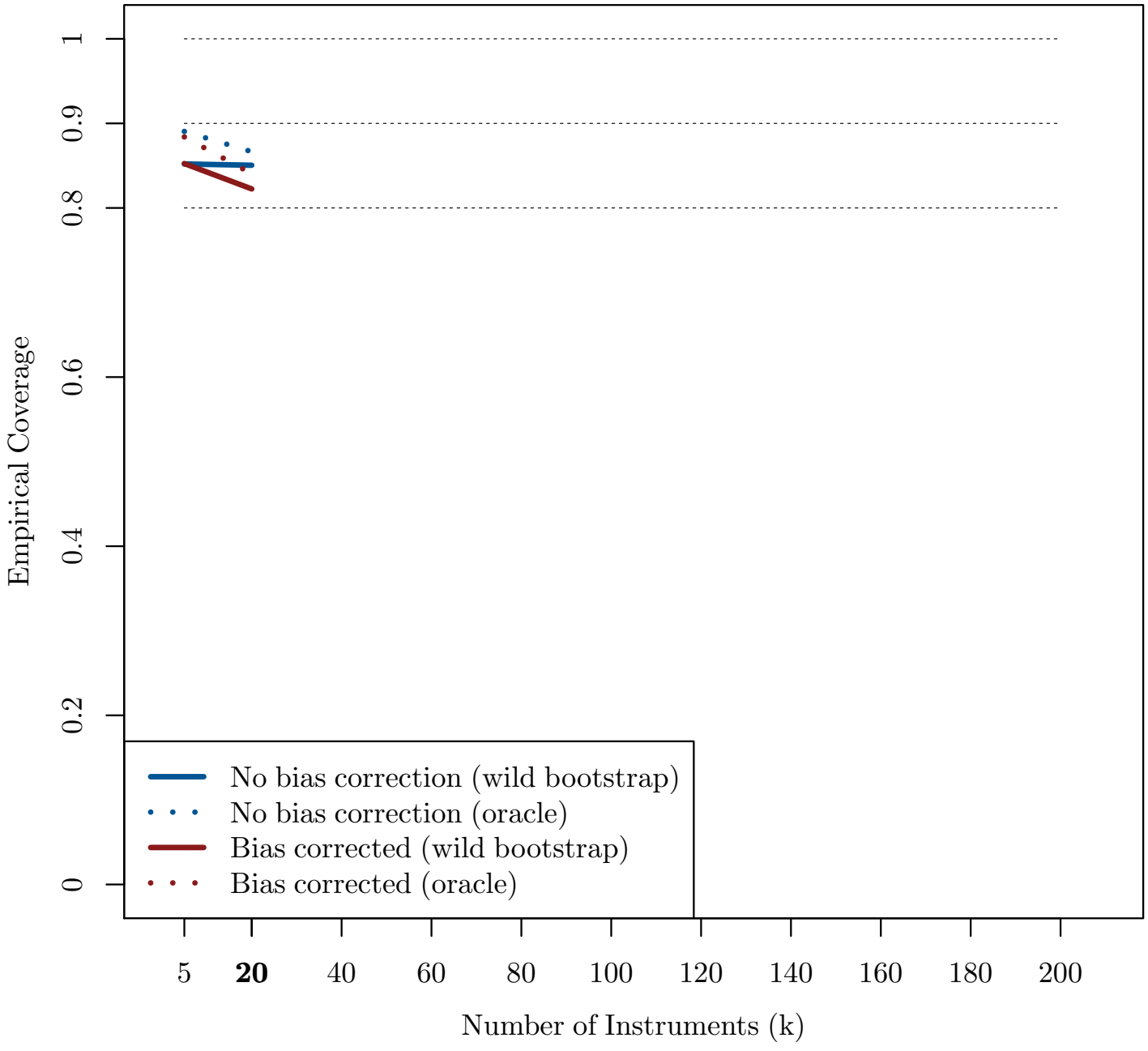
- Our proposed method: $(1-\alpha)\%$ confidence intervals for $\tau_{\texttt{mte}}$:

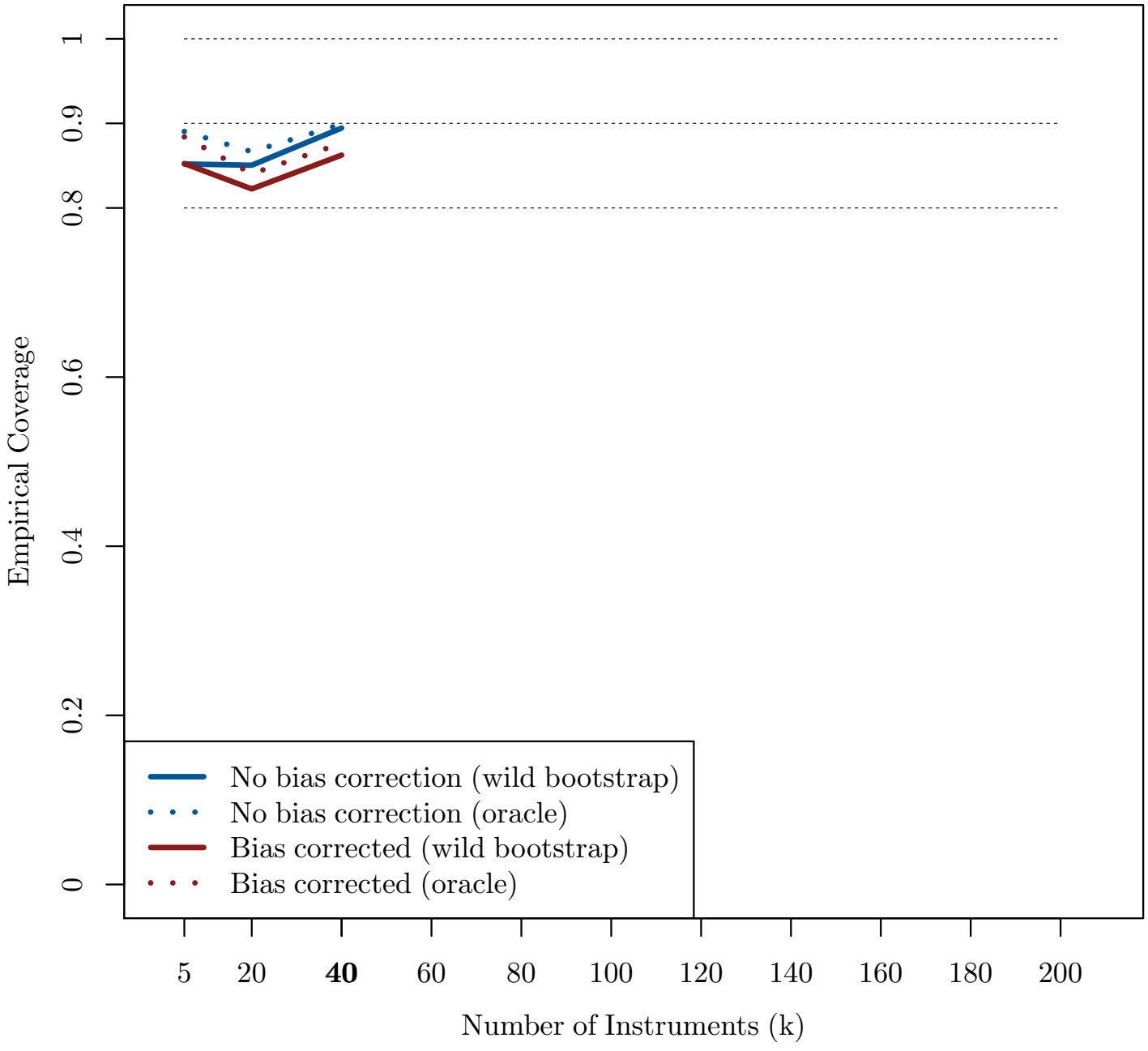
$$\left[\ \left(\hat{\tau}_{\texttt{mte}} - \frac{k}{n}\hat{B}_{\texttt{Jack}}\right) - q^*_{1-\alpha/2} \cdot \sqrt{\frac{\hat{V}_{\texttt{Jack}}}{n}} \quad , \quad \left(\hat{\tau}_{\texttt{mte}} - \frac{k}{n}\hat{B}_{\texttt{Jack}}\right) - q^*_{\alpha/2} \cdot \sqrt{\frac{\hat{V}_{\texttt{Jack}}}{n}}\ \right]$$

