11 Inference

So far we have mostly focused on estimation, e.g., constructing an estimator $\widehat{\mu}$ of a regression function μ, hopefully with some guarantee concerning its error

$$\sup_{P \in \mathcal{P}} \mathbb{E}\Big\{ d(\widehat{\mu}, \mu) \Big\} \leq C r_n,$$

where d could be the pointwise error $d(\widehat{\mu}, \mu) = |\widehat{\mu}(x) - \mu(x)|$, or the $L_2(\mathbb{P})$ norm $d(\widehat{\mu}, \mu) = \|\widehat{\mu} - \mu\|^2 = \int \{\widehat{\mu}(x) - \mu(x)\}^2 \, d\mathbb{P}(x)$, or the sup norm $d(\widehat{\mu}, \mu) = \|\widehat{\mu} - \mu\|_\infty = \sup_x |\widehat{\mu}(x) - \mu(x)|$, for example, and ideally the guarantee holds uniformly over all distributions P in our model $\mathcal{P}$. However, in practice we often do not only want an estimate $\widehat{\mu}$, we also want some quantification of the uncertainty of our estimate. Confidence sets are a natural way to formalize this.

1 Setup

A confidence set for a function μ is a set of functions promising to cover the true μ with some specified probability $1 - \alpha$, at least asymptotically, i.e.,

$$\inf_{P \in \mathcal{P}} P(\mu \in C_n) \geq 1 - \alpha + o(1).$$

Such a confidence set is said to be *honest*. Of course one could always take C_n to be arbitrarily large, ensuring $\mathbb{P}(\mu \in C_n) = 1$, so coverage by itself is not enough of a criterion. For meaningful/informative uncertainty quantification, we also want the set C_n to be as small as possible, while still being honest. A general measure of the size of a confidence set is the diameter

$$|C_n| = \sup\{d(f, g) : f, g \in C_n\}.$$

Note that the diameter is defined with respect to a particular distance d, and so is sometimes called the d-diameter.

A confidence band for μ is a set of functions of the form

$$C_n = \Big\{ f : f(x) \in [L(x), U(x)] \text{ for all } x \Big\}$$

for some lower and upper functions L, U. Confidence bands are probably the most useful and easy to interpret confidence sets. The simplest example of a confidence band is an L_∞ ball around some initial estimator $\widehat{\mu}$, i.e.,

$$\begin{aligned} C_n &= \Big\{ f : \|\widehat{\mu} - f\|_\infty \leq r_n \Big\} \\ &= \Big\{ f : f(x) \in [\widehat{\mu}(x) - r_n, \widehat{\mu}(x) + r_n] \ \text{ for all } x \Big\} \end{aligned}$$

which has diameter $2r_n$. L_2 confidence balls are also used, e.g.,

$$C_n = \Big\{ f : \|\widehat{\mu} - f\|_2 \leq r_n \Big\}$$

although they can be more difficult to visualize. Note that L_2 balls are larger than L_∞ balls with the same radius r_n, since if $\|\widehat{\mu} - f\|_\infty \leq r_n$ then $\|\widehat{\mu} - f\|_2 \leq \|\widehat{\mu} - f\|_\infty \leq r_n$.

An honest confidence set is said to be *minimax optimal* if its d-diameter $|C_n|$ shrinks at the minimax rate of estimation of μ in d-risk, i.e., if for all $\epsilon > 0$ there exists $M = M_\epsilon$ such that

$$\sup_{P \in \mathcal{P}} P(|C_n| > M r_n) < \epsilon$$

for all n, i.e., $|C_n| = O_P(1/r_n)$ uniformly over $\mathcal{P}$, where r_n is the minimax rate

$$\inf_{\widehat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \Big\{ d(\widehat{\mu}, \mu) \Big\} \asymp r_n.$$

For example if $\mathcal{P}$ is the model where μ is s-smooth, then $r_n = n^{-1/(2+d/s)}$ and so a minimax optimal interval would have diameter of this order. This definition of optimality is partly motivated by the fact that it is in some sense impossible for confidence sets to have smaller diameter, as proved for example on page 523-524 of Giné and Nickl [2021].

