Adaptive Inference Techniques for Some Irregular Problems

Inference for Linear Regression

Arun Kumar Kuchibhotla

13 June, 2025

Carnegie Mellon University

Collaborators

Kenta Takatsu (CMU)

Woonyoung Chang (CMU)

Table of contents

[1] Joint work with Woonyoung Chang (arXiv:2407.12278)

Traditional inference framework

Inference: confidence intervals

- The construction of confidence sets for functionals is a standard problem in statistics.
- Suppose $\theta(P), P \in \mathcal{P}$ is a functional of interest, for example, 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.

Example: Linear Regression

- $\star$ Suppose $(X, Y) \in \mathbb{R}^{d+1}$ is a random vector from a distribution P and we are interested in the projection parameter $\theta_0 = \theta(P)$ defined

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

- $\star$ Because of unconstrained optimization, $\theta(P)$ is also the solution to the equation

$$\mathbb{E}_P[X(Y - X^\top \theta)] = 0.$$

- $\star$ Using IID data $(X_i, Y_i), 1 \leq i \leq n$, $\theta(P)$ can be estimated using

$$\widehat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2.$$

Example: Linear Regression

⋆ Suppose $(X, Y) \in \mathbb{R}^{d+1}$ is a random vector from a distribution P and we are interested in the projection parameter $\theta_0 = \theta(P)$ defined

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

⋆ Because of unconstrained optimization, $\theta(P)$ is also the solution to the equation

$$\mathbb{E}_P[X(Y - X^\top \theta)] = 0.$$

⋆ Using IID data $(X_i, Y_i), 1 \le i \le n$, $\theta(P)$ can be estimated using

$$\widehat{\theta}_n = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2.$$

⋆ For a fixed d, assuming the invertibility of $\Sigma = \mathbb{E}[XX^\top]$, as $n \to \infty$,

$$n^{1/2}(\widehat{\theta}_n - \theta(P)) \ \stackrel{d}{\to}\ N(0, \Sigma^{-1} V \Sigma^{-1}),$$

where $V = \mathbb{E}[XX^\top (Y - X^\top \theta(P))^2]$; no linear model or Gaussianity.

Wald Inference: Linear Regression

- The asymptotic variance can be estimated as $\widehat{\Sigma}^{-1}\widehat{V}\widehat{\Sigma}^{-1}$.
- For any vector $c \in \mathbb{R}^d$, the Wald confidence interval for $c^\top\theta(P)$ can be obtained as

$$\widehat{\mathrm{CI}}_{n,\alpha}(c) := \left[c^\top\widehat{\theta}_n \pm z_{\alpha/2}\left(\frac{c^\top\widehat{\Sigma}^{-1}\widehat{V}\widehat{\Sigma}^{-1}c}{n}\right)^{1/2}\right].$$

- Again with d fixed, as $n \to \infty$, this confidence interval has an asymptotic coverage of $1-\alpha$.
- This nicety fails when dimensions grow rapidly or when constraints are placed on the projection parameter.

Failure of traditional inference: Increasing dimension

Asymptotics: Increasing dimension

- ⋆ With some algebraic manipulation, the OLS estimator satisfies

$$\widehat{\theta}_n - \theta(P) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Sigma}^{-1} X_i (Y_i - X_i^\top \theta(P)).$$

- ⋆ Asymptotic normality is claimed, for fixed d, by replacing $\widehat{\Sigma}^{-1}$ with Σ^{-1} with "negligible" error:

$$\begin{aligned}\widehat{\theta}_n - \theta(P) &= \frac{1}{n}\sum_{i=1}^{n} \Sigma^{-1} X_i (Y_i - X_i^\top \theta(P)) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} (\widehat{\Sigma}^{-1} - \Sigma^{-1}) X_i (Y_i - X_i^\top \theta(P)).\end{aligned}$$

- ⋆ The first term is mean zero and responsible for asymptotic normality, and for fixed d, the second term is negligible compared to the first.
- ⋆ But when the dimension is allowed to grow, the first term is of order $1/\sqrt{n}$ and the second is of order d/n.