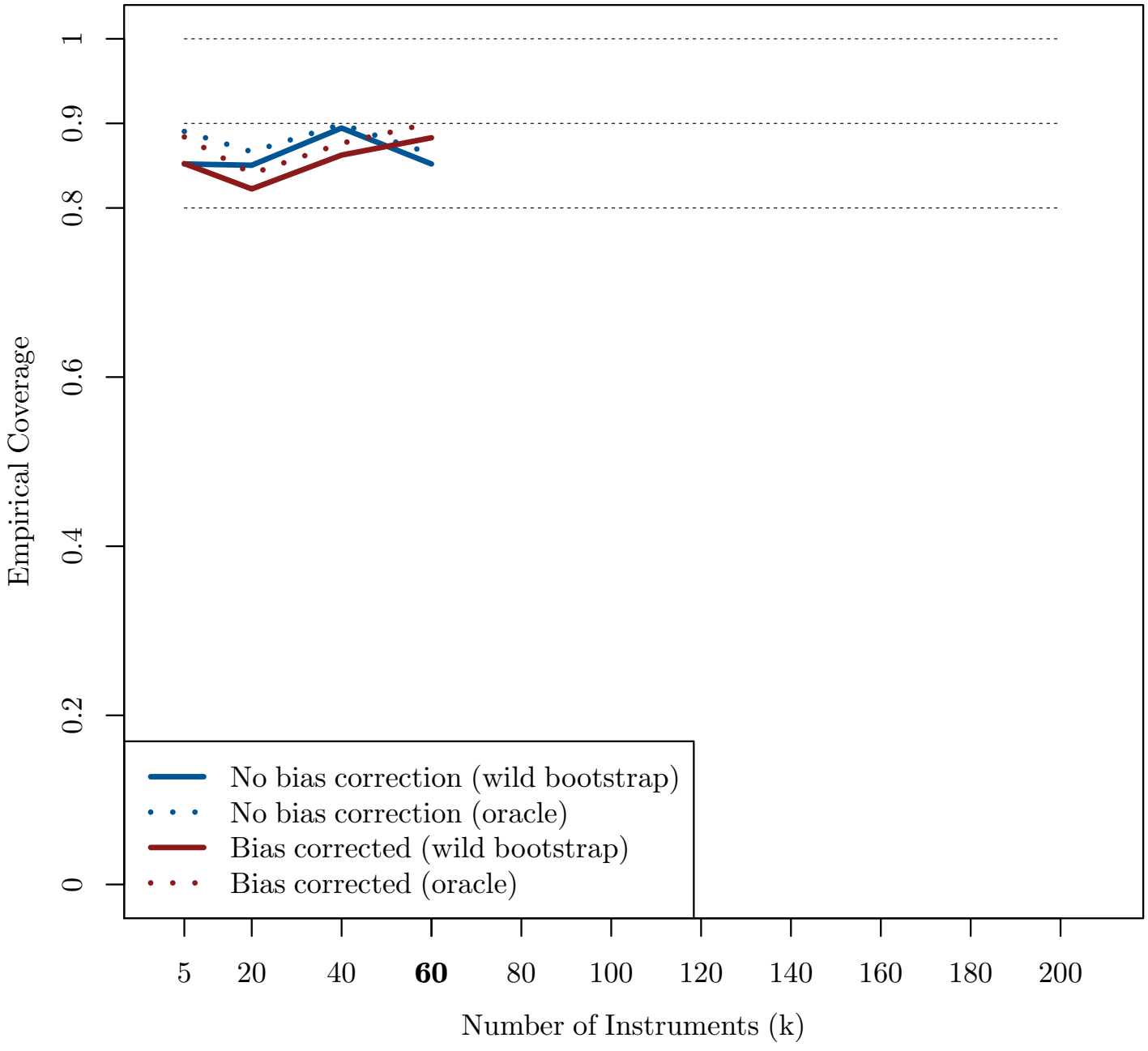
  where:

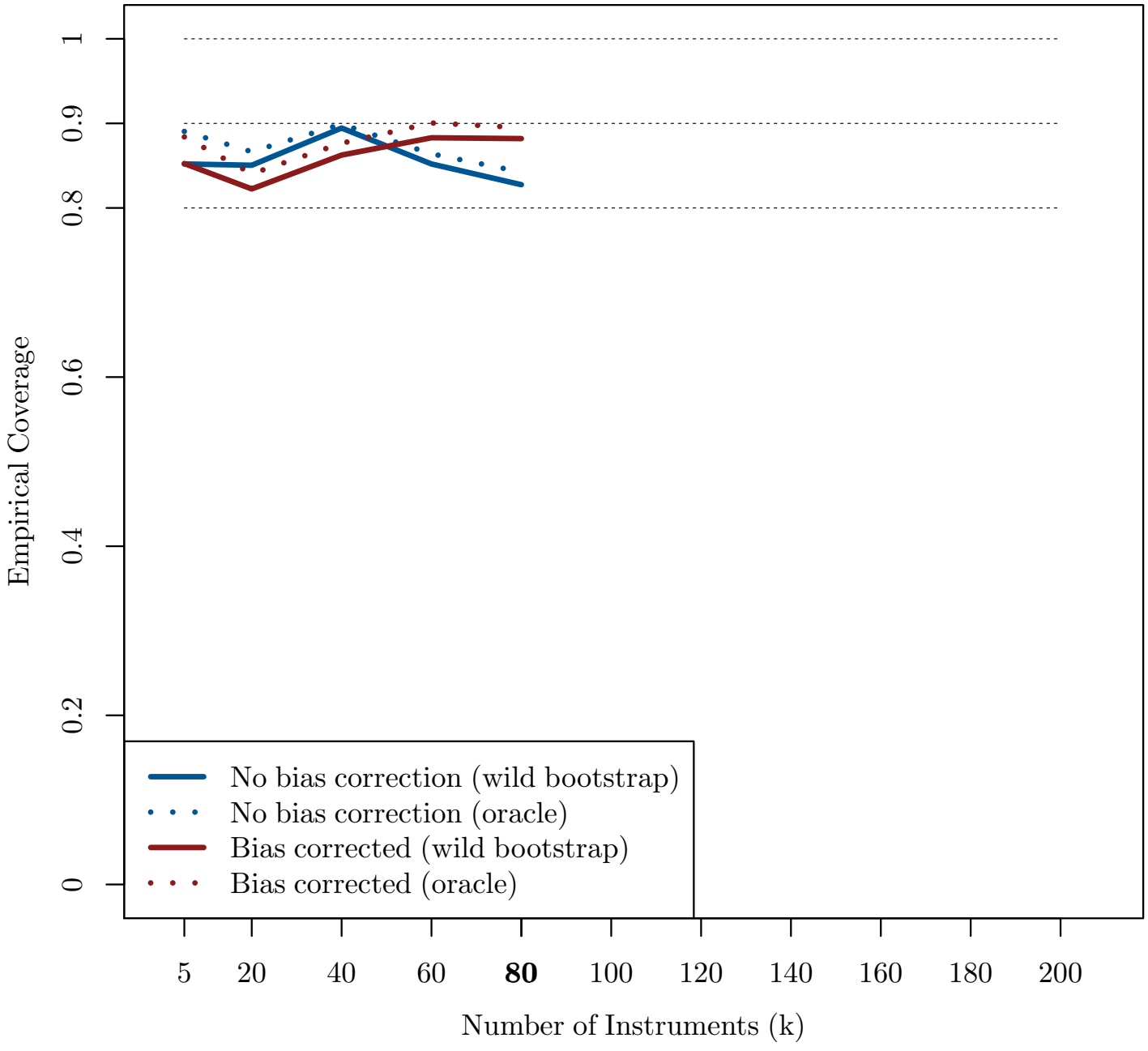
  - $\hat{B}_{\texttt{Jack}}$ and $\hat{V}_{\texttt{Jack}}$ jackknife estimators, shown consistent even when $\frac{k}{\sqrt{n}} \not\to 0$.
  - q^*_{α} is quantile from wild bootstrap distribution of jackknife-based t-test statistic:

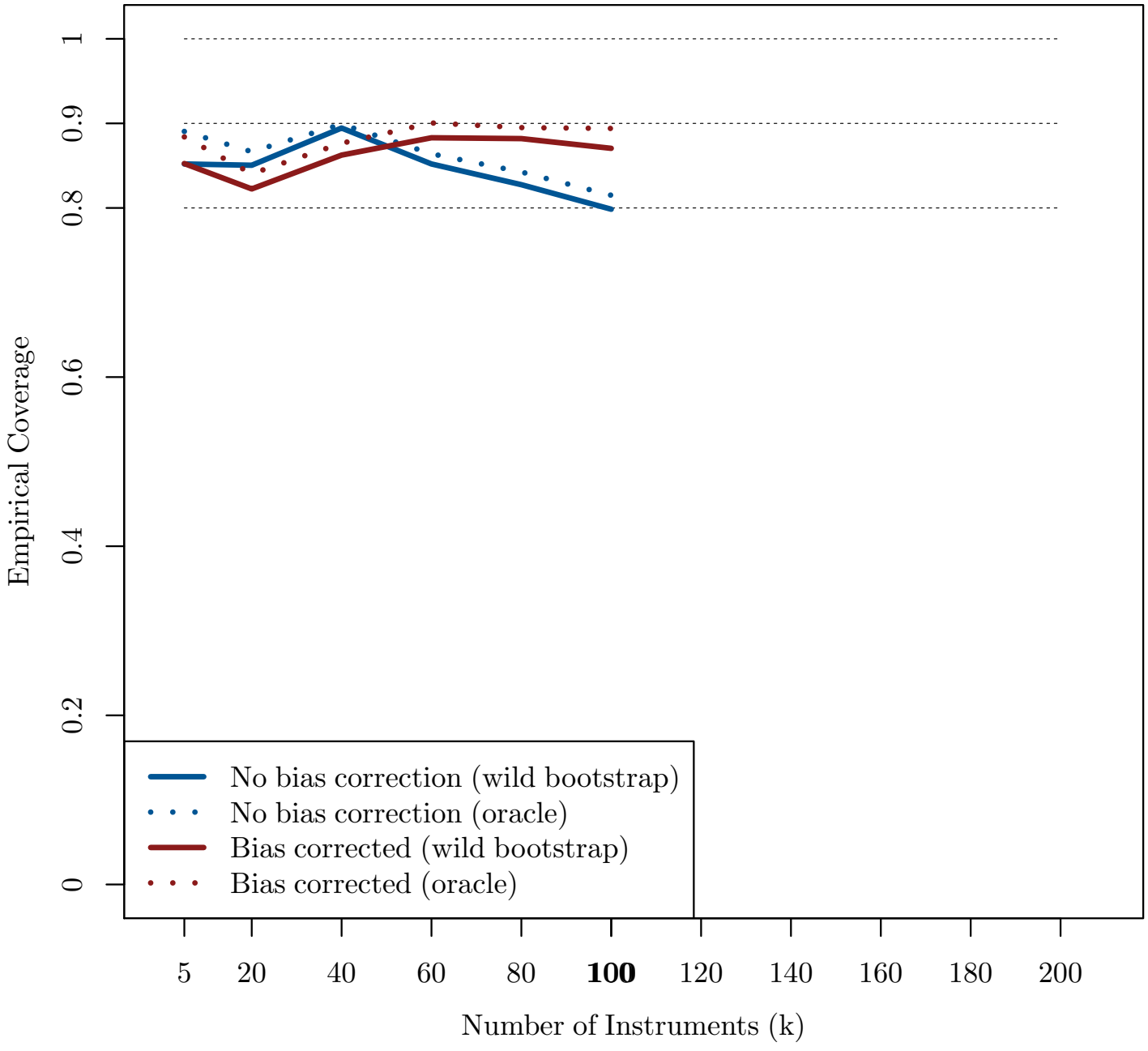
$$T^* = \frac{\sqrt{n}\left(\hat{\tau}^*_{\texttt{mte}} - \hat{\tau}_{\texttt{mte}} - \frac{k}{n}(\hat{B}^*_{\texttt{Jack}} - \hat{B}_{\texttt{Jack}})\right)}{\sqrt{\hat{V}^*_{\texttt{Jack}}}}$$

