Adaptive Inference Techniques for Some Irregular Problems

Inference! Inference! Inference!

Arun Kumar Kuchibhotla

3 Feb, 2025

Carnegie Mellon University

Joint work

This is a joint work with Larry Wasserman, Sivaraman Balakrishnan, Siddhaarth Sarkar (CMU), Kenta Takatsu (CMU), and Woonyoung Chang (CMU).

Table of contents

1. Motivation and Examples
2. Inference I: COSI Framework
3. Inference II: M-estimation Problems[1]
4. Inference III: Z-estimation Problems[2]
5. Conclusions

[1] Joint work with Kenta Takatsu

[2] Joint work with Woonyoung Chang

Motivation and Examples

Inference: confidence intervals

- $\star$ Statistical inference is the cornerstone of statistics and is a necessary ingredient in any rigorous scientific study.

- $\star$ Suppose we have a (real-valued) functional $\theta(P), P \in \mathcal{P}$, e.g., the mean of P or a coefficient in a regression model.

- $\star$ Traditional inference methods such as Wald or resampling (e.g. bootstrap or subsampling) proceed as follows.

- $\star$ Assuming the existence of an estimator $\widehat{\theta}_n$ based on n observations such that

  $$r_n(\widehat{\theta}_n - \theta(P)) \xrightarrow{d} L,$$

  a confidence interval can be constructed as

  $$\widehat{\mathrm{CI}}_{n,\alpha} := \left[\widehat{\theta}_n - \frac{\widehat{q}_{1-\alpha/2}}{\widehat{r}_n}, \widehat{\theta}_n + \frac{\widehat{q}_{\alpha/2}}{\widehat{r}_n}\right],$$

  where $\widehat{q}_\gamma$ represents an estimate of the γ-th quantile of the random variable L, and $\widehat{r}_n$ is an estimate of r_n, if unknown.

Motivating Example 1: Linear Regression

- Suppose $(X_i, Y_i), 1 \le i \le n$ are IID random vectors with $X_i \in \mathbb{R}^d$. Consider

$$\theta(P) := \underset{\theta \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}[|Y - X^\top \theta|^2].$$

- The OLS estimator $\widehat{\theta}_n$ satisfies

$$\|\widehat{\theta}_n - \theta(P)\| = O_p(\sqrt{d/n}),$$

but with $\Sigma = \mathbb{E}_P[XX^\top]$ and $V = \mathbb{E}_P[XX^\top(Y - X^\top \theta(P))^2]$,

$$n^{1/2}(\widehat{\theta}_n - \theta(P)) \overset{d}{\approx} N(0, \Sigma^{-1} V \Sigma^{-1}), \quad \text{only if} \quad d = o(n^{1/2}).$$

This implies the validity of traditional Wald or bootstrap inference if $d = o(n^{1/2})$.

- Most of the results and methods fail if $d \gg n^{1/2}$.

Motivating Example 1: Linear Regression (Contd.)

- ⋆ In general, it can be proved that

$$n^{1/2}(c^\top\widehat{\theta}_n - c^\top\theta(P)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} c^\top \mathrm{IF}(X_i, Y_i) + O_p\left(\frac{d}{\sqrt{n}}\right),$$

with the first term approximately normal.

- ⋆ Hence, if $d \gg n^{1/2}$, then

$$n^{1/2}(c^\top\widehat{\theta}_n - c^\top\theta(P)) \xrightarrow{p} \infty.$$

- ⋆ We can debias the estimator that converges to normal if $d = o(n^{2/3})$. It is not yet known whether there exists a debiased estimator that is asymptotically normal for all $d = o(n)$.
- ⋆ See Chang, Kuchibhotla, and Rinaldo (2023, arXiv:2307.00795) for details.

