Adaptive and Asymptotically Valid Inference for Regularized Problems

Arun Kumar Kuchibhotla

2 November, 2026

Carnegie Mellon University

Collaborators



Kenta Takatsu (Columbia University)

Preliminary version: https://arxiv.org/pdf/2501.07772.

Table of contents

1. Traditional inference framework

2. Our Approach: Risk Inversion

3. An Application: Regularized Linear Regression

4. Conclusions

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.

Inference for M-estimators

- ⋆ A general class of interesting functionals appear as optimizers of risk functions.

- ⋆ For a class of distributions $\mathcal{P}$, and $P \in \mathcal{P}$, let $\theta \mapsto \mathbb{M}_P(\theta)$ a risk function. Define the functional

$$\theta(P) := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \mathbb{M}_P(\theta).$$

  This can be a subset of $\mathbb{R}^d$, not just a singleton.

- ⋆ If $\theta \mapsto \widehat{\mathbb{M}}_n(\theta)$ is an estimate of the risk, then a natural estimator of $\theta(P)$ can be defined as

$$\widehat{\theta}_n := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \widehat{\mathbb{M}}_n(\theta).$$

- ⋆ When both the target and estimator are (almost surely) singleton, inference usually proceeds by proving some rate of convergence r_n and limiting distribution for $r_n(\widehat{\theta}_n - \theta(P))$.

With constraints

- ⋆ The strategy is rather difficult to implement with constraints, even if d is fixed as $n \to \infty$.

- ⋆ Suppose

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

$$= \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_P[(Y - X^\top \theta)^2] + h(\theta), \quad h(\theta) = \begin{cases} 0, & \text{if } \theta \in \Theta, \\ +\infty, & \text{otherwise,} \end{cases}$$

  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 (e.g., equality of active constraints under perturbation).

- ⋆ The limit could be a projected Gaussian; see Pflug (1995), Geyer (1994), and Shapiro (2000).

Examples with constraints

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

$$\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq k\},$$

or

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

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., $(\Delta_1\theta)_1 = \theta_2 - \theta_1$.

More generally ...

⋆ More generally, one can consider regularized/penalized risk functions:

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

for a proper closed convex function $h(\cdot)$ that is allowed to take $+\infty$ values.

⋆ One can also consider general smoother risk function:

$$\theta(P) := \underset{\theta \in \mathbb{R}^d}{\arg\min}\ \mathbb{R}_P(\theta) + h(\theta),$$

for a smooth function $\theta \mapsto \mathbb{R}_P(\theta)$.

⋆ The idea of using a penalty or a constraint set is to allow for much larger dimension d than the sample size.

⋆ In this case, how should inference proceed?

Our Approach: Risk Inversion

Recall the Setting

- ⋆ For any distribution $P \in \mathcal{P}$, define the functional of interest

$$\theta(P) \;:=\; \arg\min_{\theta \in \Theta} \mathbb{M}_P(\theta).$$

  Here Θ can be infinite-dimensional and $\theta(P)$ can be a set.

- ⋆ Clearly, for any estimator $\widetilde{\theta}_1 \in \Theta$,

$$\theta(P) \subseteq \left\{\theta \in \Theta : \mathbb{M}(\theta) - \mathbb{M}(\widetilde{\theta}_1) \leq 0\right\}.$$

- ⋆ This is a probability one statement.

- ⋆ Of course, the right hand set is not computable based on the data. However, we can construct confidence sets on the basis of this intuition.