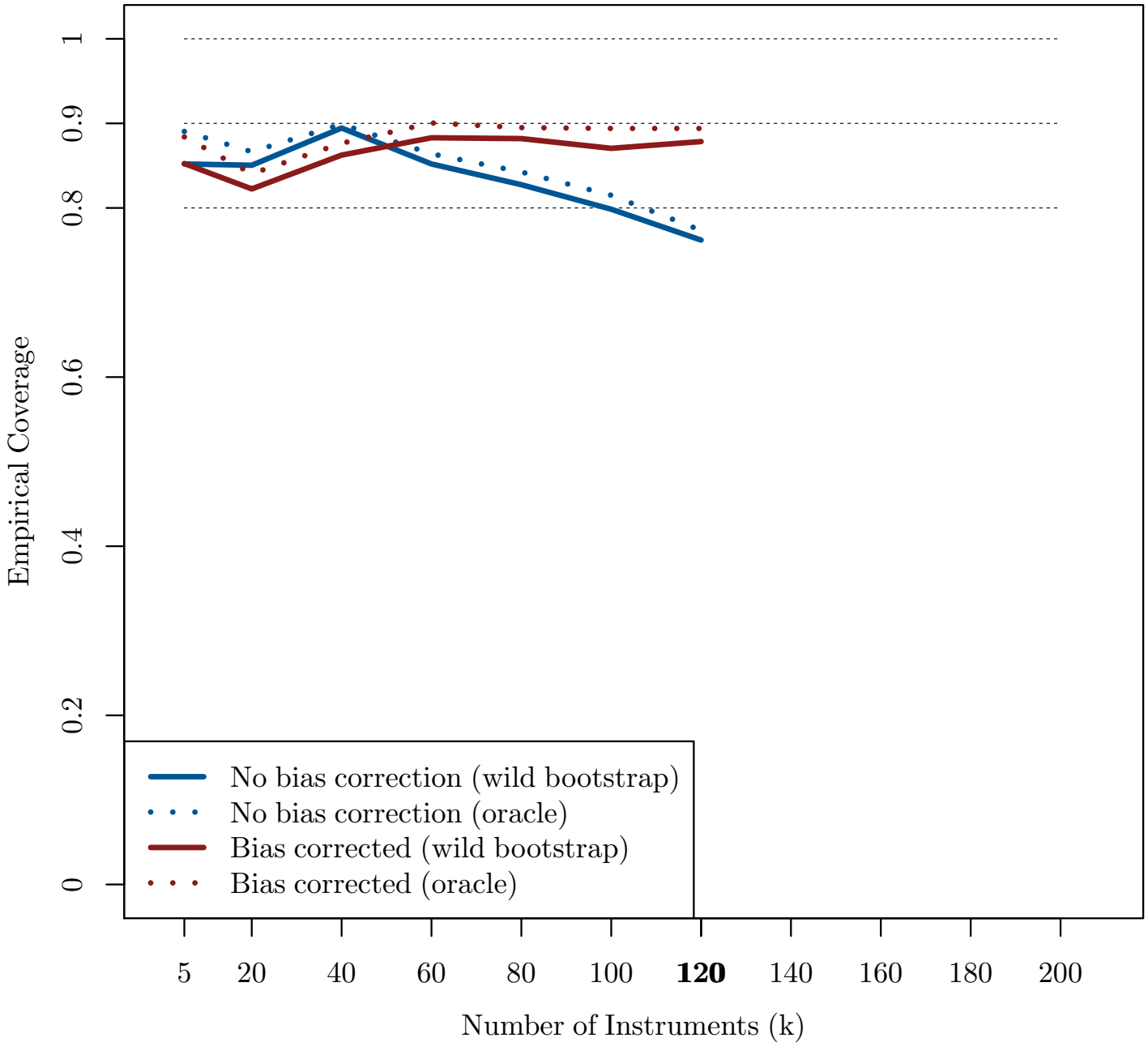

Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)

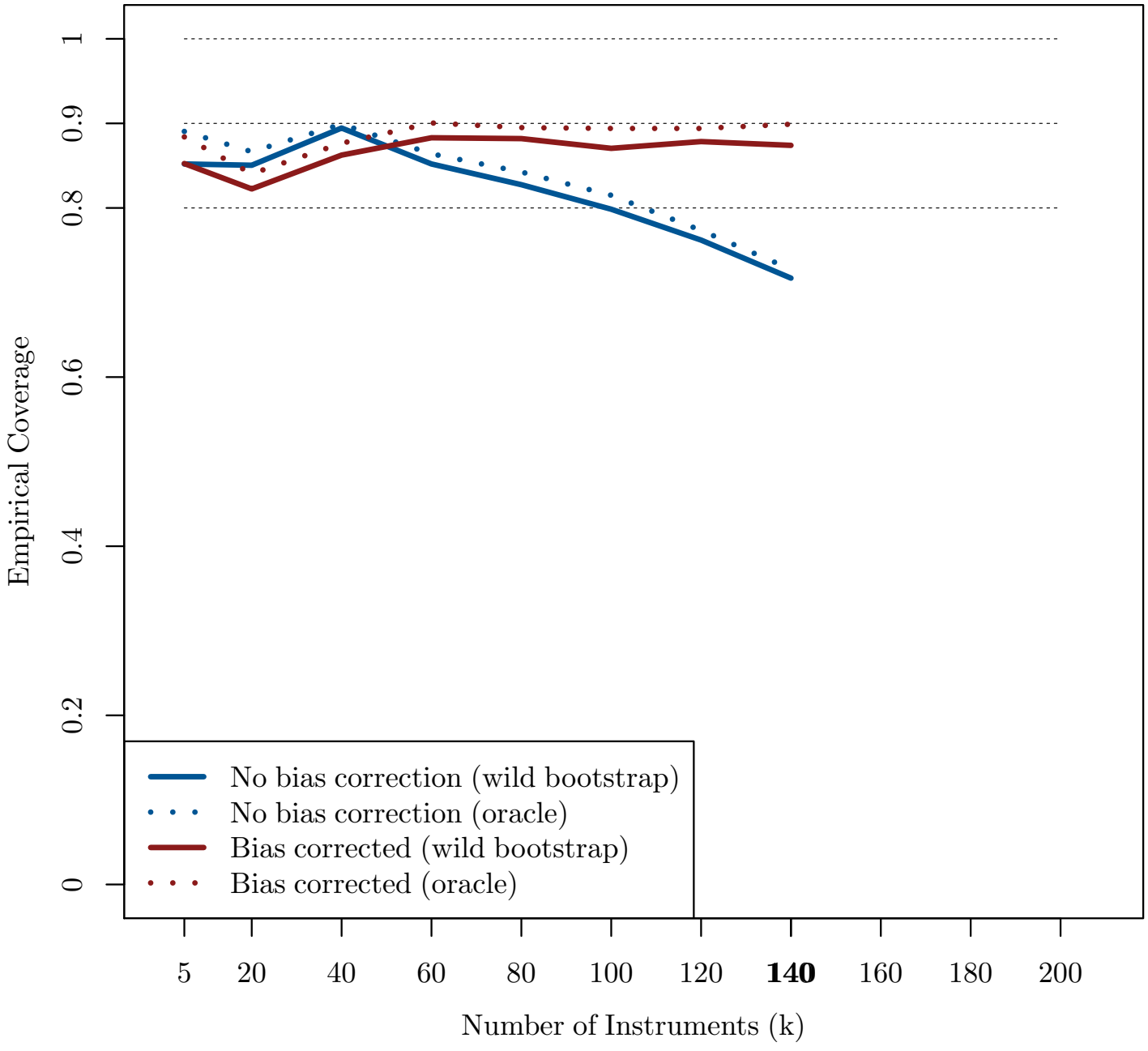

Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)