Motivating Example 2: Quantile Regression

- ⋆ Suppose $(X_i, Y_i), 1 \le i \le n$ are IID random vectors with $X_i \in \mathbb{R}^d$. Consider

$$\theta(P) := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \mathbb{E}[|Y - X^\top \theta|].$$

- ⋆ If $Y_i = X_i^\top \theta_0 + \xi_i$ for some conditionally zero median random variables ξ_i, then $\theta(P) = \theta_0$.
- ⋆ If the conditional density of ξ_i given X_i is bounded away from zero almost surely, then

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Gamma^{-1}\Sigma\Gamma^{-1}),$$

where $\Gamma = \mathbb{E}_P[f_\xi(0|X)XX^\top]$ and $\Sigma = \mathbb{E}_P[XX^\top]$. In this case, traditional Wald and bootstrap are consistent.

- ⋆ What happens if the conditional density is zero or does not exist?

Motivating Example 2: Quantile Regression

- ⋆ If $F_X(t) = \mathbb{P}(\xi \leq t|X)$, and

$$F_X(t) - F_X(0) = A_X|t|^\gamma \mathrm{sgn}(t)(1 + o(t)) \quad \text{as} \quad t \to 0,$$

then

$$n^{1/(2\gamma)}(\widehat{\theta}_n - \theta_0) \overset{d}{\to} \operatorname*{arg\,min}_{u \in \mathbb{R}^d} u^\top W + \frac{2}{\gamma + 1}\mathbb{E}[A_X|u^\top X|^{\gamma+1}].$$

- ⋆ If $\gamma = 1$, then $A_X = f_\xi(0|X)$ and this reduces to the usual asymptotic normality result.
- ⋆ The rate of convergence depends on the (unknown) smoothness of the conditional CDF around 0.
- ⋆ Bootstrap is valid if and only if $\gamma = 1$. Wald inference cannot be applied without the knowledge of γ.

Motivating Example 3: Manski's Discrete Choice Model

- Suppose $(X_i, Y_i), 1 \le i \le n$ are IID random vectors with $X_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, from Manski's model:

$$Y_i = \mathbf{1}\{X_i^\top \theta(P) + \xi_i \ge 0\} \quad \text{with} \quad \text{Median}(\xi_i | X_i) = 0.$$

- This is a semiparametric generlization of logistic regression and is used in Econometrics for discrete choice models.

- Manski's estimator of $\theta(P)$ is

$$\widehat{\theta}_n := \operatorname*{arg\,min}_{\theta \in S^{d-1}} \sum_{i=1}^{n} (Y_i - 1/2)\mathbf{1}\{X_i^\top \theta \ge 0\}.$$

- If the conditional density of ξ given X exists and is smooth, then

$$n^{1/3}(\widehat{\theta}_n - \theta(P)) \xrightarrow{d} H \times \operatorname*{arg\,min}_{s \in \mathbb{R}^{d-1}} \mathcal{G}(s) + \frac{s^\top Vs}{2},$$

  for some mean zero Gaussian process $\mathcal{G}(\cdot)$ and some matrix v.

- Wald does not apply, bootstrap is inconsistent, and subsampling is unreliable.

Motivating Example 4: Monotone Regression

- ⋆ Consider $(X_i, Y_i), 1 \le i \le n$ from the model $Y_i = f_0(X_i) + \xi_i$ where $f_0(\cdot)$ is non-decreasing.
- ⋆ The LSE is given by

$$\widehat{f}_n = \mathop{\arg\min}_{f:\, non-decreasing} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

- ⋆ If $f_0(x_0 + t) - f_0(x_0) = A|t|^\gamma \mathrm{sgn}(t)(1 + o(1))$ as $t \to 0$, then

$$n^{\gamma/(2\gamma+1)}(\widehat{f}_n(x_0) - f_0(x_0)) \xrightarrow{d} \left(\frac{\sigma^2(x_0)A^{1/\gamma}}{h(x_0)(\gamma+1)^{1/\gamma}}\right)^{\gamma/(2\gamma+1)} \mathbb{C}_\gamma,$$

  where $\mathbb{C}_\gamma$ is related to a drifted two-sided Brownian motion.