2 Constructing confidence sets

Here we show that an estimator coupled with a (known) optimal risk guarantee can be turned into an honest and optimal confidence set. Specifically, if for some estimator $\widehat{\mu}$ we have

$$\sup_{P \in \mathcal{P}} \mathbb{E} \Big\{ d(\widehat{\mu}, \mu) \Big\} \leq C r_n$$

where r_n is the minimax rate in d-risk, then the confidence set

$$C_n = \Big\{ f : d(\widehat{\mu}, f) \leq C r_n / \alpha \Big\}$$

is honest and minimax optimal. For honesty note that by Markov's inequality we have

$$\begin{aligned} \inf_{P \in \mathcal{P}} P(\mu \in C_n) &= 1 - \sup_{P \in \mathcal{P}} P \Big\{ d(\widehat{\mu}, \mu) > C r_n / \alpha \Big\} \\ &\geq 1 - \frac{\sup_{P \in \mathcal{P}} \mathbb{E}\{ d(\widehat{\mu}, \mu) \}}{C r_n / \alpha} \geq 1 - \alpha \end{aligned}$$

and similarly for minimax optimality, if we take $M = (2C/\alpha)/\epsilon$ we have

$$\sup_{P \in \mathcal{P}} P(|C_n| > M r_n) \leq \frac{\sup_{P \in \mathcal{P}} \mathbb{E}(|C_n|)}{M r_n} \leq \frac{2 C r_n / \alpha}{M r_n} \leq \epsilon.$$

Of course in practice the constant C and/or rate r_n may not be known; even if they were, the above Markov bound could be too imprecise, making the interval wider than necessary, and providing more than $1 - \alpha$ coverage.

2.1 Asymptotic distribution

A classic way of constructing confidence sets is to use the asymptotic distribution of the estimator, after centering and scaling (we resort to asymptotics since the finite sample distribution is often intractable). For example one might show that

$$\frac{\widehat{\mu}(x) - \mu(x)}{\widehat{\sigma}(x)}$$

converges in distribution to a standard normal, say, where $\widehat{\sigma}(x)$ is some estimate of the standard error of $\widehat{\mu}(x)$. Then asymptotically exact intervals for $\mu(x)$ would be given by

$$C_n = \Big[\widehat{\mu}(x) \pm z_{1-\alpha}\widehat{\sigma}(x)\Big].$$

For an analog of this story for confidence bands, one could show that

$$\frac{\widehat{\mu}(x) - \mu(x)}{\widehat{\sigma}(x)}$$

converges to a Gaussian process $W(x)$ so that intervals of the form

$$C_n = \Big[\widehat{\mu}(x) \pm c\widehat{\sigma}(x)\Big]$$

would have coverage

$$\mathbb{P}(\mu \in C_n) = \mathbb{P}\left\{\frac{|\widehat{\mu}(x) - \mu(x)|}{\widehat{\sigma}(x)} \leq c \;\text{ for all } x\right\} \to \mathbb{P}\Big\{\sup_x |W(x)| \leq c\Big\}$$

so one could pick $c = c_\alpha$ based on quantiles of the supremum of a Gaussian process, in order to ensure the above probability is $1-\alpha$. See for example Sun and Loader [1994].

Alternatively if one can derive a concentration inequality of the form

$$\mathbb{P}\Big\{\|\widehat{\mu} - \mu\|_\infty \geq \widehat{\tau}(z)\Big\} \leq 2\exp(-z)$$

for some $\widehat{\tau}(z)$ then the band

$$C_n = \Big[\widehat{\mu}(x) \pm \widehat{\tau}(\log(2/\alpha))\Big]$$

has coverage

$$\mathbb{P}(\mu \in C_n) = \mathbb{P}\Big\{\|\widehat{\mu} - \mu\|_\infty \leq \widehat{\tau}(\log(2/\alpha))\Big\} \geq 1 - \alpha.$$

For an example of this approach see Corollary 6.4.5 in Giné and Nickl [2021].