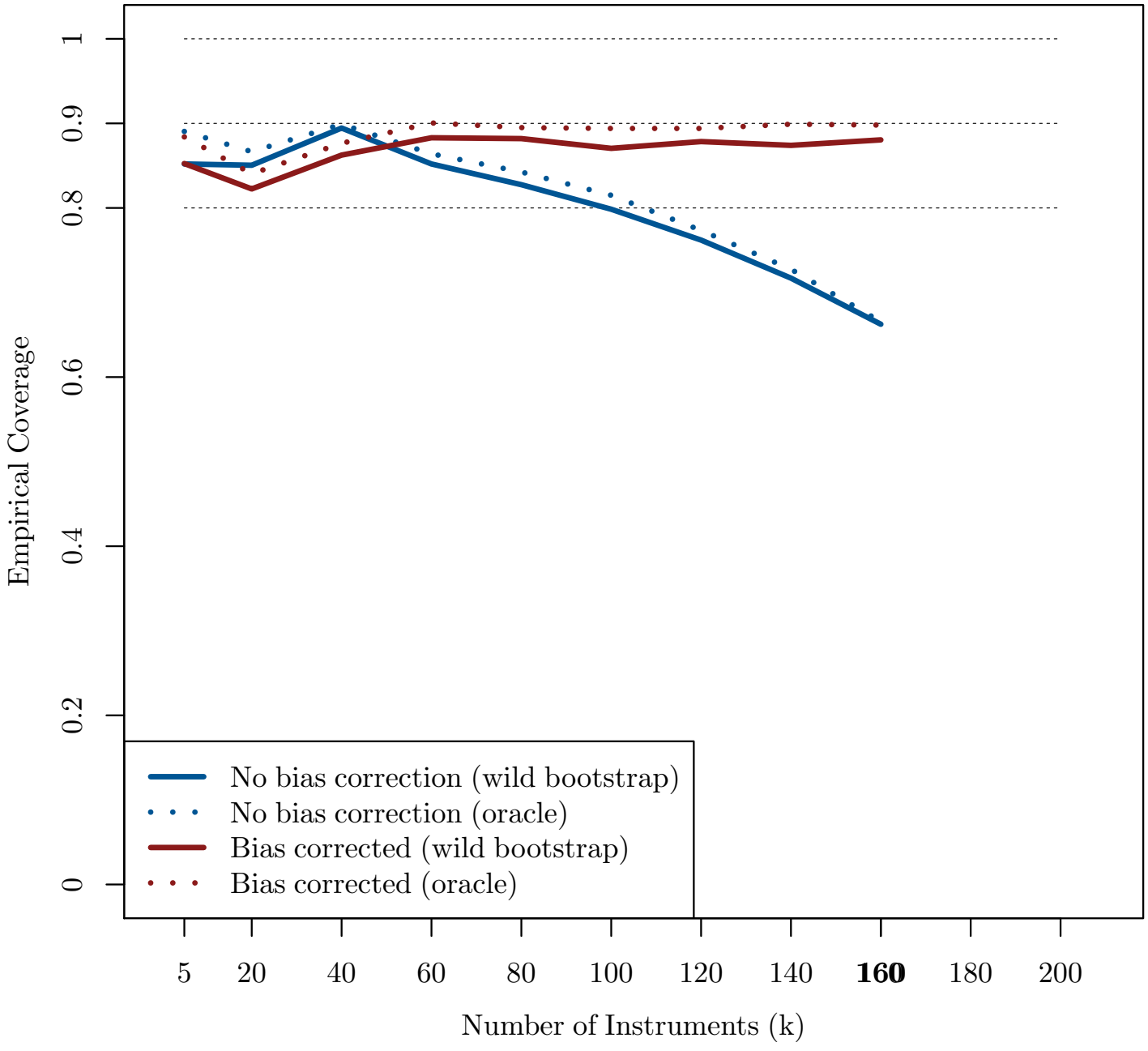

Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)

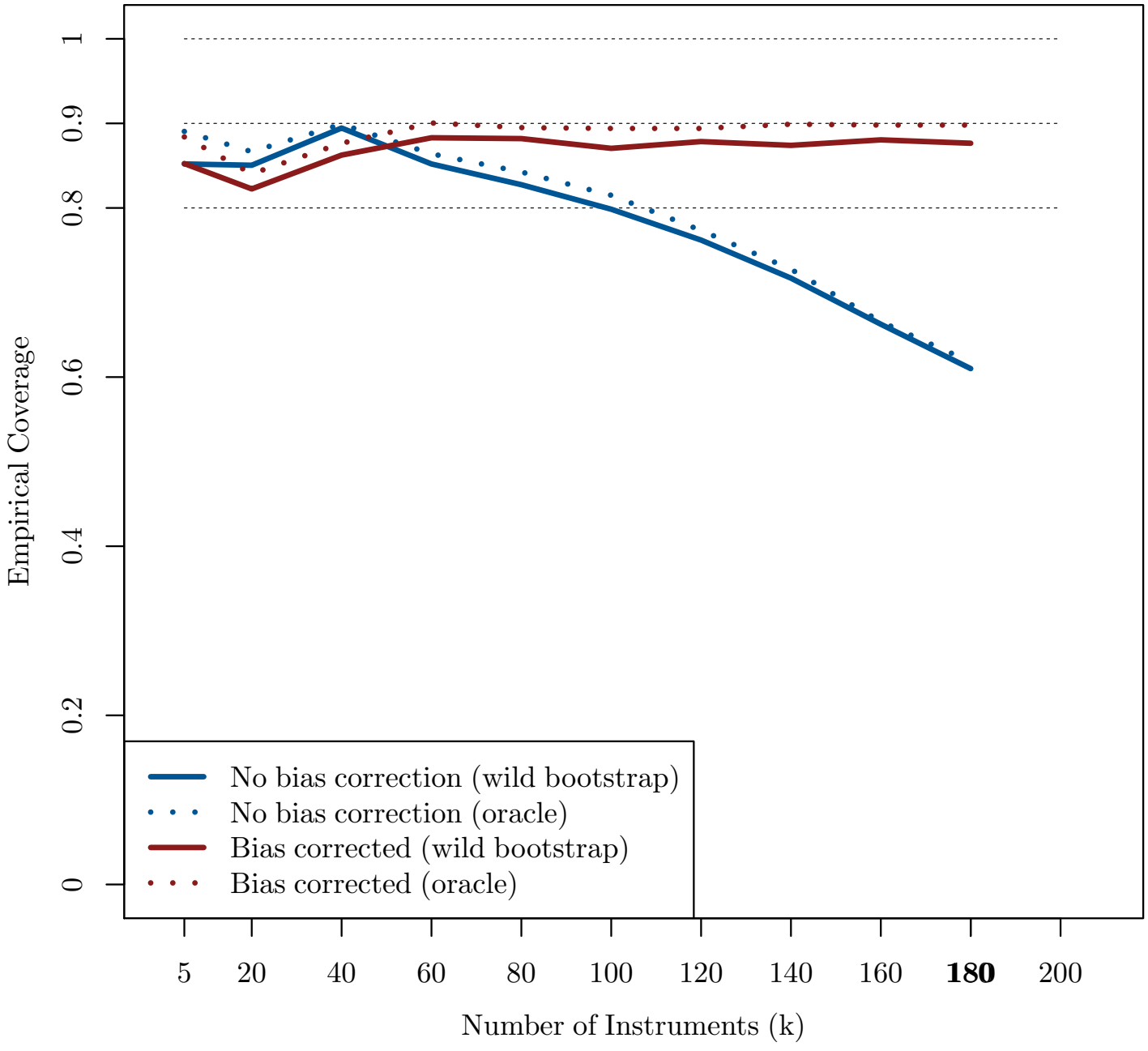

Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)

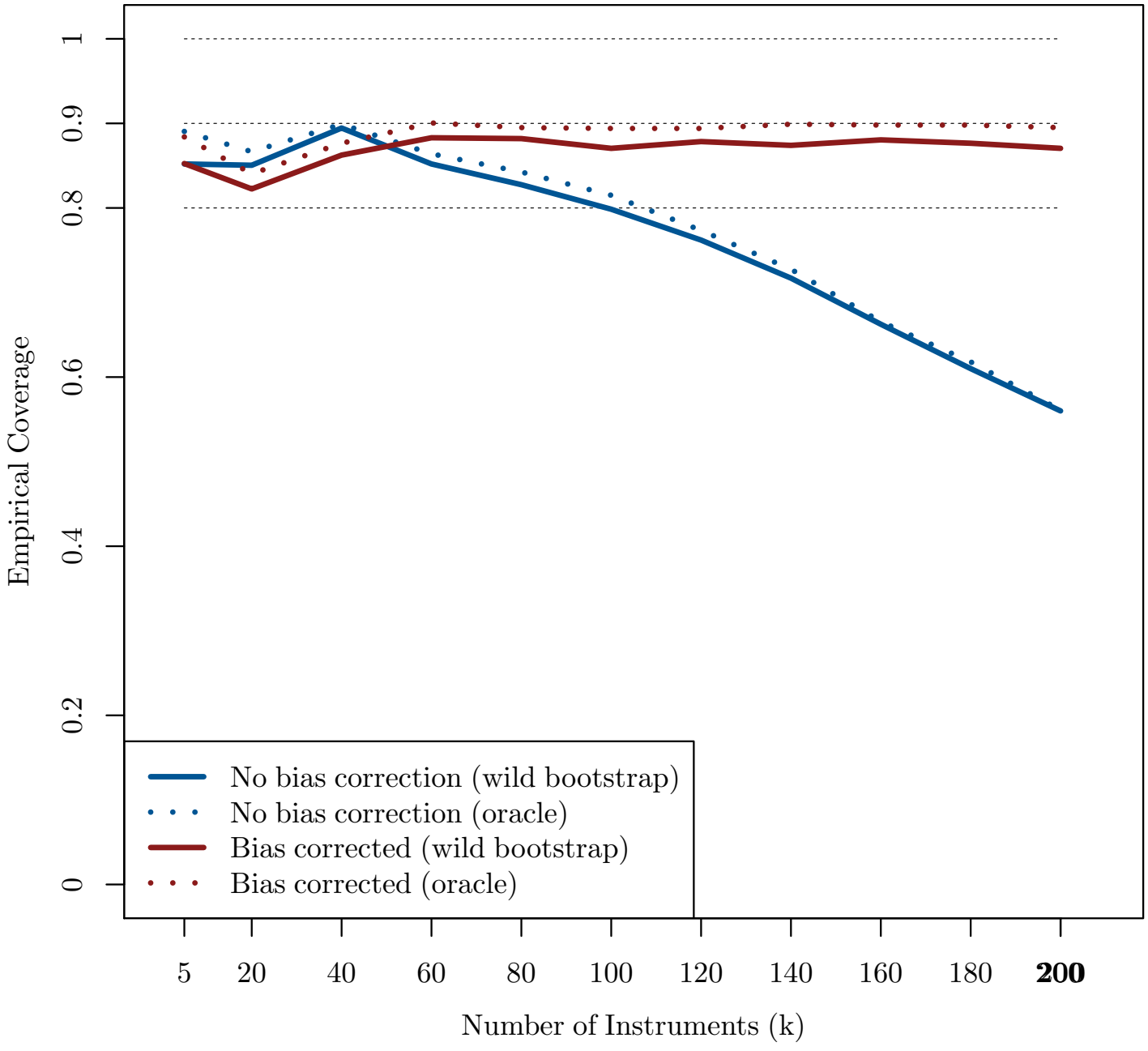

Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
5
20
40
60
80
100
120
140
160
180
200
Number of Instruments (k)

