Nonparametric Inference via Bootstrapping the Debiased Estimator

Yen-Chi Chen

Department of Statistics, University of Washington

ICSA-Canada Chapter Symposium 2017

Problem Setup

- Let $X_1, \cdots, X_n$ be an IID random sample from an unknown distribution function with a density function p.
- For simplicity, we assume p is supported on $[0,1]^d$.
- Goal: given a level α, we want to find $L_\alpha(x), U_\alpha(x)$ using the random sample such that

$$P\left(L_\alpha(x) \leq p(x) \leq U_\alpha(x) \ \forall x \in [0,1]^d\right) \geq 1 - \alpha + o(1).$$

- Namely, $[L_\alpha(x), U_\alpha(x)]$ forms an asymptotic simultaneous confidence band of $p(x)$.

Simple Approach: using the KDE

- A classical approach is to construct $L_\alpha(x), U_\alpha(x)$ using the kernel density estimator (KDE).
- Let
$$\widehat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$
be the KDE where $h > 0$ is the smoothing bandwidth and $K(x)$ is a smooth function such as a Gaussian.
- We pick t_α such that
$$L_\alpha(x) = \widehat{p}_h(x) - t_\alpha, \quad U_\alpha(x) = \widehat{p}_h(x) + t_\alpha.$$
As long as we choose t_α wisely, the resulting confidence band is asymptotically valid.

Simple Approach: the L_∞ Error

- How do we choose t_α to obtain a valid confidence band?
- A simple idea: inverting the L_∞ error.

Simple Approach: the L_∞ Error

- How do we choose t_α to obtain a valid confidence band?
- A simple idea: inverting the L_∞ error.
- Let $F_n(t)$ be the CDF of $\|\widehat{p}_h - p\|_\infty = \sup_x |\widehat{p}_h(x) - p(x)|$.
- Then the value $t^*_\alpha = F_n^{-1}(1-\alpha)$ has a nice property:

$$P(\|\widehat{p}_h - p\|_\infty \leq t^*_\alpha) = 1 - \alpha.$$

Simple Approach: the L_∞ Error

- How do we choose t_α to obtain a valid confidence band?
- A simple idea: inverting the L_∞ error.
- Let $F_n(t)$ be the CDF of $\|\widehat{p}_h - p\|_\infty = \sup_x |\widehat{p}_h(x) - p(x)|$.
- Then the value $t^*_\alpha = F_n^{-1}(1-\alpha)$ has a nice property:

$$P(\|\widehat{p}_h - p\|_\infty \leq t^*_\alpha) = 1 - \alpha.$$

- This implies

$$P(|\widehat{p}_h(x) - p(x)| \leq t^*_\alpha \ \forall x \in [0,1]^d) = 1 - \alpha.$$

- Thus,

$$L^*_\alpha(x) = \widehat{p}_h(x) - t^*_\alpha, \quad U^*_\alpha(x) = \widehat{p}_h(x) + t^*_\alpha$$

leads to a simultaneous confidence band.

Simple Approach: the Bootstrap - 1

- The previous method is great – it works even in a finite sample case.
- However, it has a critical problem: we do not know the distribution F_n! So we cannot compute the quantile.

Simple Approach: the Bootstrap - 1

- The previous method is great – it works even in a finite sample case.
- However, it has a critical problem: we do not know the distribution F_n! So we cannot compute the quantile.
- A simple solution: using the bootstrap (we will use the empirical bootstrap).

Simple Approach: the Bootstrap - 2

- Let $X_1^*, \cdots, X_n^*$ be a bootstrap sample.
- We first compute the bootstrap KDE:

$$\widehat{p}_h^*(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{X_i^* - x}{h}\right).$$

- Then we compute the bootstrap L_∞ error $W = \|\widehat{p}_h^* - \widehat{p}_h\|_\infty$.
- After repeating the bootstrap procedure B times, we obtain realizations

$$W_1, \cdots, W_B.$$

- Compute the empirical CDF

$$\widehat{F}_n(t) = \frac{1}{B} \sum_{\ell=1}^B I(W_\ell \leq t).$$

- Finally, we use $\widehat{t}_\alpha^* = \widehat{F}_n^{-1}(1-\alpha)$ and construct the confidence band as

$$\widehat{L}_\alpha^*(x) = \widehat{p}_h(x) - \widehat{t}_\alpha^*, \quad \widehat{U}_\alpha^*(x) = \widehat{p}_h(x) + \widehat{t}_\alpha^*.$$

Simple Approach: the Bootstrap - 3

- Does the bootstrap approach work?