2.2 The bias problem

Unfortunately there is a real challenge with characterizing the distribution of something like

$$\widehat{\mu}(x) - \mu(x)$$

in nonparametric regression/density estimation settings, which Wasserman [2006] termed "the bias problem." The issue is the following. Note we can write

$$\frac{\widehat{\mu}(x) - \mu(x)}{\sqrt{\mathrm{var}\{\widehat{\mu}(x)\}}} = \underbrace{\frac{\widehat{\mu}(x) - \mathbb{E}\{\widehat{\mu}(x)\}}{\sqrt{\mathrm{var}\{\widehat{\mu}(x)\}}}}_{\text{mean-zero noise}} + \underbrace{\frac{\mathbb{E}\{\widehat{\mu}(x)\} - \mu(x)}{\sqrt{\mathrm{var}\{\widehat{\mu}(x)\}}}}_{\text{deterministic bias}}.$$

The first term will often converge to a standard normal distribution, depending on the estimator; however the second term, which is the bias scaled by the standard deviation, will typically behave like an unknown constant, since for optimal tuning we balance the squared bias with the variance. This is true even for large sample sizes.

As Wasserman [2006] states:

> The presence of this second, nonvanishing term introduces a bias into the Normal limit. The result is that the confidence interval will not be centered around the true function μ due to the smoothing bias $\mathbb{E}\{\widehat{\mu}(x)\} - \mu(x)$.

Wasserman [2006] suggests three ways forward, none of which are particularly satisfying:

1. live with it, i.e., give up on covering $\mu(x)$ and instead settle for covering $\mathbb{E}\{\widehat{\mu}(x)\}$,
2. estimate the bias term with $\widehat{b}(x)$ and instead use a debiased estimator $\widehat{\mu}(x) - \widehat{b}(x)$,
3. undersmooth, so that the squared bias is of smaller order than the variance, and so the second term vanishes with growing sample size.

The first option is unsatisfying since we are giving up on our original (admittedly ambitious) goal of confidence sets for μ itself, and settling instead for covering a somewhat unnatural quantity $\mathbb{E}\{\widehat{\mu}(x)\}$, which is dependent on the estimator we are using, etc. In some cases, this mean has some interpretation as a smoothed version of the regression function, but not always. The second option is unsatisfying since to estimate the bias one typically has to make stronger assumptions than are needed to estimate $\mu(x)$ itself (e.g., assume more smoothness to estimate higher-order derivatives), which is somewhat circular. The third option is unsatisfying because undersmoothing results in a suboptimal estimator (we are no longer balancing bias/variance), and because there are not really any practical data-driven ways of undersmoothing (we can always just divide an optimal bandwidth choice by $\log n$ but this is pretty arbitrary - why $\log n$ and not $C \log n$ or something else?).

2.3 Risk estimation

Another different approach to constructing confidence sets is based on estimating the risk $d(\widehat{\mu}, \mu)$ of an estimator directly, with sample splitting. This is especially useful when d is something like an L_2 norm that involves some averaging and so can be estimated at faster rates than μ itself. This method comes from Robins and Van Der Vaart [2006].

Split the sample into two parts D_1^n and D_2^n. First estimate $\widehat{\mu}$ on the first sample D_1^n, and then second let $\widehat{R}(\widehat{\mu})$ denote an estimator of the squared $L_2(\mathbb{P})$ norm

$$\|\widehat{\mu} - \mu\|_2^2 = \int \Big\{\widehat{\mu}(x) - \mu(x)\Big\}^2 \, d\mathbb{P}(x)$$

built from the second sample D_2^n. Note unbiased and root-n consistent estimators are available for

$$\int \widehat{\mu}(x)^2 \, d\mathbb{P}(x) \quad \text{and} \quad 2\int \mu(x)\widehat{\mu}(x) \, d\mathbb{P}(x)$$

and so it is only estimation of

$$\int \mu(x)^2 \, d\mathbb{P}(x)$$

that is challenging in this $L_2(\mathbb{P})$ norm. Then the proposed confidence set is given by

$$C_n = \Big\{f : \|\widehat{\mu} - f\|_2 \leq \sqrt{\widehat{R}(\widehat{\mu}) + z_\alpha \widehat{\tau}}\Big\}$$

where z_α and $\widehat{\tau}$ are suitable quantiles and scale estimators, respectively. Robins and Van Der Vaart [2006] show that for a particular U-statistic estimator of $\widehat{R}(\widehat{\mu})$, built using series regression with a basis of dimension k, under a Holder smoothness model, this confidence set is honest and its diameter is of the order

$$\begin{aligned} |C_n| &= O_{\mathbb{P}}\left(\sqrt{\widehat{R}(\widehat{\mu})} + \sqrt{\widehat{\tau}} + \frac{1}{\sqrt{n}}\right) \\ &= O_{\mathbb{P}}\left(\|\widehat{\mu} - \mu\| + \left(\frac{k^{1/4}}{\sqrt{n}} + k^{-s/d}\right) + \frac{1}{\sqrt{n}}\right). \end{aligned}$$