Intuition Behind Results

- Recall that

$$\hat{\tau}_{\texttt{mte}} = \frac{\partial}{\partial a} m(\mathbf{x}, a, \hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i, \hat{p}_{n,i}, \boldsymbol{\theta}))^2,$$

$$\hat{p}_{n,i} = \mathbf{z}_i' \hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (t_i - \mathbf{z}_i' \boldsymbol{\beta})^2,$$

- $\hat{\tau}_{\texttt{mte}} \to_{\mathbb{P}} \tau_{\texttt{mte}}$, even when $\frac{k}{\sqrt{n}} \not\to 0$, if $\max_{1 \le i \le n} |\hat{p}_{n,i} - p_i| = o_{\mathbb{P}}(1)$.

- Distribution theory is about "local" behavior:

$$\begin{aligned}
&\sqrt{n}(\hat{\tau}_{\texttt{mte}} - \tau_{\texttt{mte}}) \\
&\approx \underset{\downarrow}{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(\mathbf{x}_i, p_i)} + \underset{\downarrow}{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial g(\mathbf{x}_i, p_i)}{\partial p_i} \cdot [\hat{p}_{n,i} - p_i]} + \underset{\downarrow}{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial^2 g(\mathbf{x}_i, p_i)}{\partial p_i^2} \cdot [\hat{p}_{n,i} - p_i]^2} \\
&\qquad \mathcal{N}(0, V_1) \qquad\qquad \mathcal{N}(0, V_2) + \tfrac{k}{\sqrt{n}} B_1 \qquad\qquad \tfrac{k}{\sqrt{n}} B_2
\end{aligned}$$

Application to Treatment Effect Extrapolation

- Large class of treatment effects can be recovered from MTE:

$$\tau_\omega(\mathbf{x}) = \int_0^1 \tau_{\texttt{mte}}(a|\mathbf{x})\omega(a|\mathbf{x})da$$

- (Policy, Extrapolation, etc.) Estimator:

$$\hat{\tau}_\omega = \hat{\tau}_\omega(\mathbf{x}) = \int_0^1 \hat{\tau}_{\texttt{mte}}(a|\mathbf{x})\omega(a|\mathbf{x})da$$

- Our results apply directly (using the "delta method"):

$$\sqrt{n}(\hat{\tau}_\omega - \tau_\omega) = \frac{k}{\sqrt{n}}B_\omega + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i + o_{\mathbb{P}}(1) \to_d \mathcal{N}(\rho B_\omega, V_\omega)$$

  where

$$\frac{k}{\sqrt{n}} \to \rho \in [0, \infty)$$

- **Jackknife** bias and variance estimation works, even when $\rho > 0$.

- **Wild bootstrap** works after jackknife estimation, even when $\rho > 0$.

(Backup) Empirical Illustration: Card (1993)

- We try out a preliminary (backup) empirical application, using Card (1993)'s returns to college education paper.

- We will soon have access to the NLSY Geocode data files, to revisit Carneiro, Heckman and Vytlacil (2011)'s approach to estimate marginal returns to education, with a much richer set of raw and expanded instruments.

- Card (1993) Empirical Illustration:
  - Subsample of National Longitudinal Survey of Young Men (NLSYM).
  - Initial survey 1966, sample survey 1976; for example, $n = 2,307$ white males.
  - Treatment: $t_i = \mathbb{1}(\text{education in } 1976 \geq 13)$.
  - Consider specifications from 25 to 59 instruments.
  - Consider 2SLS, ATE, ATE with bias-correction.
  - Excluded instrument: college proximity measures.
  - **Main empirical finding**: results are robust to "many instruments" bias.

Table: Returns to College Education, Whites Subsample, $n = 2,307$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
2SLS	0.125	0.129	0.138	0.131	0.136	0.143	0.144	0.170	0.186	0.192	0.177	0.186
	(0.018)	(0.017)	(0.016)	(0.017)	(0.016)	(0.015)	(0.062)	(0.046)	(0.032)	(0.055)	(0.034)	(0.027)
ATE	0.128	0.136	0.142	0.139	0.137	0.145	0.106	0.142	0.158	0.162	0.148	0.164
	(0.030)	(0.026)	(0.022)	(0.029)	(0.022)	(0.018)	(0.067)	(0.045)	(0.034)	(0.053)	(0.031)	(0.027)
ATE (bc)	0.143	0.164	0.156	0.148	0.171	0.158	0.118	0.188	0.172	0.170	0.213	0.170
	(0.035)	(0.033)	(0.026)	(0.034)	(0.030)	(0.024)	(0.075)	(0.061)	(0.041)	(0.064)	(0.047)	(0.037)
Outcome Eqn.												
Baseline and Geographic[(i)]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Parental Education[(ii)]							✓	✓	✓	✓	✓	✓
Selection Eqn.												
Baseline and Geographic[(i)]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Parental Education[(ii)]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family Structure[(iii)]			✓			✓			✓			✓
4-yr College Proximity[(iv)]	✓	✓	✓				✓	✓	✓			
Interactions[(v)]		✓	✓					✓	✓			
Three College Proximities[(vi)]				✓	✓	✓				✓	✓	✓
Interactions[(v)]					✓	✓					✓	✓
k	25	35	37	27	57	59	25	35	37	27	57	59
k/n	0.01	0.02	0.02	0.01	0.02	0.03	0.01	0.02	0.02	0.01	0.02	0.03
$k/\sqrt{n}$	0.52	0.73	0.77	0.56	1.19	1.23	0.52	0.73	0.77	0.56	1.19	1.23

Discussion and Conclusion

1. Motivation: Elucidate the finite-sample sensitivity of distributional approximations

2. Some examples:
   - HAC Estimation: $b_n \propto n$
   - Weak IV: $\pi \propto n^{-1/2}$
   - Many IV: $\dim(\pi) \propto n$
   - FE Panels: $T \propto n$
   - NN-based ("matching"): $M_n \propto 1$
   - Kernel-based semiparametrics: $h_n \propto n^{\gamma}$
   - Series-based semiparametrics: $K_n \propto n^{\gamma}$
   - High-dimensional linear regression models: $K_n \propto n$

3. Today:
   - Two-step estimation with high-dimensional first step: $k \propto \sqrt{n}$
   - Gaussian distributional approximation not valid: first-order bias!
   - Jackknife bias and variance estimation works: fully automatic!
   - Wild bootstrap applied to jackknife-based t-statistic works very well: fully automatic!

THANK YOU !

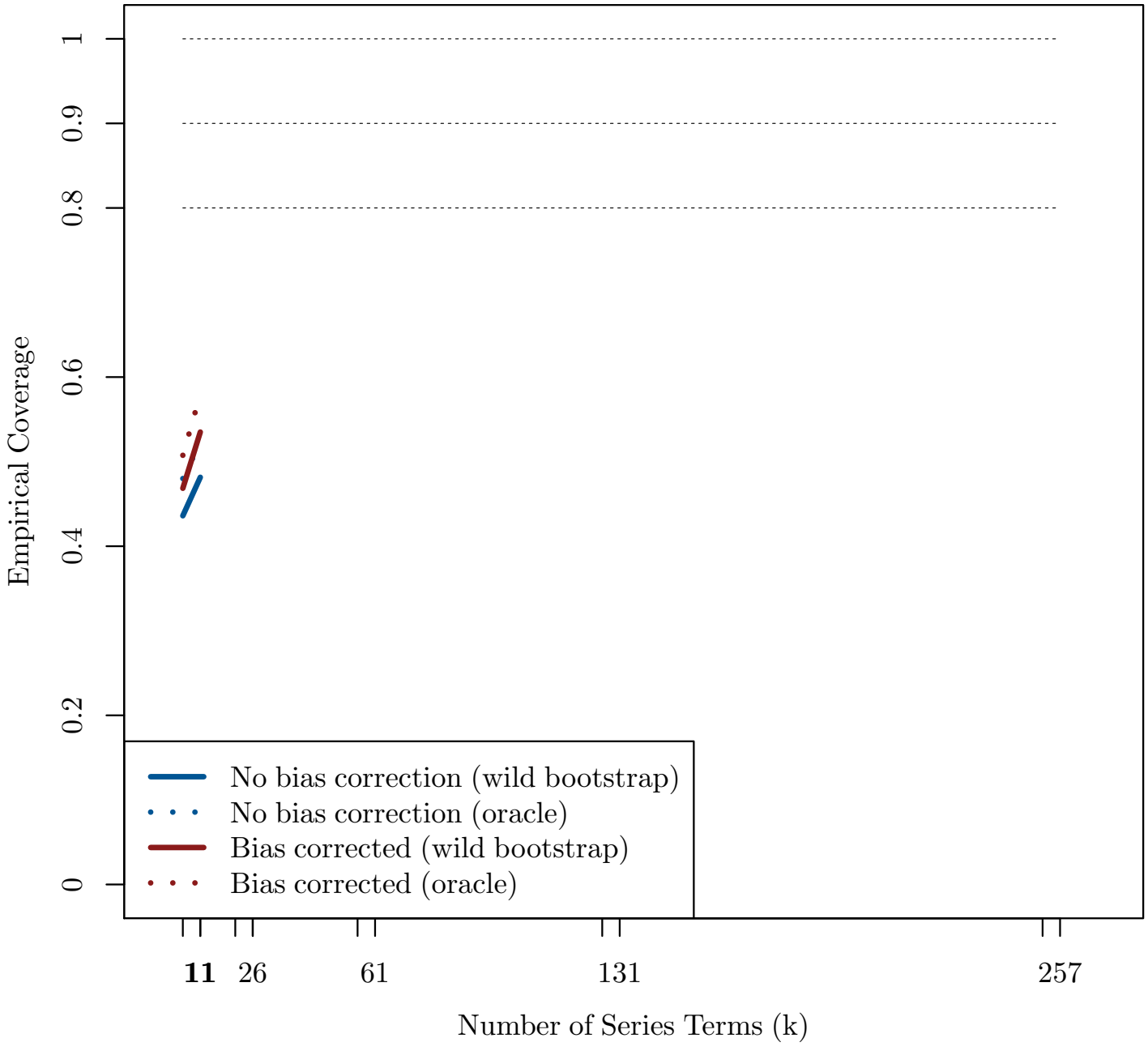
Appendix: Simulations with Misspecification Error