- ⋆ **Weak miscoverage:**

$$\mathrm{MC}(\widehat{\mathrm{CI}}_{n,\alpha}) = \sup_{\theta^* \in \theta(P)} \mathbb{P}_P\left(\theta^* \notin \widehat{\mathrm{CI}}_{n,\alpha}\right).$$

Confidence Set: first construction

- ⋆ Randomly split $\{1, 2, \ldots, 2n\}$ into two equal parts $\mathcal{I}_1$ and $\mathcal{I}_2$.
- ⋆ Compute any estimator $\widetilde{\theta}_1$ using $\{Z_i : i \in \mathcal{I}_1\}$.
- ⋆ On the second half, compute some estimator of the risk $\theta \mapsto \widehat{\mathbb{M}}_n(\theta)$.
- ⋆ Construct the set

$$\widehat{\mathrm{CI}}_n := \left\{\theta \in \Theta : \widehat{\mathbb{M}}_n(\theta) - \widehat{\mathbb{M}}_n(\widetilde{\theta}_1) \leq 0\right\}. \tag{1}$$

- ⋆ Compare with the (corresponding) inactionable set

$$\left\{\theta \in \Theta : \mathbb{M}(\theta) - \mathbb{M}(\widetilde{\theta}_1) \leq 0\right\}.$$

- ⋆ We are simply replacing unknown risk with estimated risk.

History

Statistics	Operations Research
(1) Without sample splitting	
Stein (1981)	Pflug (1991, 1995, 2003)
Li (1981)	Vogel (2008)
Beran (1996)	Vogel and Seeger (2017)
Geyer (1996)	Guigues et al. (2017)
Beran and Dumbgen (1998)	
(2) With sample splitting	
Robins and van der Vaart (2006)	
Hoffmann and Nickl (2011)	
Carpentier (2013)	
Chakravarti et al. (2019)	
Kim and Ramdas (2024)	
Park et al. (2023)	

Validity

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

$$\mathrm{MC}(\widehat{\mathrm{CI}}_n) \;\le\; \sup_{\theta^* \in \theta(P)} \mathbb{E}\left[\frac{\sigma^2(\widetilde{\theta}_1)}{\sigma^2(\widetilde{\theta}_1) + n\mathbb{C}^2(\widetilde{\theta}_1)}\right],$$

where

$$\sigma^2(\theta') := \mathrm{Var}_P(m(\theta^*, Z) - m(\theta', Z)),$$
$$\mathbb{C}(\theta') := \mathbb{E}[m(\theta', Z)] - \min_{\theta \in \Theta} \mathbb{E}[m(\theta, Z)] \;\ge\; 0.$$

⋆ Note that if $\widetilde{\theta}_1$ is inconsistent for $\theta(P)$, i.e., $\mathrm{dist}(\widetilde{\theta}_1, \theta(P)) \neq o_p(1)$ as $n \to \infty$, then

$$\mathrm{MC}(\widehat{\mathrm{CI}}_n) \;\to\; 0, \quad \text{as} \quad n \to \infty.$$

⋆ In fact, if $\sigma^2(\theta') \le \|\theta^* - \theta'\|^{2\eta}$ and $\mathbb{C}(\theta') \ge \|\theta^* - \theta'\|^{1+\beta}$, then

$$\mathrm{MC}(\widehat{\mathrm{CI}}_n) \le \sup_{\theta^* \in \theta(P)} \mathbb{E}\left[\left(1 + n\|\widetilde{\theta}_1 - \theta^*\|^{2+2\beta-2\eta}\right)^{-1}\right].$$

Validity

- ⋆ There is a much sharper result in the manuscript which shows that even if $\widetilde{\theta}_1$ converges to $\theta(P)$ at fast rate, its miscoverage is in most cases at most $1/2$ (asymptotically).
- ⋆ A confidence interval with asymptotic miscoverage at most $1/2$ can always be boosted to construct a new confidence interval with asymptotic miscoverage of $1-\alpha$ for any $\alpha \in (0, 1/2)$, by confidence boosting (next slide).
- ⋆ In the manuscript, we also prove that these sets are minimax rate optimal for a large class of problems.
- ⋆ For brevity, in the following, we discuss the properties of $\widehat{\mathrm{CI}}_n$ for regularized problems.