Asymptotics: Increasing dimension

- ⋆ If $d = o(n^{1/2})$, then

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

  The asymptotic variance can be consistently estimated as if d were fixed.

- ⋆ If $d \gg n^{1/2}$, then

$$n^{1/2}(\widehat{\theta}_n - \theta(P) - B(P)) \overset{d}{\approx} N(0, \Sigma^{-1} V \Sigma^{-1}),$$

  where

$$B(P) = n^{-1}\mathbb{E}[\Sigma^{-1}(XX^\top - \Sigma)\Sigma^{-1}X(Y - X^\top\theta(P))].$$

- ⋆ Interestingly, $B(P) = 0$ if $\mathbb{E}[Y|X] = X^\top\theta(P)$.

Inference with increasing dimension

- If $\widehat{B}_n$ is a consistent estimator for $B(P)$ satisfying

$$n^{1/2}(\widehat{B}_n - B(P)) = o_p(1),$$

then the debiased estimator $\widehat{\theta}_n^{\texttt{debias}} = \widehat{\theta}_n - \widehat{B}_n$ satisfies

$$n^{1/2}(\widehat{\theta}_n^{\texttt{debias}} - \theta(P)) \overset{d}{\approx} N(0, \Sigma^{-1} V \Sigma^{-1}).$$

- Even if such a bias estimator exists, traditional inference still relies on estimating the variance.

- Unfortunately, consistent bias estimation may not be possible for all of $d = o(n)$. Chang et al. (2023) proposed a "good" bias estimator when $d = o(n^{2/3})$, and also proved the consistency of the classical variance estimator.

- Hence, traditional Wald inference is only valid for $d = o(n^{2/3})$. We do not know the limiting distribution of $\widehat{\theta}_n^{\texttt{debias}}$ for $d \gg n^{2/3}$.

Failure of traditional inference: Constraints

With constraints

- ⋆ Summarizing the unconstrained case, we do not know of an estimator for $\theta(P)$ with a tractable (estimable) limiting distribution for all of $d = o(n)$.
- ⋆ The situation is much worse with constraints, even if d is fixed as $n \to \infty$.
- ⋆ Suppose
$$\theta(P) = \arg\min_{\theta \in \Theta} \mathbb{E}[(Y - X^\top \theta)^2],$$
for some set $\Theta \subseteq \mathbb{R}^d$.
- ⋆ The limiting distribution of the sample estimator $\widehat{\theta}_n$ is highly dependent on the regularity of $\theta(P)$ with respect to Θ. The limit could be a projected Gaussian; see Pflug (1995), Geyer (1994), and Shapiro (2000).

With constraints

- Summarizing the unconstrained case, we do not know of an estimator for $\theta(P)$ with a tractable (estimable) limiting distribution for all of $d = o(n)$.
- The situation is much worse with constraints, even if d is fixed as $n \to \infty$.
- Suppose
$$\theta(P) = \underset{\theta \in \Theta}{\arg\min}\ \mathbb{E}[(Y - X^\top \theta)^2],$$
for some set $\Theta \subseteq \mathbb{R}^d$.
- The limiting distribution of the sample estimator $\widehat{\theta}_n$ is highly dependent on the regularity of $\theta(P)$ with respect to Θ. The limit could be a projected Gaussian; see Pflug (1995), Geyer (1994), and Shapiro (2000).
- If Θ is a closed convex set, then $\theta(P)$ is characterized by
$$(\theta - \theta(P))^\top \mathbb{E}[X(Y - X^\top \theta(P))] \le 0 \quad \text{for all} \quad \theta \in \Theta.$$

Examples with constraints

- Examples with constraints are relevant in practice.
- **Sparsity** inducing least squares:

$$\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq t\},$$

or

$$\Theta = \left\{\theta \in \mathbb{R}^d : \sum_{j=1}^{k} \|\theta_{G_j}\|_2 \leq t\right\}.$$

- **Shape** inducing least squares:

$$\Theta = \{\theta \in \mathbb{R}^d : \theta \succeq 0\},$$

or

$$\Theta = \{\theta \in \mathbb{R}^d : \Delta_1\theta \succeq 0\},$$

where $\Delta_1\theta$ yields the first order differences of θ; e.g., $(\Delta_1\theta)_1 = \theta_2 - \theta_1$.

