Robustly Valid Inference for M- and Z-estimation Problems

Inference! Inference! Inference!

Arun Kumar Kuchibhotla

22 May, 2025

Carnegie Mellon University

Collaborators

Manit Paul (UPenn)

Kenta Takatsu (CMU)

Woonyoung Chang (CMU)

Selina Carter (CMU)

Table of contents

[1] Joint work with Larry Wasserman and Siva Balakrishnan
[2] Joint work with Kenta Takatsu (arXiv:2501.07772)
[3] Joint work with Woonyoung Chang (arXiv:2407.12278)

Motivation and Examples

Inference: confidence intervals

- ⋆ Statistical inference is the cornerstone of statistics and is a necessary ingredient in any rigorous scientific study.
- ⋆ 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.
- ⋆ Traditional inference methods such as Wald or resampling (e.g. bootstrap or subsampling) proceed as follows.
- ⋆ 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.

Limitations of Traditional Inference

⋆ Estimation of quantiles or even the asymptotic variance can be difficult.

⋆ The rate of convergence of the estimator can depend on the underlying data generating process.

⋆ The limiting distribution may be intractable.

⋆ Finally, traditional methods can be computationally expensive.

The robust statistics literature includes several examples that fit one or more of the above limitations. For example, sample median or quantile regression is robust but

– its asymptotic variance involves (conditional) densities hard to estimate;
– rate of convergence can be unknown if density becomes zero; and
– limiting distribution can be intractable.

Finally, except for simpler models, robust estimators are derived from non-convex optimization and can be computationally intensive for resampling purposes.

Inference I: HulC[a]

[a] Joint work with Larry Wasserman and Siva Balakrishnan

Inference I: HulC (Hull-based Confidence Regions)