Confidence Boosting

- ⋆ Fix $\alpha \in (0, 1/2)$. Set $B = \lceil \log_2(2/\alpha) \rceil$.
- ⋆ Suppose $\widehat{\mathrm{CI}}^{(j)}, 1 \le j \le B$ are independent confidence sets for a functional $\theta(P)$ such that

$$\limsup_{n\to\infty} \max_{1\le j\le B} \mathbb{P}_P \left(\theta(P) \notin \widehat{\mathrm{CI}}^{(j)} \right) \le \frac{1}{2}.$$

- ⋆ Then

$$\widehat{\mathrm{CI}} := \bigcup_{j=1}^{B} \widehat{\mathrm{CI}}^{(j)},$$

satisfies

$$\limsup_{n\to\infty} \mathbb{P}_P \left(\theta(P) \notin \widehat{\mathrm{CI}} \right) \le \alpha.$$

- ⋆ One can get independent confidence sets by simply splitting independent data into B non-overlapping batches.

An Application: Regularized Linear Regression

Regularized Linear Regression

- ⋆ Suppose $h : \mathbb{R}^d \to (-\infty, \infty]$ is a proper (lsc) convex function. Consider

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

- ⋆ If $\mathbb{E}_P[XX^\top]$ is non-singular, then $\theta(P)$ is a singleton.

- ⋆ Suppose we have $2n$ IID random vectors, split as

$$(X_i, Y_i), i \in \mathcal{I}_1 \quad \text{and} \quad (X_i, Y_i), i \in \mathcal{I}_2,$$

  with $|\mathcal{I}_1| = |\mathcal{I}_2| = n$.

- ⋆ Set

$$\widehat{R}_n(\theta) := \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} (Y_i - X_i^\top \theta)^2.$$

- ⋆ Define

$$\widehat{\mathrm{CI}}_n := \left\{ \theta : \ \widehat{R}_n(\theta) + h(\theta) \leq \widehat{R}_n(\widetilde{\theta}_1) + h(\widetilde{\theta}_1) \right\}.$$

  This is a bounded set if $h(\cdot)$ has bounded level sets.

Structure of the Set

⋆ Let

$$\widehat{\Sigma}_2 := \frac{1}{|\mathcal{I}_2|}\sum_{i\in\mathcal{I}_2} X_i X_i^\top \quad \text{and} \quad \widehat{\theta}^{\mathrm{LS}} \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \sum_{i\in\mathcal{I}_2} (Y_i - X_i^\top\theta)^2.$$

(Choose any one if there is more than one minimizer.)

⋆ Then

$$\widehat{\mathrm{CI}}_n := \{\theta\in\mathbb{R}^d : \|\theta - \widehat{\theta}^{\mathrm{LS}}\|^2_{\widehat{\Sigma}_2} + h(\theta) \le \|\widehat{\theta}^{\mathrm{LS}} - \widetilde{\theta}_1\|^2_{\widehat{\Sigma}_2} + h(\widetilde{\theta}_1)\}.$$

⋆ In particular, if $\widetilde{\theta}_1 \in K$ and

$$h(\theta) = \begin{cases} 0, & \text{if } \theta\in K, \\ +\infty, & \text{if } \theta\notin K \end{cases} \quad \Rightarrow \quad \widehat{\mathrm{CI}}_n = \mathcal{B}_{\widehat{\Sigma}_2}(\widehat{\theta}^{\mathrm{LS}}, \|\widehat{\theta}^{\mathrm{LS}} - \widetilde{\theta}_1\|_{\widehat{\Sigma}_2}) \cap K,$$

and generally,

$$\widehat{\mathrm{CI}}_n \subseteq \begin{cases} \mathcal{B}_{\widehat{\Sigma}_2}\left(\widehat{\theta}^{\mathrm{LS}}, \sqrt{\|\widehat{\theta}^{\mathrm{LS}} - \widetilde{\theta}_1\|^2_{\widehat{\Sigma}_2} + h(\widetilde{\theta}_1)}\right), \\ \{\theta : h(\theta) \le \|\widehat{\theta}^{\mathrm{LS}} - \widetilde{\theta}_1\|^2_{\widehat{\Sigma}_2} + h(\widetilde{\theta}_1)\}. \end{cases}$$