New Approach: Self-normalization[a]

[a] Joint work with Woonyoung Chang (arXiv:2407.12278)

Without constraints

- ⋆ Without constraints, $\theta(P)$ solves the equation

$$\mathbb{E}_P[\psi(Z;\theta(P))] = 0, \quad \text{where} \quad \psi(Z;\theta) = X(Y - X^\top\theta).$$

Hence, $u^\top\psi(Z;\theta(P))$ is a mean zero random variable for any $u \in \mathbb{R}^d$.

- ⋆ This implies that

$$\mathrm{CI}_{n,\alpha}(u) := \left\{\theta \in \mathbb{R}^d : \frac{|\sum_{i=1}^n u^\top\psi(Z_i;\theta)|}{\sqrt{\sum_{i=1}^n (u^\top\psi(Z_i;\theta))^2}} \le z_{\alpha/2}\right\},$$

is an asymptotically valid $(1-\alpha)$ confidence set. In fact, for all $u \in \mathbb{R}^d$ and $n \ge 1$,

$$\mathbb{P}(\theta(P) \notin \mathrm{CI}_{n,\alpha}(u)) \le \alpha + \frac{1}{\sqrt{n}} \times \frac{\mathbb{E}_P[|u^\top\psi(Z;\theta(P))|^3]}{(\mathbb{E}_P[(u^\top\psi(Z;\theta(P))^2])^{3/2}}.$$

- ⋆ This proves dimension-agnostic validity guarantee and holds for any Z-estimation problem. Note: no variance estimation, no bootstrap, no rate of convergence are needed.

Without constraints

- ⋆ Although valid, this confidence set is not practically viable because it is unbounded in all but one direction. This is useful for inference for linear contrasts.
- ⋆ This comes from the fact that $\mathbb{E}_P[u^\top \psi(Z;\theta)] = 0$ does not imply $\mathbb{E}_P[\psi(Z;\theta)] = 0$.
- ⋆ Alternatively, vectors u that depend on θ yield bounded confidence sets. Formally,

$$\widehat{\mathrm{CI}}_{n,\alpha}^{*} := \left\{\theta \in \mathbb{R}^d : \frac{|\sum_{i=1}^{n}(\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\theta)|}{\sqrt{\sum_{i=1}^{n}((\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\theta))^2}} \le z_{\alpha/2}\right\},$$

  is also an asymptotically valid $(1-\alpha)$ confidence set. Here, $\widetilde{\theta}_1$ is any estimator independent of $Z_1, \ldots, Z_n$.
- ⋆ The validity does not depend on the consistency of $\widetilde{\theta}_1$, but the diameter depends on it.

Without constraints

- In the context of linear regression, this confidence set is easy to compute because it is a quadratic inequality.

- It is clear that

$$\widetilde{\theta}_1, \widehat{\theta}_n \in \widehat{\mathrm{CI}}^*_{n,\alpha}.$$

  Hence, the diameter of the confidence set cannot shrink faster than the rate of convergence of the Z-estimator.

- Chang and Kuchibhotla (2025) prove that, for linear regression,

$$\mathrm{diam}(\widehat{\mathrm{CI}}^*_{n,\alpha}) = O_p\left(\sqrt{d/n}\right).$$

  Similar result holds for GLMs.

- For the functional of interest $c^\top \theta(P)$, we propose

$$c^\top \left(\widehat{\mathrm{CI}}^*_{n,\alpha/n} \cap \widehat{\mathrm{CI}}_{n,\alpha}(\widetilde{\Sigma}^{-1} c)\right),$$

  as the confidence set. This has dimension-agnostic validity and, moreover, its diameter scales as $n^{-1/2} + d/n$.