Balancing the second term in k amounts to taking

$$k^{(4s+d)/4d} \sim \sqrt{n} \implies k \sim n^{2d/(4s+d)} \implies k^{-s/d} \sim n^{-1/(2+d/2s)}.$$

Note this term is smaller than the estimation rate for μ when it is s-smooth; in fact it is the estimation rate for a function that is twice as smooth. This will be important for the next section. Therefore when μ is s-smooth the diameter is of the order

$$|C_n| = O_{\mathbb{P}}\left(\|\widehat{\mu} - \mu\| + n^{-1/(2+d/2s)}\right) = O_{\mathbb{P}}\left(n^{-1/(2+d/s)}\right),$$

due to the first term, which is indeed minimax optimal for the s-smooth model.

3 Adaptivity

We have seen in the estimation context that one can adapt to unknown smoothness or other structure using, for example, cross-validation-based model selection. Recall if we set up a grid of different tuning parameters we can achieve the optimal rate over varying smoothness classes, plus some typically smaller order gap term. Unfortunately, it turns out that the possibility of adaptation in the inference setting is much more limited.

Cai and Low [2004] and Genovese and Wasserman [2008] show results similar to the following, which is adapted from Giné and Nickl [2021].

Theorem 1. *Consider the model where μ is t-smooth, where $t \in \{r, s\}$ for some $s > r$. Assume C_n is an asymptotically honest confidence set over the full model, i.e.,*

$$\liminf_n \inf_{P \in \mathcal{P}_r} P(\mu \in C_n) \geq 1 - \alpha.$$

Then the L_∞ diameter cannot shrink at the minimax optimal rate, i.e., we do not have that for any $\epsilon > 0$ there is some M such that

$$\sup_{P \in \mathcal{P}_t} P\left(|C_n| > M\,(n/\log n)^{-1/(2+d/t)}\right) \leq \epsilon$$

for both $t \in \{r, s\}$. In fact any confidence set that is honest over $\mathcal{P}_r$ cannot satisfy $|C_n| = O_P(r_n)$ uniformly over $\mathcal{H}(s)$ for $s > r$ for any rate r_n faster than the slower rate $(n/\log n)^{-1/(2+d/r)}$.

In short the above result shows that adaptive confidence bands do not exist over the whole model $\mathcal{P}_r$, and that the diameter can at best shrink at the slowest rate $(n/\log n)^{-1/(2+d/r)}$, meaning we have to pay the full price of the slowest estimation rate over our model. Given the contrast to positive results for adaptive estimation, Genovese and Wasserman [2008] state, "We are left knowing that we are close to the true function but being unable to demonstrate it inferentially".

These negative results for bands have led people to consider different setups where more extensive adaptation is possible, essentially by moving the goalposts in some way. Some examples include:

1. Changing the target, e.g., from the regression function μ to some simpler surrogate μ^* [Genovese and Wasserman, 2008] .
2. Weakening the notion of coverage, e.g., to average coverage [Juditsky and Lambert-Lacroix, 2003] (also see below) or coverage over "most" x values [Cai et al., 2014].
3. Restricting the model, e.g., by adding self-similarity or other assumptions [Giné and Nickl, 2010].

Somewhat surprisingly, the story is slightly more optimistic for L_2 confidence balls. Consider again the setting where μ is t-smooth, where $t \in \{r, s\}$ for some $s > r$. And recall the bound on the L_2 diameter using the risk estimation method given by

$$|C_n| = O_{\mathbb{P}}\left(\|\widehat{\mu} - \mu\| + n^{-1/(2+d/2r)}\right).$$

Since adaptive estimation is possible we can construct an estimator satisfying for example

$$\sup_{\mu \in \mathcal{H}(t)} \mathbb{E}\|\widehat{\mu} - \mu\| \leq n^{-1/(2+d/t)}$$

for any unknown t. Therefore when μ is s-smooth, $s > r$, the diameter shrinks at the rate

$$|C_n| = O_{\mathbb{P}}\left(n^{-1/(2+d/s)} + n^{-1/(2+d/2r)}\right)$$

which matches the faster s-smooth rate when $s \leq 2r$. This indicates that adaptation is possible for L_2 confidence balls, albeit over the limited range $[r, 2r]$.