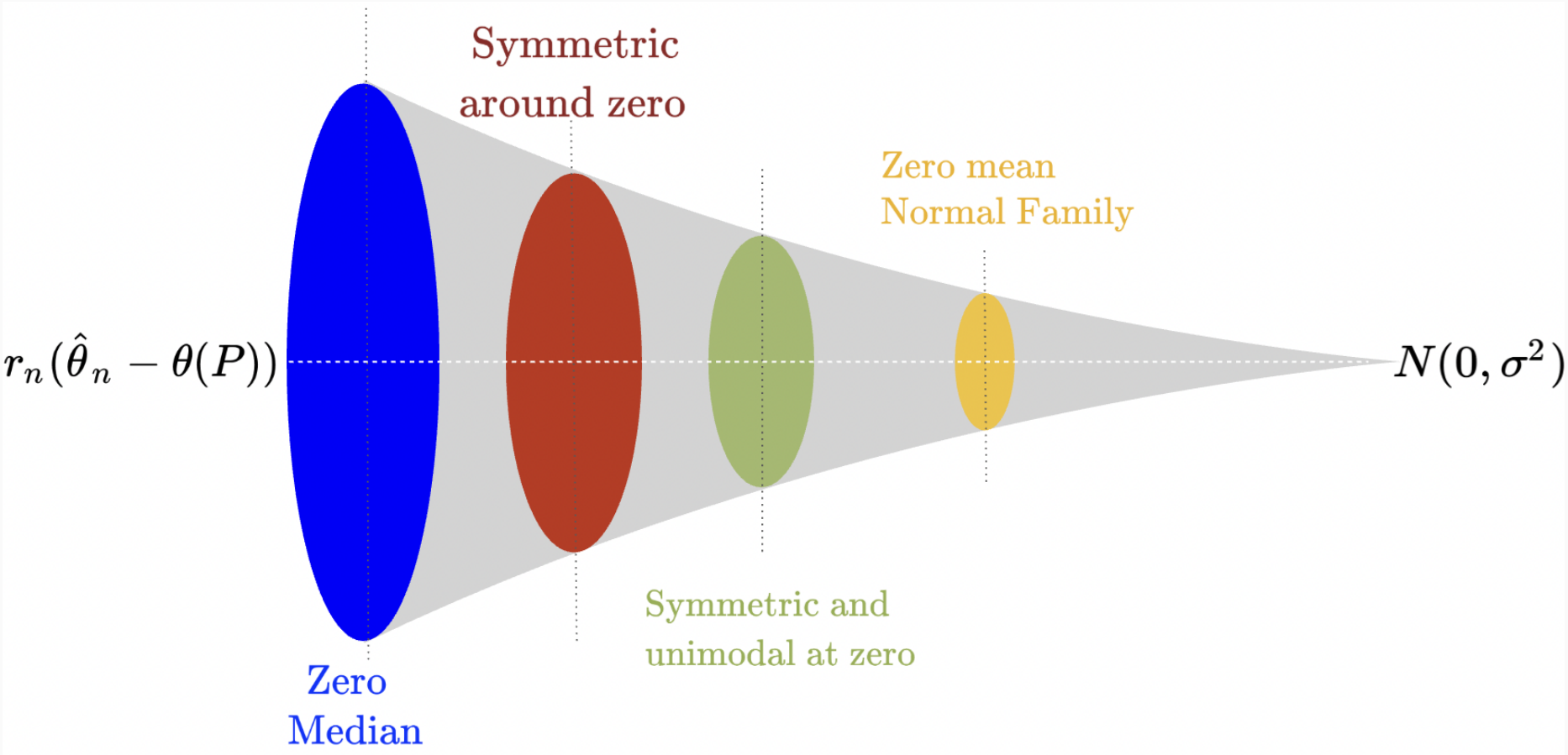


Figure 1: Illustration of Nested Structure of Limiting Distributions

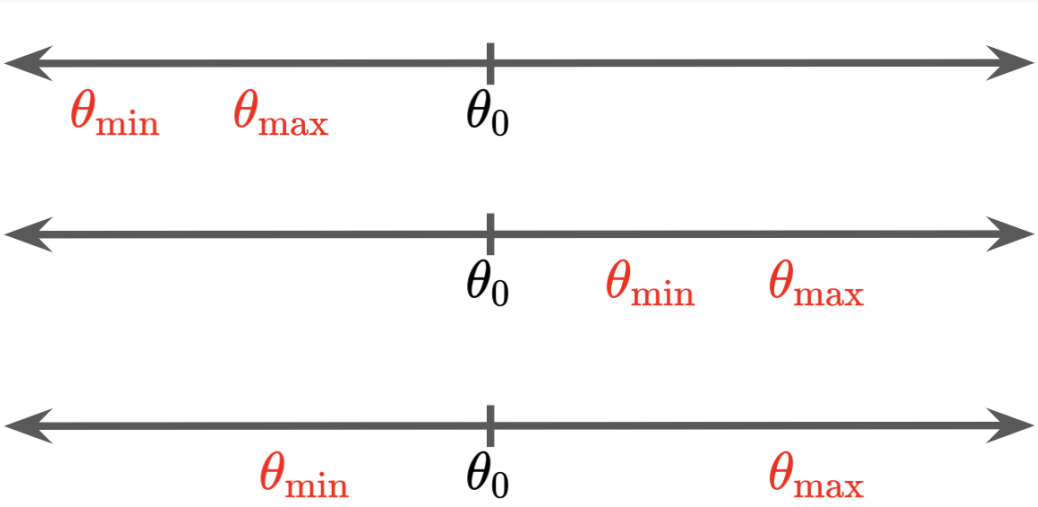
Inference I: HulC

⋆ Suppose that we have two *independent* estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for the parameter/functional $\theta_0 = \theta(P)$ satisfying

$$\mathbb{P}(\widehat{\theta}_j \geq \theta(P)) = \mathbb{P}(\widehat{\theta}_j \leq \theta(P)) = 1/2 \quad \text{for} \quad j = 1, 2.$$

⋆ Define

$$\theta_{\mathsf{min}} = \min\{\widehat{\theta}_1, \widehat{\theta}_2\} \quad \text{and} \quad \theta_{\mathsf{max}} = \max\{\widehat{\theta}_1, \widehat{\theta}_2\}.$$



$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

Hence, $[\theta_{\mathsf{min}}, \theta_{\mathsf{max}}]$ is a valid 50% confidence interval.