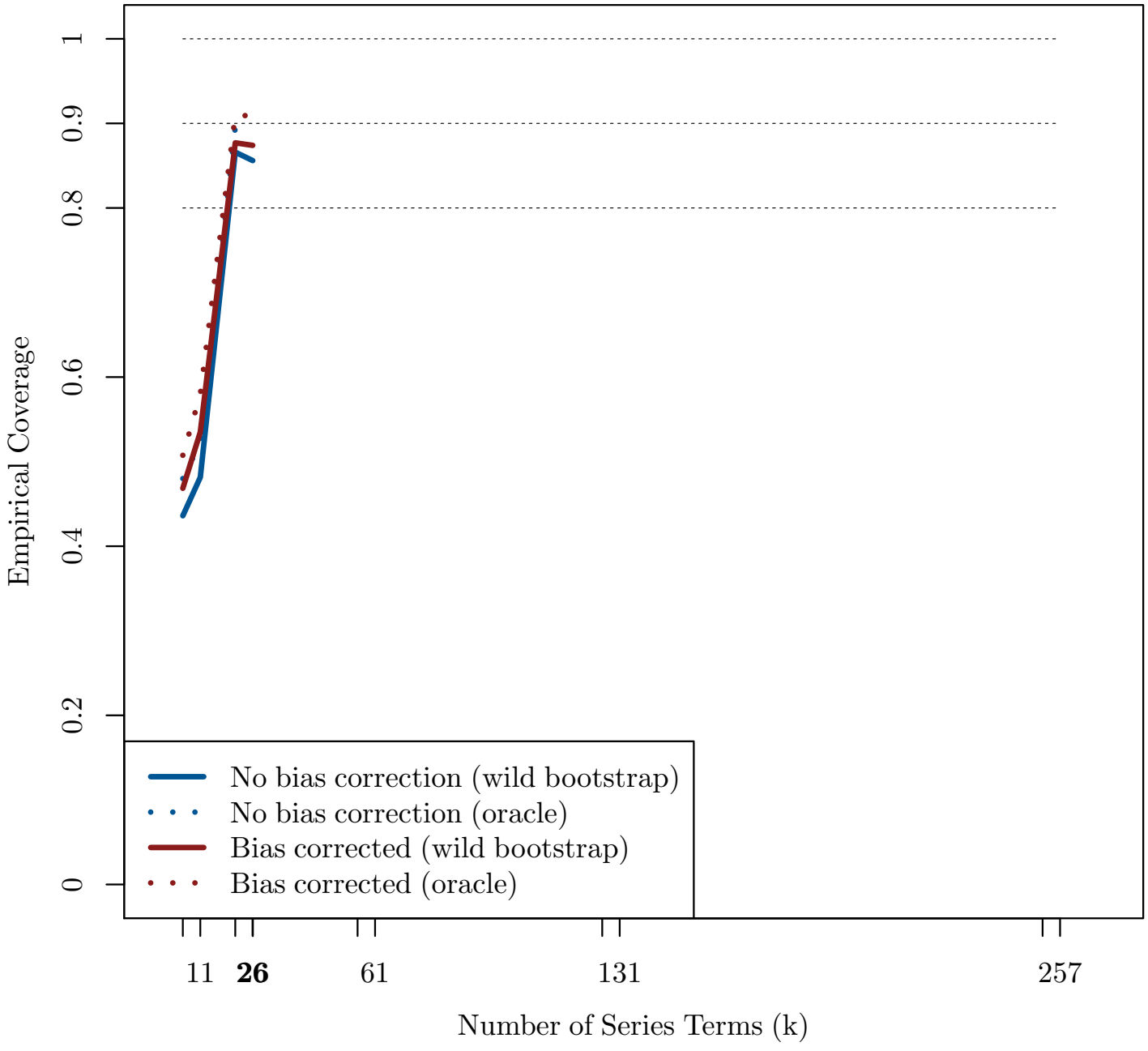
- $n = 2,000$

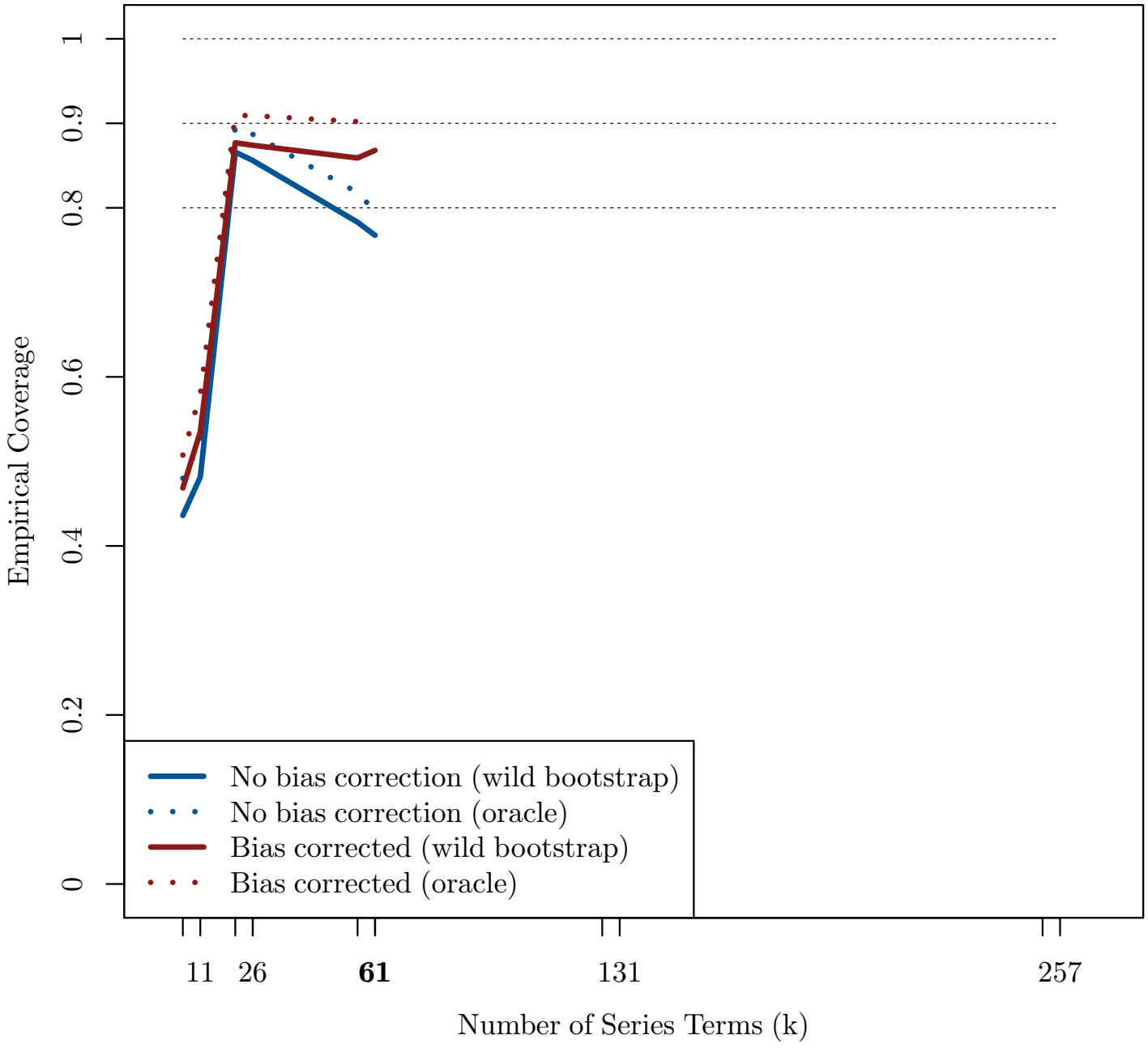
Table: Polynomial Basis Expansion.