Another important fact about L_2 confidence balls is that they can be used to construct bands with average coverage. The next result illustrates this, following page 94 of Wasserman [2006].

Theorem 2. *Let $C_n(\alpha) = \{f : \|\widehat{\mu} - f\| \leq r_n(\alpha)\}$ denote an honest $L_2(\mathbb{P})$ confidence ball satisfying $\mathbb{P}(C_n(\alpha) \ni \mu) \geq 1 - \alpha$. Then the band*

$$\widehat{\mu}(x) \pm r_n(\alpha/2)\sqrt{2/\alpha}$$

has average coverage at least $1 - \alpha$, i.e.,

$$\int \mathbb{P}\left\{\mu(x) \in \left[\widehat{\mu}(x) \pm r_n(\alpha/2)\sqrt{2/\alpha}\right]\right\} \, d\mathbb{P}(x) \geq 1 - \alpha.$$

Proof. Note that the average coverage equals $\mathbb{P}\{\mu(X) \in [\widehat{\mu}(X) \pm r_n(\alpha/2)\sqrt{2/\alpha}]\}$ for an independent draw X. Let $\mathcal{E}$ denote the event that $\mu \in C_n(\alpha/2)$, and $\overline{\mathcal{E}}$ its complement, and also note that on $\mathcal{E}$ we have $\|\widehat{\mu} - \mu\| \leq r_n(\alpha/2)$. Then one minus the coverage is

$$\begin{aligned}
\mathbb{P}\Big\{\mu(X) \notin [\widehat{\mu}(X) \pm r_n(\alpha/2)\sqrt{2/\alpha}]\Big\} &= \mathbb{P}\Big\{|\widehat{\mu}(X) - \mu(X)| > r_n(\alpha/2)\sqrt{2/\alpha}\Big\} \\
&= \mathbb{P}\Big\{|\widehat{\mu}(X) - \mu(X)| > r_n(\alpha/2)\sqrt{2/\alpha} \ \cap \ \mathcal{E}\Big\} + \mathbb{P}\Big\{|\widehat{\mu}(X) - \mu(X)| > r_n(\alpha/2)\sqrt{2/\alpha} \ \cap \ \overline{\mathcal{E}}\Big\} \\
&\leq \mathbb{P}\Big\{|\widehat{\mu}(X) - \mu(X)| > r_n(\alpha/2)\sqrt{2/\alpha} \ \cap \ \mathcal{E}\Big\} + \mathbb{P}(\overline{\mathcal{E}}) \\
&\leq \frac{\mathbb{E}\{\mathbb{1}_{\mathcal{E}}|\widehat{\mu}(X) - \mu(X)|^2\}}{2r_n(\alpha/2)^2/\alpha} + \alpha/2 = \frac{\mathbb{E}\left(\mathbb{1}_{\mathcal{E}}\|\widehat{\mu} - \mu\|^2\right)}{2r_n(\alpha/2)^2/\alpha} + \alpha/2 \\
&\leq \frac{r_n(\alpha/2)^2}{2r_n(\alpha/2)^2/\alpha} + \alpha/2 = \alpha
\end{aligned}$$

where in the fourth line we used Markov's inequality and the fact that by definition $\mathbb{P}(\overline{\mathcal{E}}) = 1 - \mathbb{P}(\mu \in C_n(\alpha/2)) \leq 1 - \alpha/2$. □

References

T. T. Cai and M. G. Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32:1805–1840, 2004.

T. T. Cai, M. Low, and Z. Ma. Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association*, 109(507):1054–1070, 2014.

C. Genovese and L. Wasserman. Adaptive confidence bands. *The Annals of Statistics*, 36: 875–905, 2008.

E. Giné and R. Nickl. Confidence bands in density estimation. 2010.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2021.

A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Mathematical Methods of Statistics*, 12(4):410–428, 2003.

J. Robins and A. Van Der Vaart. Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34:229–253, 2006.

J. Sun and C. R. Loader. Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, pages 1328–1345, 1994.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.