Shape of the Set

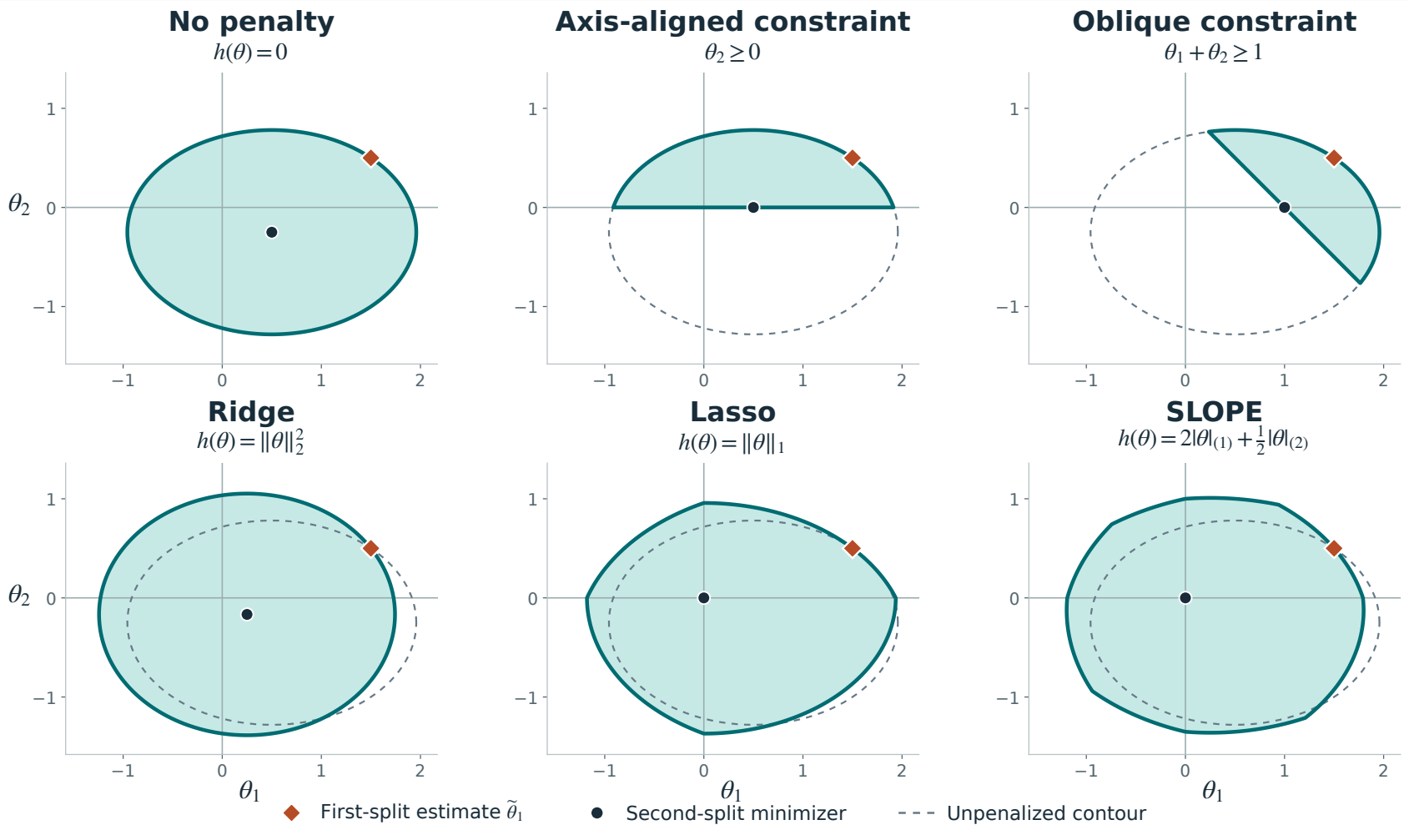


Figure 1: Shape of the set for different penalties

Validity

- ⋆ Suppose $\Sigma_P = \mathbb{E}_P[XX^\top]$ and suppose, for all $v \in \mathbb{R}^d$,

$$\begin{aligned} \mathbb{E}_P[(Y - X^\top \theta(P))^2 (X^\top v)^2] &\leq L v^\top \Sigma_P v, \quad \text{and} \\ \mathbb{E}_P[(X^\top v)^4] &\leq K (v^\top \Sigma_P v)^2. \end{aligned}$$