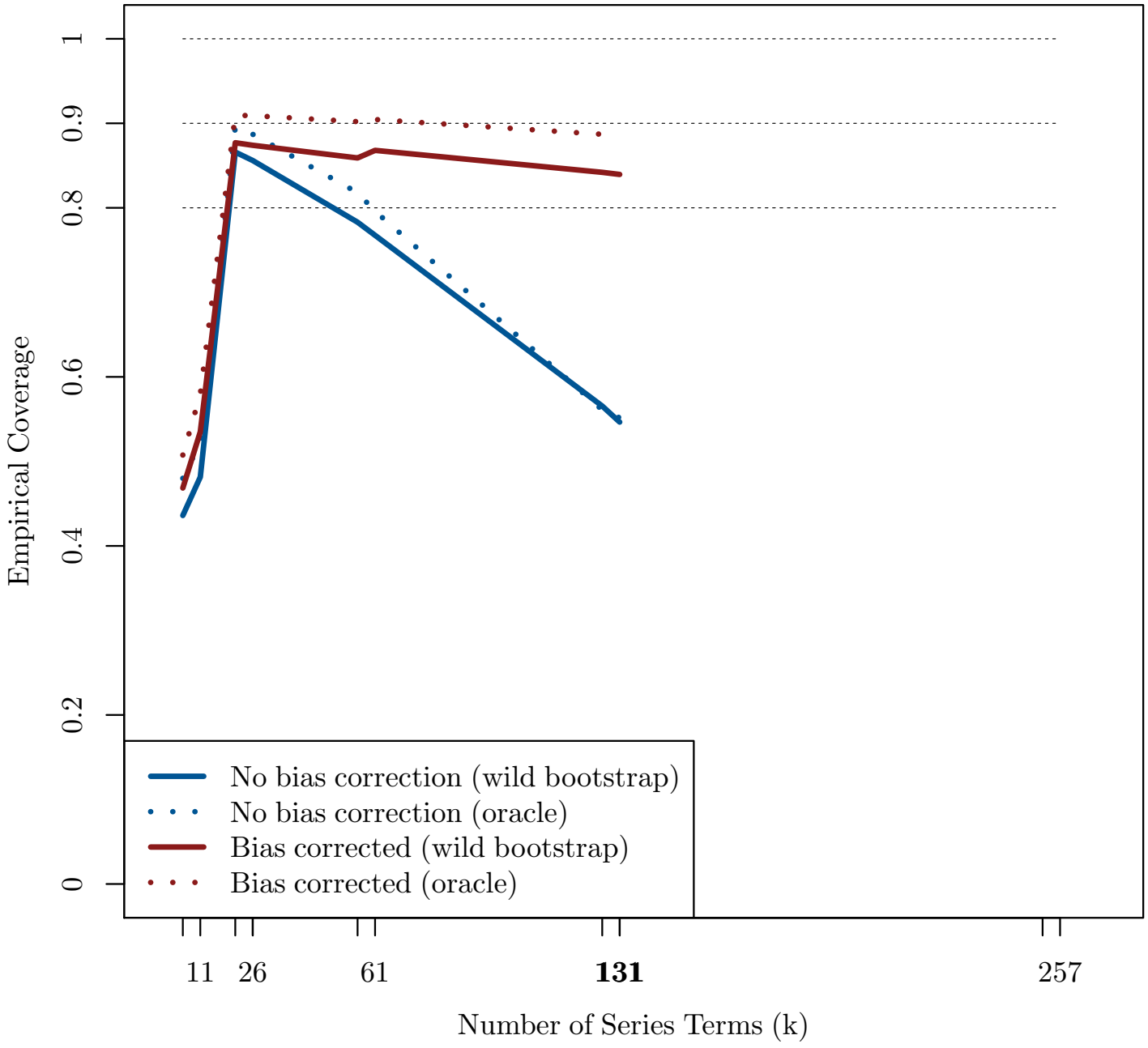
Series terms (k)	$\mathbf{s}^k(\mathbf{z}_i)$
6	1 and $\mathbf{z}_i$
11	1, $\mathbf{z}_i$ and $[z_{1i}^2, z_{2i}^2, \cdots, z_{5i}^2]$
21	$\mathbf{s}^{11}(\mathbf{z}_i)$ and 2^{nd}-order interactions
26	$\mathbf{s}^{21}(\mathbf{z}_i)$ and $[z_{1i}^3, z_{2i}^3, \cdots, z_{5i}^3]$
56	$\mathbf{s}^{26}(\mathbf{z}_i)$ and 3^{rd}-order interactions
61	$\mathbf{s}^{56}(\mathbf{z}_i)$ and $[z_{1i}^4, z_{2i}^4, \cdots, z_{5i}^4]$
126	$\mathbf{s}^{61}(\mathbf{z}_i)$ and 4^{th}-order interactions
131	$\mathbf{s}^{126}(\mathbf{z}_i)$ and $[z_{1i}^5, z_{2i}^5, \cdots, z_{5i}^5]$
252	$\mathbf{s}^{131}(\mathbf{z}_i)$ and 5^{th}-order interactions
257	$\mathbf{s}^{252}(\mathbf{z}_i)$ and $[z_{1i}^6, z_{2i}^6, \cdots, z_{5i}^6]$

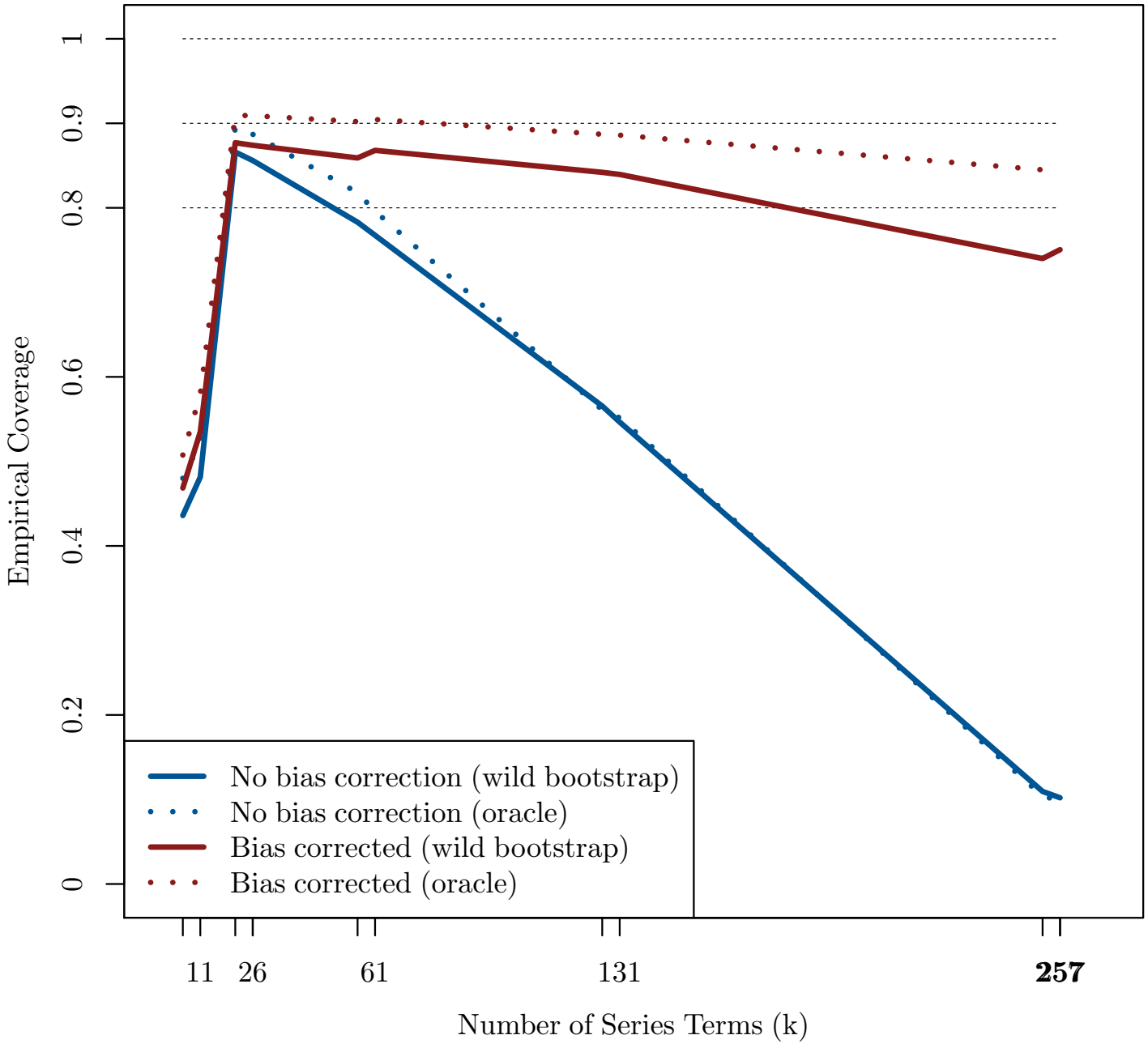
Notes. In DGP 3, the propensity score is a nonlinear function of five raw instruments, denoted by $\mathbf{z}_i \in \mathbb{R}^5$. We use the series method to estimate the propensity score, by including interactions and higher moments of the raw instruments, where the series basis is denoted by $\mathbf{s}^k(\mathbf{z}_i)$. An intercept is also included.