- ⋆ Wald is not applicable, bootstrap is inconsistent, and subsampling is unreliable.
- ⋆ Any L-Lipschitz g can be written as $g(x) = f_0(x) - Lx$ for some non-decreasing f_0. Hence, inference for f_0 implies inference for Lipschitz functions.

Some Observations

- $\star$ Asymptotic normality is only one of the many possibilities for limiting distributions.

- $\star$ Often with non-normal limiting distributions, the rate of convergence of the estimator is not $n^{1/2}$. More importantly, the rate can depend on the underlying properties of the data-generating process.

- $\star$ The limiting distribution in many cases can be written as the minimizer of some stochastic process. Pflug (1995, Math. of OR) identified three classes of such stochastic processes that appear in limiting distributions. See Bhowmick and Kuchibhotla (2024, arXiv:2411.17087) for details.

- $\star$ Bootstrap and other resampling methods tend to fail when the limiting distribution is non-normal.

- $\star$ Wald interval also becomes difficult to implement.

Inference I: COSI Framework

Inference I: COSI (COnfidence sets using Scale Invariance)

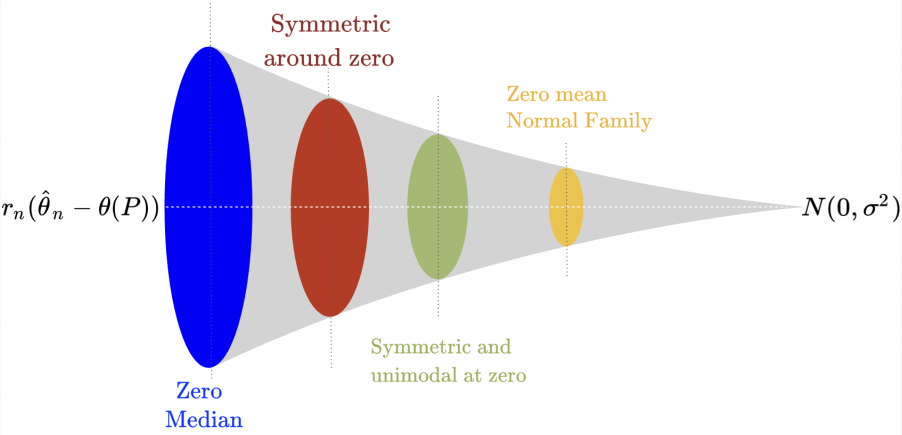

Figure 1: Illustration of Nested Structure of Limiting Distributions

Inference I: COSI Framework

- ⋆ Many estimators have a limiting distribution that meets a scale-invariant property.

- ⋆ A property $\mathfrak{P}$ is called *scale invariant*, if for every random variable W satisfying $\mathfrak{P}$, cW also satisfies $\mathfrak{P}$ for any $c \geq 0$.

- ⋆ Here are a few examples:

Name	Definition
Central Symmetry	$W \overset{d}{=} -W$
Angular Symmetry	$W/\|W\| \overset{d}{=} -W/\|W\|$
Unimodality at 0	Density maximized uniquely at 0
Normal with mean zero	$W \sim N(0, \Sigma)$

- ⋆ Zero is the "center" for any distribution satisfying a scale invariant property.

- ⋆ If $r_n(\widehat{\theta}_n - \theta(P)) \xrightarrow{d} L$ and L satisfies some scale-invariant property, then $\widehat{\theta}_n - \theta(P)$ also *approximately* satisfies the scale-invariant property.