- ⋆ Let $\mathcal{P}_{K,L}$ be the collection of all such distributions. Then

$$\mathbb{P}_P(\theta(P) \notin \widehat{\mathrm{CI}}_n) \leq \mathbb{E}_P \left[\frac{8L + 2K\|\widetilde{\theta}_1 - \theta(P)\|^2_{\Sigma_P}}{8L + (n + 2K)\|\widetilde{\theta}_1 - \theta(P)\|^2_{\Sigma_P}} \right].$$

- ⋆ In particular, if $n\|\widetilde{\theta}_1 - \theta(P)\|^2_{\Sigma_P} \to \infty$ (in probability), then miscoverage converges to zero.

Validity

- ⋆ Suppose $\Sigma_P = \mathbb{E}_P[XX^\top]$ and suppose, for all $v \in \mathbb{R}^d$,

$$\mathbb{E}_P[(Y - X^\top \theta(P))^2 (X^\top v)^2] \;\leq\; L v^\top \Sigma_P v, \quad \text{and}$$
$$\mathbb{E}_P[(X^\top v)^4] \;\leq\; K (v^\top \Sigma_P v)^2.$$

- ⋆ Let $\mathcal{P}_{K,L}$ be the collection of all such distributions. Then

$$\mathbb{P}_P(\theta(P) \notin \widehat{\mathrm{CI}}_n) \;\leq\; \mathbb{E}_P \left[\frac{8L + 2K \|\widetilde{\theta}_1 - \theta(P)\|_{\Sigma_P}^2}{8L + (n + 2K) \|\widetilde{\theta}_1 - \theta(P)\|_{\Sigma_P}^2} \right].$$

- ⋆ In particular, if $n\|\widetilde{\theta}_1 - \theta(P)\|_{\Sigma_P}^2 \to \infty$ (in probability), then miscoverage converges to zero.

- ⋆ With a more refined argument, it can be proved that

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{K,L}} \sup_{\widetilde{\theta}_1} \mathbb{P}_P(\theta(P) \notin \widehat{\mathrm{CI}}_n) \leq \frac{1}{2},$$

  irrespective of how fast/slow d changes with n.

- ⋆ Note that there is no condition here on the penalty $h(\cdot)$, except for convexity.

Diameter: lower bounds

- ⋆ The diameter of the confidence set depends on the "complexity" of $h(\cdot)$ and the consistency of $\widetilde{\theta}_1$.
- ⋆ Let us consider a few simpler lower bounds.
- ⋆ We know $\widetilde{\theta}_1 \in \widehat{\mathrm{CI}}_n$, the confidence set $\widehat{\mathrm{CI}}_n$ is asymptotically valid, and

$$\widehat{\theta}_2 \in \widehat{\mathrm{CI}}_n, \quad \text{where} \quad \widehat{\theta}_2 := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \widehat{R}_n(\theta) + h(\theta).$$

- ⋆ Therefore,

$$\mathrm{diam}(\widehat{\mathrm{CI}}_n) \;\geq\; \max\{\|\widehat{\theta}_2 - \widetilde{\theta}_1\|, \|\widetilde{\theta}_1 - \theta(P)\|, \|\widehat{\theta}_2 - \theta(P)\|\},$$

  with a positive probability.
- ⋆ Hence, the diameter cannot converge to zero unless $\widetilde{\theta}_1$ is consistent for $\theta(P)$. The suboptimality of the M-estimator implies the suboptimality of the confidence set.