The HulC Algorithm

- ⋆ Suppose we have n IID observations $Z_1, \ldots, Z_n$.
- ⋆ Randomly split into $B = \log_2(2/\alpha)$ batches of approximately equal size and compute the estimator on each batch: $\widehat{\theta}^{(j)}, 1 \le j \le B$.
- ⋆ Return the confidence set

$$\widehat{\mathrm{CI}}_{n,\alpha} := \left[\min_{1 \le j \le B} \widehat{\theta}^{(j)}, \max_{1 \le j \le B} \widehat{\theta}^{(j)} \right].$$

– If the median bias

$$\Delta = \max_{1 \le j \le B} \left(\frac{1}{2} - \min \left\{ \mathbb{P}(\widehat{\theta}^{(j)} \ge \theta_0), \mathbb{P}(\widehat{\theta}^{(j)} \le \theta_0) \right\} \right)_+ \to 0,$$

then

$$\limsup_{n \to \infty} \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{n,\alpha}) \le \alpha.$$

– Note that the computational cost is always smaller than the cost of computing the estimator on n observations. ($B \approx 6$ for $\alpha = 0.05$.)

– No need for knowledge of the rate of convergence and the limiting distribution.

Applications

- ⋆ Every asymptotically normal estimator satisfies $\Delta \to 0$.
- ⋆ Many non-normal estimators also satisfy $\Delta \to 0$:
  - Quantile regression
  - Least median of squares
  - Shorth estimator
  - Estimators of mode
  - Robust non-parametric monotone regression
  - MM-estimators.
- ⋆ The great advantage of this method is that one does not need to estimate the variance/rate to construct the confidence interval.
- ⋆ Some limitations (not just of HuIC) include the fact that convergence in distribution happens at a slow rate for reliable performance in practice.
- ⋆ The next two methods provide inference without requiring convergence in distribution of the estimator.

Inference II: M-estimation problems[a]

[a] Joint work with Kenta Takatsu (arXiv:2501.07772)

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

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

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

$$\widehat{\mathrm{CI}}_n^{\dagger} := \left\{\theta \in \Theta : \widehat{\mathbb{M}}_n(\theta) - \widehat{\mathbb{M}}_n(\widehat{\theta}_1) \leq 0\right\},$$
$$\widehat{\mathrm{CI}}_{n,\alpha} := \left\{\theta \in \Theta : \widehat{\mathbb{M}}_n(\theta) - \widehat{\mathbb{M}}_n(\widehat{\theta}_1) \leq \frac{z_{\alpha/2}\widehat{\sigma}(\theta, \widehat{\theta}_1)}{n^{1/2}}\right\}, \quad (1)$$

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 $m(\theta, Z_i) - m(\widehat{\theta}_1, Z_i), 1 \leq i \leq n$.

⋆ Clearly,

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

⋆ 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

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

  where

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

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

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

- ⋆ 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

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

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

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

⋆ This would be a penalized/constrained reweighted least squares estimator and can be efficiently computed.

⋆ However, the limiting distribution 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 (arXiv:2407.12278)

Z-estimation Problems

- ⋆ 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 $\Theta \subseteq \mathbb{R}^d$).
- ⋆ 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.
- ⋆ Clearly, for any vector $a \in \mathbb{R}^d$, we have $\mathbb{E}[a^\top \Psi(\theta(P), Z)] = 0$, which implies

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

  is an asymptotically valid confidence set. Of course, this is inefficient in $d-1$ dimensions (i.e., its Lebesgue measure is infinite), in general.

Z-estimation Problems

⋆ Instead, split the data into two parts and compute estimator $\widehat{\theta}$ from first half. Then compute this set with $a = \widehat{\theta} - \theta$, which makes it full dimensional:

$$\widehat{\mathrm{CI}}_{n,\alpha} = \left\{ \theta \in \Theta : \frac{|\sum_{i=1}^{n} (\widehat{\theta} - \theta)^\top \Psi(\theta, Z_i)|}{\sqrt{\sum_{i=1}^{n} ((\widehat{\theta} - \theta)^\top \Psi(\theta, Z_i))^2}} \leq z_{\alpha/2} \right\}.$$

⋆ If

$$\sup_{a \in \mathbb{R}^d} \frac{\mathbb{E}[|a^\top \Psi(\theta(P), Z)|^3]}{(\mathbb{E}[|a^\top \Psi(\theta(P), Z)|^2)^{3/2}} \leq L,$$

then

$$\limsup_{n \to \infty} \mathbb{P}(\theta(P) \notin \widehat{\mathrm{CI}}_{n,\alpha}) \leq \alpha.$$

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.
- ⋆ Robust estimation does not readily imply robust inference.
- ⋆ 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.
- ⋆ Robust estimation does not readily imply robust inference.
- ⋆ 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!