The COSI Algorithm

- ⋆ Suppose we have n IID observations $Z_1, \dots, Z_n$.
- ⋆ Randomly split into B batches of approximately equal size and compute the estimator on each batch. We get

$$\begin{pmatrix} r_{n/B}(\widehat{\theta}^{(1)} - \theta(P)) \\ \vdots \\ r_{n/B}(\widehat{\theta}^{(B)} - \theta(P)) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} L^{(1)} \\ \vdots \\ L^{(B)} \end{pmatrix}.$$

- ⋆ Note that $L^{(1)}, \dots, L^{(B)}$ are IID and if the limiting distribution satisfies a scale invariant property $\mathfrak{P}$, then we can think of $\widehat{\theta}^{(j)} - \theta(P), 1 \le j \le B$ as IID observations from a distribution that satisfies $\mathfrak{P}$ approximately.
- ⋆ Return the confidence set

$$\widehat{\mathrm{CI}}_{n,\alpha} := \left\{\theta : \text{test for } \mathfrak{P} \text{ using } \{\widehat{\theta}^{(j)} - \theta\}_{j=1}^{B} \text{ is not rejected}\right\},$$

- ⋆ Specific scenarios for univariate functionals to follow.

Scenario I: Zero mean Normality

- For $\theta(P) \in \mathbb{R}$, consider the scale invariant property of zero mean normality. A random variable $W \in \mathbb{R}$ is zero mean normal if

$$W \sim N(0, \sigma^2) \quad \text{for some } \sigma \geq 0.$$

- This is commonly occurring case for "regular" parametric and semiparametric models.

- A classical test for testing zero mean of a normal is the t-test yielding the COSI confidence interval

$$\widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{t-test}} := \left[\frac{1}{B}\sum_{j=1}^{B}\widehat{\theta}_j \ \pm \ \frac{s_B}{\sqrt{B}}t_{B-1,\alpha}\right],$$

where $s_B^2 = \sum_{j=1}^{B}(\widehat{\theta}_j - \bar{\theta}_B)^2/(B-1)$.

- This is precisely the t-test confidence interval of Ibragimov and Muller (2010, Journal of Business and Economic Statistics).

Scenario II: Zero median

- For $\theta(P) \in \mathbb{R}$, consider the scale invariant property of zero median. A random variable $W \in \mathbb{R}$ has zero median if

$$\min\{\mathbb{P}(W \geq 0), \mathbb{P}(W \leq 0)\} \geq 1/2.$$

- Asymptotic zero median is same as "estimator is equally likely to over-estimate and under-estimate $\theta(P)$."

- A classical test for zero median is the sign test yielding the COSI confidence interval

$$\widehat{\mathrm{CI}}_{n,\alpha}^{\mathtt{HulC}} := \left[\widehat{\theta}^{(\lfloor B/2\rfloor - c_{B,\alpha})}, \widehat{\theta}^{(\lceil B/2\rceil + c_{B,\alpha}+1)}\right], \quad \text{if } B \geq \log_2(2/\alpha).$$

Here $c_{B,\alpha}$ is the $(1-\alpha/2)$-th quantile of $\mathrm{Bin}(B, 1/2) - \lfloor B/2\rfloor$.

- This is a generalization of the HulC confidence intervals proposed in Kuchibhotla et al. (2024, JRSS-B).

Scenario III: Symmetry around zero

- ⋆ For $\theta(P) \in \mathbb{R}$, consider the scale invariant property of symmetry around zero. A random variable $W \in \mathbb{R}$ is symmetric around zero if

$$W \overset{d}{=} -W \quad \text{or equivalently} \quad |W| \perp \text{sign}(W).$$

- ⋆ Next to normality, this is the most common case. Quantile regression, Monotone regression, Grenander estimator, and Manski's estimator all satisfy this invariance property.

- ⋆ A classical test for symmetry around zero is the sign-rank test yielding the COSI confidence interval

$$\widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{Sym}} := \left[A_{\lfloor 2^{B-1}\alpha\rfloor}, A_{2^B - \lfloor 2^{B-1}\alpha\rfloor}\right],$$

where $A_1 \le A_2 \le \cdots \le A_{2^B-1}$ is the ordered sequence of all subset averages $\{|S|^{-1}\sum_{j\in S}\widehat{\theta}_j : S \subseteq \{1, \ldots, B\}\}$. See Hartigan (1969, JASA) and Maritz (1979, Biometrika).