Diameter: upper bounds

- ⋆ Without assumptions on $h(\cdot)$, we have

$$\text{diam}(\widehat{\mathrm{CI}}_n) = O_p\left(\sqrt{\frac{\text{tr}(\Sigma_P)}{n}} + \|\widetilde{\theta}_1 - \theta(P)\|\right),$$

- ⋆ If Σ_P is non-singular, then $\text{tr}(\Sigma_P) \geq \lambda_{\min} \times d$ and hence, this diameter bound is $O_p(\sqrt{d/n})$, in general.
- ⋆ This is the minimax rate if the penalty $h(\cdot)$ does not impose any structural restrictions.
- ⋆ However, penalties like LASSO, SLOPE, or nuclear norm regularization in matrix regression, impose sparsity.
- ⋆ This is a low-dimensional structure and hence, we expect a diameter rate that depends on this "low dimension".

Diameter: upper bounds for LASSO penalty

- ⋆ Suppose the observed data follows a sparse linear model:

$$Y_i = X_i^\top \theta^* + \xi_i,$$

  for a vector $\theta^* \in \mathbb{R}^d$ such that $\|\theta^*\|_0 \leq s$.

- ⋆ Without the knowledge of s, the LASSO estimator with a "good" choice of penalty parameter yields an estimator $\widehat{\theta}_n$ satisfying

$$\|\widehat{\theta}_n - \theta^*\|_2 = O_p\left(\sqrt{\frac{s\log(d)}{n}}\right).$$

- ⋆ Unfortunately, it is impossible to construct confidence sets for θ^* with diameter of this rate, unless s is known.

- ⋆ It is important to recall that our target of inference is NOT θ^* but

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

- ⋆ Our confidence set is valid for $\theta(P)$, not θ^*.

Diameter: upper bounds for LASSO penalty

⋆ Suppose the observed data follows a sparse linear model:

$$Y_i = X_i^\top \theta^* + \xi_i,$$

for a vector $\theta^* \in \mathbb{R}^d$ such that $\|\theta^*\|_0 \leq s$.

⋆ If $h(\theta) = \lambda_n \|\theta\|_1$ with $\lambda_n \asymp \sqrt{\log(d)/n}$ and $\widetilde{\theta}_1$ is the LASSO estimator, then

$$\text{diam}(\widehat{\text{CI}}_n) = O_p\left(\sqrt{\frac{s\log(d)}{n}}\right).$$

⋆ Note that our confidence set does not require the knowledge of s and the choice of $\lambda = \lambda_n$ also does not require this knowledge.

⋆ Hence, the resulting confidence set adapts to the sparsity and is minimax rate optimal.

Diameter: upper bounds for SLOPE penalty

⋆ Consider the sparse linear setting before.

⋆ Consider the SLOPE penalty, i.e.,

$$h(\theta) = \sum_{j=1}^{d} \lambda_j |\theta|_{(j)},$$

for a sequence of penalties $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$.

⋆ If $\lambda_j \asymp \sqrt{\log(d/j)/n}$, and $\widetilde{\theta}_1$ is the SLOPE estimator, then

$$\text{diam}(\widehat{\text{CI}}_{n,\alpha}) = O_p\left(\sqrt{\frac{s\log(d/s)}{n}}\right).$$

⋆ As with LASSO, this confidence set adapts to unknown sparsity.

Conclusions

Conclusions

- ⋆ Construction of valid confidence sets can be difficult even for seemingly innocuous functionals.
- ⋆ We proposed confidence sets for M-estimation problems along with confidence boosting.
- ⋆ For regularized/constrained linear regression problems, these confidence sets are asymptotically valid irrespective of the dimension asymptotics.
- ⋆ Note that for problems with constraints the asymptotic limit theory is still unavailable.
- ⋆ When the sparse linear model is the true data generating process, then the resulting sets can be adaptive.
- ⋆ The resulting sets are for the entire parameter vector. There are refined sets for low-dimensional functions of the parameter vector but adaptivity is unknown.