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
11
26
61
131
257
Number of Series Terms (k)

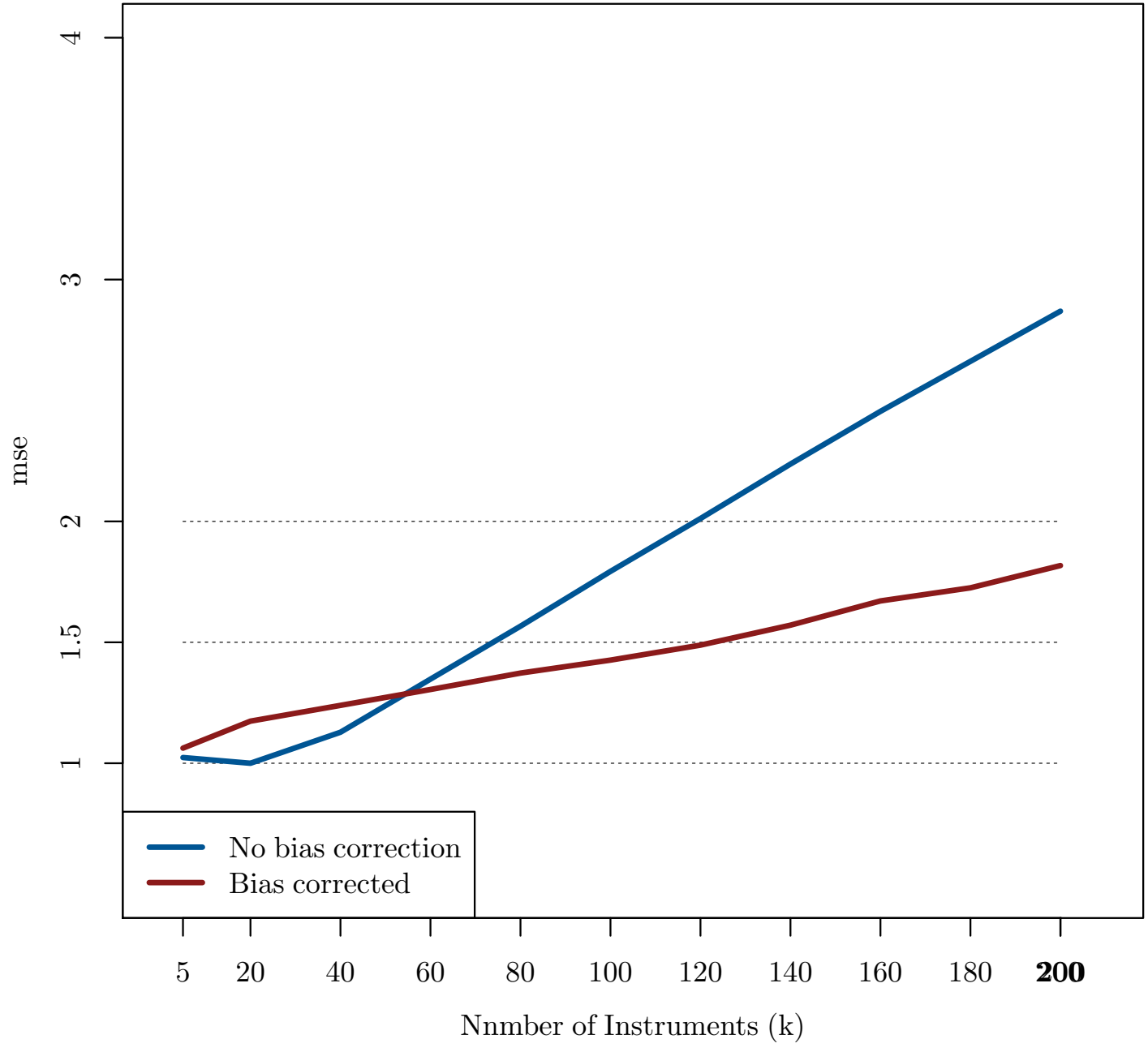

Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
11
26
61
131
257
Number of Series Terms (k)


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
11
26
61
131
257
Number of Series Terms (k)

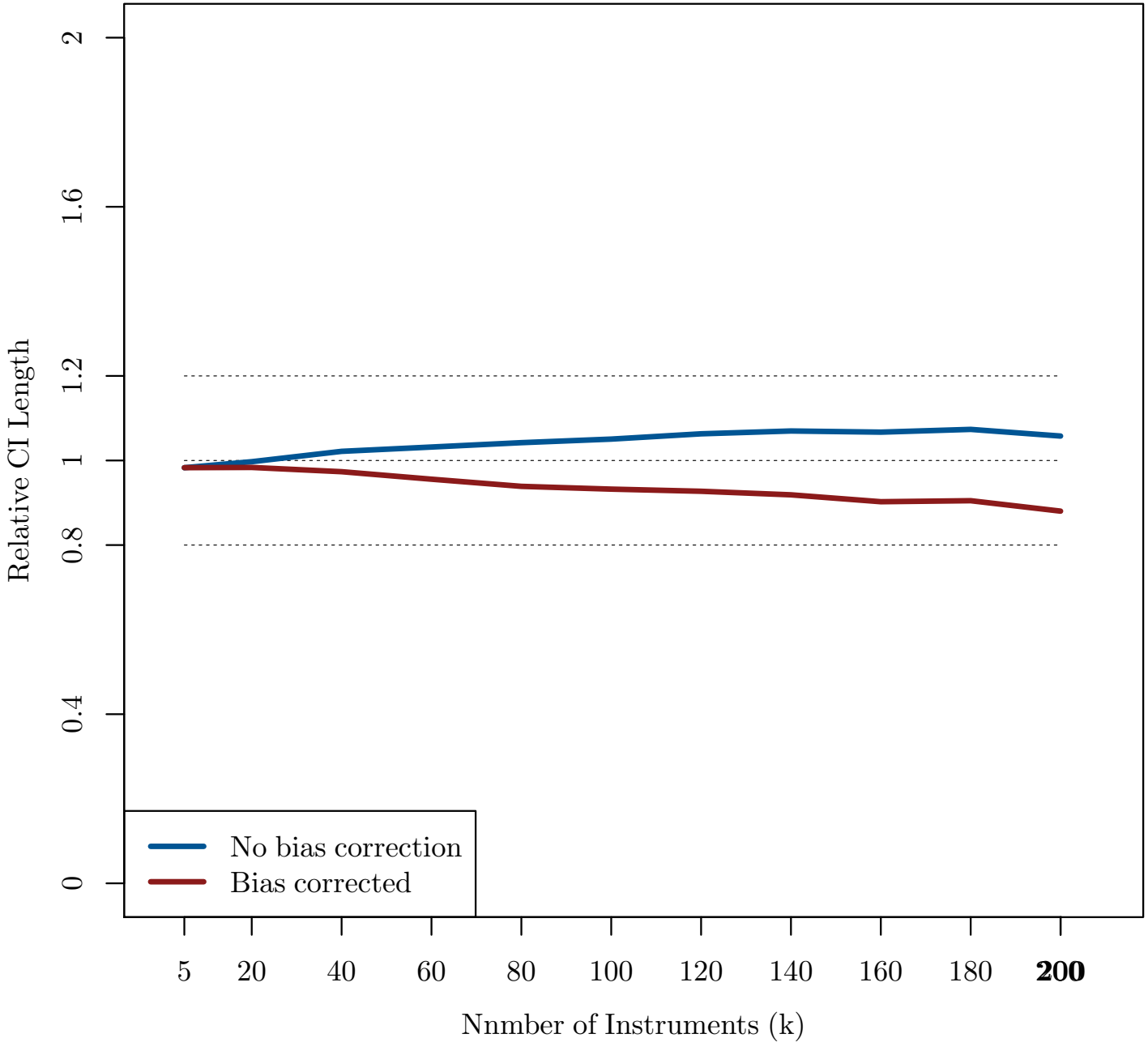

Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
11
26
61
131
257
Number of Series Terms (k)


Empirical Coverage
0
0.2
0.4
0.6
0.8
0.9
1
No bias correction (wild bootstrap)
No bias correction (oracle)
Bias corrected (wild bootstrap)
Bias corrected (oracle)
11
26
61
131
257
Number of Series Terms (k)

Appendix: Simulations, MSE



Appendix: Simulations, Interval Length



Appendix: Jackknife Estimation

- **Step 1**. For each observation $j = 1, 2, \ldots, n$ estimate the propensity score without using the j-th observation, denoted by $\hat{p}_{n,i}^{(j)}$, and compute the leave-j-out estimator by solving

$$\hat{\boldsymbol{\theta}}^{(j)} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n-1} \sum_{i=1, i\neq j}^{n} \left(y_i - m(\mathbf{x}_i, \hat{p}_{n,i}^{(j)}, \boldsymbol{\theta}) \right)^2 .$$

  Define $\hat{\boldsymbol{\theta}}^{(\cdot)} = \frac{1}{n} \sum_{j=1}^{n} \hat{\boldsymbol{\theta}}^{(j)}$.

- **Step 2**. The jackknife bias estimator is defined as

$$\hat{\boldsymbol{\mathcal{B}}} = (n-1) \cdot \sqrt{n} \left(\hat{\boldsymbol{\theta}}^{(\cdot)} - \hat{\boldsymbol{\theta}} \right) = \frac{n-1}{n} \sum_{j=1}^{n} \sqrt{n} \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}} \right),$$

  and the bias corrected estimator is $\hat{\boldsymbol{\theta}}_{\mathsf{bc}} = \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mathcal{B}}}/\sqrt{n}$.

- **Step 3**. The jackknife variance estimator is

$$\hat{\boldsymbol{\mathcal{V}}} = (n-1) \sum_{j=1}^{n} \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)} \right) \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)} \right)' .$$

RECALL. The MTE and functionals thereof are

$$\hat{\tau}_{\mathtt{mte}} = \frac{\partial}{\partial a} m(\mathbf{x}, a, \hat{\boldsymbol{\theta}}) \qquad \textit{and} \qquad \hat{\tau}_{\omega} = \int_0^1 \hat{\tau}_{\mathtt{mte}}(a|\mathbf{x}) \omega(a|\mathbf{x}) da$$