- ⋆ This yields a generalization of randomization based tests under approximate symmetry of Canay et al. (2017, Econometrica).

Finite-sample Micoverage Bounds

⋆ For any scale-invariance property $\mathfrak{P}$, we have

$$\mathbb{P}\left(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{COSI}}\right) \le \alpha + B \times \max_{1\le j\le B} \mathrm{dist}(\widehat{\theta}^{(j)} - \theta(P), \mathfrak{P}).$$

Finite-sample Micoverage Bounds

- ⋆ For any scale-invariance property $\mathfrak{P}$, we have

$$\mathbb{P}\left(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{COSI}}\right) \le \alpha + B \times \max_{1\le j\le B} \mathrm{dist}(\widehat{\theta}^{(j)} - \theta(P), \mathfrak{P}).$$

- ⋆ For zero median property, we have

$$\mathbb{P}(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{HulC}}) \le \alpha \left(1 + 2B\Delta e^{2B\Delta}\right),$$

where

$$\Delta := \max_{1\le j\le B} \left(\frac{1}{2} - \min_{s\in\{\pm 1\}} \mathbb{P}(s(\widehat{\theta}^{(j)} - \theta(P)) \ge 0)\right)_+ .$$

Finite-sample Micoverage Bounds

- ⋆ For any scale-invariance property $\mathfrak{P}$, we have

$$\mathbb{P}\left(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{COSI}}\right) \le \alpha + B \times \max_{1\le j\le B} \mathrm{dist}(\widehat{\theta}^{(j)} - \theta(P), \mathfrak{P}).$$

- ⋆ For zero median property, we have

$$\mathbb{P}(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{HulC}}) \le \alpha\left(1 + 2B\Delta e^{2B\Delta}\right),$$

where

$$\Delta := \max_{1\le j\le B}\left(\frac{1}{2} - \min_{s\in\{\pm1\}} \mathbb{P}(s(\widehat{\theta}^{(j)} - \theta(P)) \ge 0)\right)_+ .$$

- ⋆ For symmetry around zero, we have

$$\mathbb{P}(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}^{\texttt{Sym}}) \le \alpha\left(1 + 2\Delta\right)^B,$$

where

$$\Delta := \max_{1\le j\le B} \mathbb{E}\left[\left(\frac{1}{2} - \min_{s\in\{\pm1\}} \mathbb{P}(s(\widehat{\theta}^{j} - \theta(P)) \ge 0 \big| |\widehat{\theta}^{(j)} - \theta(P)|)\right)_+\right].$$