With constraints

⋆ The approach can be seamlessly extended to the case with constraints. Recall that if Θ is a closed convex set and $\widetilde{\theta}_1 \in \Theta$ is some initial estimator, then

$$(\widetilde{\theta}_1 - \theta(P))\mathbb{E}_P[X(Y - X^\top \theta(P))] \leq 0.$$

⋆ Hence, a valid confidence set for $\theta(P)$ is

$$\widehat{\mathrm{CI}}^*_{n,\alpha} := \left\{ \theta \in \Theta : \frac{\sum_{i=1}^n (\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\theta)}{\sqrt{\sum_{i=1}^n ((\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\theta))^2}} \leq z_{\alpha/2} \right\},$$

⋆ Once again, the validity is agnostic to the dimension d. The study of the diameter is in progress.

⋆ Similarly, confidence intervals can be constructed for $c^\top \theta(P)$ for any $c \in \mathbb{R}^d$.

Comment: Assumptions

Set $\Sigma = \mathbb{E}[XX^\top]$ and $V = \mathbb{E}[XX^\top(Y - X^\top\theta_0)^2]$.

(LM1) There exist $q_x \geq 8, q_y, K_x, K_y \geq 1$ such that

$$\sup_{u \in \mathbb{S}^{d-1}} \mathbb{E}[|u^\top \Sigma^{-1/2} X|^{q_x}] \leq K_x^{q_x},$$

and

$$\mathbb{E}[|Y - X^\top \theta(P)|^{q_y}] \leq K_y^{q_y}.$$

Moreover, $q_{xy} := (1/q_x + 1/q_y)^{-1} \geq 4$,

(LM2) There exist positive constants $\underline{\lambda}_\Sigma$, $\overline{\lambda}_\Sigma$, $\underline{\lambda}_V$, and $\overline{\lambda}_V$ such that

$$0 < \underline{\lambda}_\Sigma \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \overline{\lambda}_\Sigma < \infty$$

and

$$0 < \underline{\lambda}_V \leq \lambda_{\min}(V).$$

Comment: General Z-estimators

- In a more general context of Z-estimation (beyond linear regression), we have $\theta(P)$ defined by

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

for some estimating function $\psi(Z;\cdot)$.

- The proposed confidence set

$$\widehat{\mathrm{CI}}^*_{n,\alpha} := \left\{ \theta \in \mathbb{R}^d : \frac{|\sum_{i=1}^n (\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\theta)|}{\sqrt{\sum_{i=1}^n ((\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\theta))^2}} \le z_{\alpha/2} \right\},$$

continues to be an asymptotically valid $(1-\alpha)$-confidence set.

- However, this is analytically and computationally intractable for general ψ. Tractability can be improved using the initial estimator $\widetilde{\theta}_1$.

- Define the alternative confidence set

$$\widehat{\mathrm{CI}}^*_{n,\alpha} := \left\{ \theta \in \mathbb{R}^d : \frac{|\sum_{i=1}^n (\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\theta)|}{\sqrt{\sum_{i=1}^n ((\widetilde{\theta}_1 - \theta)^\top \psi(Z_i;\widetilde{\theta}_1))^2}} \le z_{\alpha/2} \right\},$$

Conclusions

Conclusions

- ⋆ Construction of valid confidence sets can be difficult even for seemingly innocuous functionals.
- ⋆ For the linear regression problem, our confidence sets are valid regardless of dimension and have a minimax diameter of $\sqrt{d/n}$.
- ⋆ This continues to hold for GLMs as well, including logistic regression.
- ⋆ Our proposal can be seamlessly extended to problems with constraints for which asymptotic limit theory is still unavailable.
- ⋆ For linear contrasts (one-dimensional functionals), our self-normalization confidence set has a diameter of order $n^{-1/2} + d/n$. In contrast, our debiasing approach yields a confidence interval with the width of $n^{-1/2}$ whenever $d = o(n^{2/3})$.
- ⋆ Characterizing the minimax width of confidence sets for linear contrasts is of interest.