Illustration: Normal means

Suppose $\widehat{\theta}_j - 0 \sim N(0, 0.5^2), 1 \leq j \leq 10$. Then

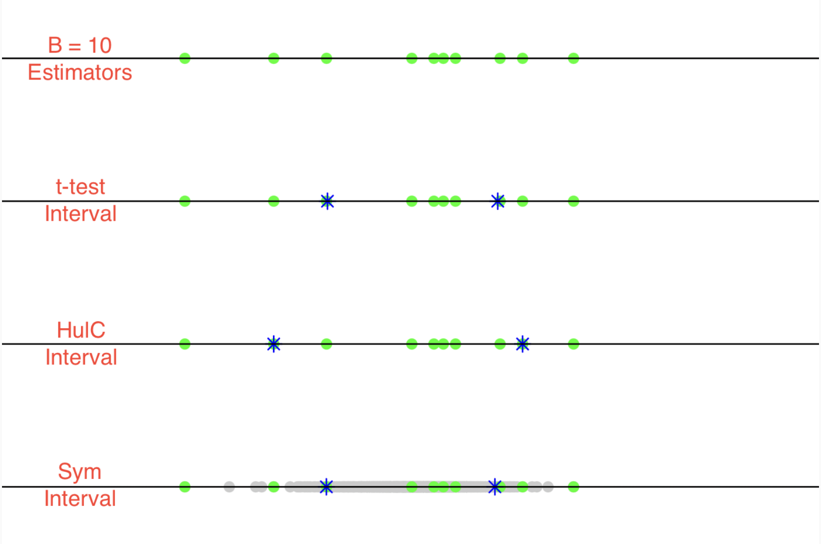

Illustration: Quantile Regression

$$Y_i = X_i^\top \beta_0 + \xi_i, X_i \sim N(0, 0.8I_3 + 0.2\mathbf{1}\mathbf{1}^\top),$$
$$F_X(t) = 0.5 + 0.5\text{sgn}(t)|t|^\gamma, t \in [-1, 1].$$

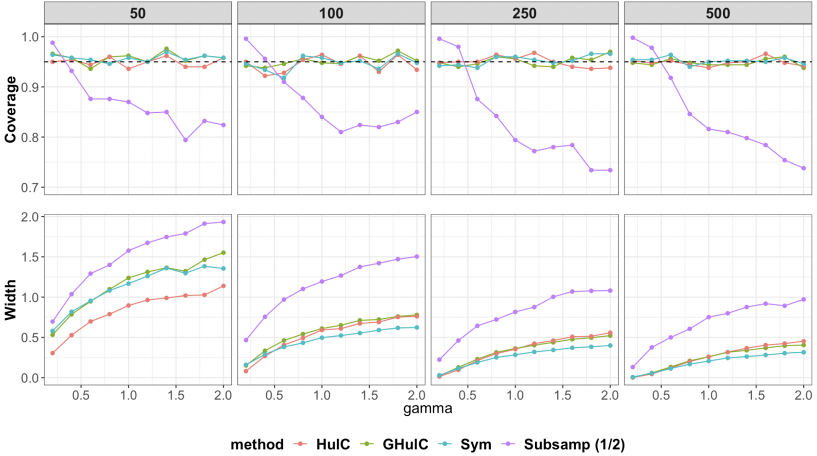

Figure 2: Illustration of Coverage and Width in Quantile Regression.

Pros and Cons

- ⋆ The procedure needs neither the rate of convergence nor the form of the limiting distribution.
- ⋆ For many scale-invariant properties, finite-sample (or distribution-free) tests can be constructed. This includes central symmetry, angular symmetry, unimodality at zero, and normality with zero mean.
- ⋆ Based on the test used for the scale-invariant property, the resulting confidence sets can have second-order accuracy.
- ⋆ The disadvantage is that one needs to understand the limiting distribution of the estimator to conclude the existence of a scale-invariant property.
- ⋆ This can be difficult, especially in non-parametric or high-dimensional problems (e.g., Lasso or non-parametric regression). Even if one knows the exact limiting distribution, it may not have any scale-invariant property.

Inference II: M-estimation Problems[a]

[a] Joint work with Kenta Takatsu

M-estimation Inference

- ⋆ Most functionals encountered in practice can be written as

$$\theta(P) := \arg\min_{\theta\in\Theta} \mathbb{E}_P[m(\theta, Z)],$$

for some loss function $m(\theta, Z)$. OLS, Quantile regression, Manski's model, MLE are some examples.

- ⋆ Setting $\mathbb{M}(\theta) = \mathbb{E}_P[m(\theta, Z)]$, we know that

$$\theta(P) \subseteq \left\{\theta \in \Theta : \mathbb{M}(\theta) \le \mathbb{M}(\widehat{\theta})\right\},$$

for any estimator $\widehat{\theta} \in \Theta$.

- ⋆ Of course, the right hand set is not computable based on the data. But we can construct two sets based on this intuition and prove their validity.

M-estimation Inference

- Consider

$$\begin{aligned}\widehat{\mathrm{CI}}_n^{\dagger} \;&:=\; \left\{\theta \in \Theta:\; \widehat{\mathbb{M}}_n(\theta) - \widehat{\mathbb{M}}_n(\widehat{\theta}_1) \;\le\; 0\right\},\\ \widehat{\mathrm{CI}}_{n,\alpha} \;&:=\; \left\{\theta \in \Theta:\; \widehat{\mathbb{M}}_n(\theta) - \widehat{\mathbb{M}}_n(\widehat{\theta}_1) \;\le\; \frac{z_{\alpha/2}\widehat{\sigma}(\theta, \widehat{\theta}_1)}{n^{1/2}}\right\},\end{aligned} \tag{1}$$

  where $\widehat{\mathbb{M}}_n(\theta) = n^{-1}\sum_{i=1}^n m(\theta, Z_i)$ and $\widehat{\theta}_1$ is obtained from an independent sample, and $\widehat{\sigma}(\theta, \widehat{\theta}_1)$ is the sample standard deviation of $m(\theta, Z_i) - m(\widehat{\theta}_1, Z_i), 1 \le i \le n$.

- Clearly,

$$\widehat{\mathrm{CI}}_n^{\dagger} \subseteq \widehat{\mathrm{CI}}_{n,\alpha} \quad \text{for any} \quad \alpha \in (0,1), n \ge 1.$$

- Note that the definition of the confidence sets have no restrictions on Θ or $\widehat{\theta}_1$ except for $\widehat{\theta}_1 \in \Theta$.

- This idea exists in the operations research literature (Vogel (2008, J. of Opt.)) where $\widehat{\theta}_1$ and $\widehat{\mathbb{M}}_n(\cdot)$ are computed on the same data.

Validity

- ⋆ For any $\widehat{\theta}_1$, we have

$$\mathbb{P}(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}) \le \mathbb{P}(\theta(P) \notin \widehat{\mathrm{CI}}_n^{\dagger}) \le \mathbb{E}\left[\frac{\sigma_P^2(\theta(P), \widehat{\theta}_1)}{\sigma_P^2(\theta(P), \widehat{\theta}_1) + n\mathbb{C}_P^2(\widehat{\theta}_1)}\right],$$

where

$$\sigma_P^2(\theta, \theta') := \mathrm{Var}(m(\theta, Z) - m(\theta', Z)),$$
$$\mathbb{C}_P(\theta') := \mathbb{E}[m(\theta, Z)] - \min_{\theta \in \Theta} \mathbb{E}[m(\theta, Z)].$$

- ⋆ If $\widehat{\theta}_1$ is consistent for $\theta(P)$, then

$$\mathbb{P}(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}) \ge 1 - \alpha - o(1) \quad \text{as} \quad n \to \infty.$$

- ⋆ Neither guarantee depends on Θ or the dimension/definition of $\widehat{\theta}_1$.
- ⋆ With a slight modification, we can obtain finite sample validity for these confidence intervals if the loss is bounded (with a known bound).
- ⋆ Interestingly, we can show that the confidence region $\widehat{\mathrm{CI}}_{n,\alpha}$ shrinks to a singleton at the optimal rate. It adapts!!

Simple, non-trivial example

- ⋆ Consider

$$\theta(P) := \arg\min \mathbb{E}[|Y - X^\top\theta|^2] + h(\theta),$$

where $h(\cdot)$ is some non-stochastic penalty, such as

$$h(\theta) = \quad \lambda\|\theta\|_\rho^\rho, \quad \rho \geq 0 \quad \text{or} \quad \begin{cases} 0, & \text{if } A\theta \leq b, \\ +\infty, & \text{if } A\theta \not\leq b \end{cases}$$

- ⋆ The OLS would be a penalized/constrained least squares estimator and can be efficiently computed.

- ⋆ However, the limiting distribution of the OLS is incomprehensible because it depends on the derivative of penalty at $\theta(P)$ and/or inequalities that are active at $\theta(P)$, i.e., the coordinates j such that $a_j^\top\theta(P) = b_j$.

- ⋆ To my knowledge, no uniformly valid inference procedure exists except $\widehat{\mathrm{CI}}_{n,\alpha}$. Also, note that our procedure does not require a well-specified linear model.

Inference III: Z-estimation Problems[a]

[a] Joint work with Woonyoung Chang

Z-estimation Problems

- $\star$ Z-estimation problems refer to functionals defined as solutions to equations:

$$\mathbb{E}_P[\Psi(\theta(P), Z)] = 0,$$

for some estimating equation $\Psi : \Theta \otimes \mathcal{Z} \to \mathbb{R}^d$ (assuming $\Theta \subseteq \mathbb{R}^d$).

- $\star$ In general, we can consider $\theta(P)$ defined by a set of moment equalities and inequalities. Such weakly/partially identified parameters are common in econometrics.
- $\star$ For any set $\mathcal{A} \subseteq S^{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$, consider the set

$$\widehat{\mathrm{CI}}_{n,\alpha} = \left\{\theta \in \Theta : \sup_{a \in \mathcal{A}} \frac{|\sum_{i=1}^n a^\top \Psi(\theta, Z_i)|}{\sqrt{\sum_{i=1}^n (a^\top \Psi(\theta, Z_i))^2}} \le \kappa_\alpha \right\},$$

where $\kappa_\alpha = \kappa_\alpha(\mathcal{A})$ is the quantile of the maximum of a sequence of Gaussian random variables.

Z-estimation Problems

- ⋆ Validity follows from an application of high-dimensional or infinite-dimensional CLT, and hence, the validity guarantee is tied to the "complexity" of $\mathcal{A}$.
- ⋆ Note that unlike the procedure for M-estimation problem, no pilot estimator is needed for the construction of the confidence set. In this respect, this confidence set may be more powerful as it uses all available data for inference and none for estimation.
- ⋆ Choosing $\mathcal{A}$ to be a singleton has some interesting implications for a one-dimensional functional of $\theta(P)$; e.g., $a = \widehat{\Sigma}^{-1}e_j$ for $e_j^\top\theta(P)$.
- ⋆ In the context of linear regression, with $\mathcal{A} = \{e_j, 1 \le j \le d\}$, this confidence set becomes

$$\widehat{\mathrm{CI}}_{n,\alpha} = \left\{\theta \in \mathbb{R}^d : \max_{a\in\mathcal{A}} \frac{|\sum_{i=1}^n (a^\top X_i)(Y_i - X_i^\top\theta)|}{\sqrt{\sum_{i=1}^n (a^\top X_i)^2 (Y_i - X_i^\top\theta)^2}} \le \kappa_\alpha\right\}.$$

This is valid as long as $d = o(n)$ and has a diameter shrinking as $\sqrt{d/n}$.

Conclusions

Conclusions

- ⋆ Estimation has received a lot of focus in both regular and irregular settings.
- ⋆ Traditionally, the construction of tests or confidence sets is mostly based on some estimation procedure and its limiting distribution.
- ⋆ We have discussed three new inference procedures, two of which completely avoid the study of intricate limiting behavior of the pilot estimator.
- ⋆ The validity of all three methods is relatively easy, especially compared to that of resampling methods.
- ⋆ Although the methods are not developed with optimality as a goal, all of them yield optimal adaptive confidence sets.

Conclusions

- ⋆ Estimation has received a lot of focus in both regular and irregular settings.
- ⋆ Traditionally, the construction of tests or confidence sets is mostly based on some estimation procedure and its limiting distribution.
- ⋆ We have discussed three new inference procedures, two of which completely avoid the study of intricate limiting behavior of the pilot estimator.
- ⋆ The validity of all three methods is relatively easy, especially compared to that of resampling methods.
- ⋆ Although the methods are not developed with optimality as a goal, all of them yield optimal adaptive confidence sets.

Thank You!