Simple Approach: the Bootstrap - 3

- Does the bootstrap approach work?
- It depends.
- The bootstrap works if

$$\|\widehat{p}_h^* - \widehat{p}_h\|_\infty \approx \|\widehat{p}_h - p\|_\infty$$

in the sense that

$$\sup_t |P(\|\widehat{p}_h^* - \widehat{p}_h\|_\infty < t) - P(\|\widehat{p}_h - p\|_\infty < t)| = o(1).$$

Simple Approach: the Bootstrap - 3

- Does the bootstrap approach work?
- It depends.
- The bootstrap works if

$$\|\widehat{p}_h^* - \widehat{p}_h\|_\infty \approx \|\widehat{p}_h - p\|_\infty$$

in the sense that

$$\sup_t \left| P(\|\widehat{p}_h^* - \widehat{p}_h\|_\infty < t) - P(\|\widehat{p}_h - p\|_\infty < t) \right| = o(1).$$

- However, the above bound holds if we *undersmooth* the data (Neumann and Polzehl 1998, Chernozhukov et al. 2014). Namely, we choose the smoothing bandwidth $h = o(n^{-\frac{1}{4+d}})$.

Simple Approach: the Bootstrap - 4

- Why do we need to undersmooth the data?

- Why do we need to undersmooth the data?
- The L_∞ error has a bias-variance tradeoff:

$$\|\widehat{p}_h - p\|_\infty = \underbrace{O(h^2)}_{\text{Bias}} + \underbrace{O_P\left(\sqrt{\frac{\log n}{nh^d}}\right)}_{\text{stochastic error}}.$$

Simple Approach: the Bootstrap - 4

- Why do we need to undersmooth the data?
- The L_∞ error has a bias-variance tradeoff:

$$\|\widehat{p}_h - p\|_\infty = \underbrace{O(h^2)}_{\text{Bias}} + \underbrace{O_P\left(\sqrt{\frac{\log n}{nh^d}}\right)}_{\text{stochastic error}}.$$

- The bootstrap L_∞ error is capable of capturing the errors in the stochastic part. However, it does not capture the bias.

Simple Approach: the Bootstrap - 4

- Why do we need to undersmooth the data?
- The L_∞ error has a bias-variance tradeoff:

$$\|\widehat{p}_h - p\|_\infty = \underbrace{O(h^2)}_{\text{Bias}} + \underbrace{O_P\left(\sqrt{\frac{\log n}{nh^d}}\right)}_{\text{stochastic error}}.$$

- The bootstrap L_∞ error is capable of capturing the errors in the stochastic part. However, it does not capture the bias.
- Undersmooth guarantees that the bias is of a smaller order so we can ignore it.

Problem of Undersmoothing

$$\|\widehat{p}_h - p\|_\infty = \underbrace{O(h^2)}_{\text{Bias}} + \underbrace{O_P\left(\sqrt{\frac{\log n}{nh^d}}\right)}_{\text{stochastic error}}.$$

- Undermoothing has a problem: we do not have the optimal convergence rate.
- The optimal rate occurs when we balance the bias and stochastic error: $h = h_{\text{opt}} \asymp n^{-\frac{1}{d+4}}$ (ignoring the log n factor).

Problem of Undersmoothing

$$\|\widehat{p}_h - p\|_\infty = \underbrace{O(h^2)}_{\text{Bias}} + \underbrace{O_P\left(\sqrt{\frac{\log n}{nh^d}}\right)}_{\text{stochastic error}}.$$

- Undermoothing has a problem: we do not have the optimal convergence rate.
- The optimal rate occurs when we balance the bias and stochastic error: $h = h_{\text{opt}} \asymp n^{-\frac{1}{d+4}}$ (ignoring the $\log n$ factor).
- A remedy to this problem: choose h optimally but **correct** the bias (debiased method).

The Debiased Method - 1

- The idea of the debiased method is based on the fact that a leading term of $O(h^2)$ is

$$\frac{h^2}{2} C_K \cdot \nabla^2 p(x),$$

where C_K is a known constant depending on the kernel function and ∇^2 is the Laplacian operator.

The Debiased Method - 1

- The idea of the debiased method is based on the fact that a leading term of $O(h^2)$ is
$$\frac{h^2}{2} C_K \cdot \nabla^2 p(x),$$
where C_K is a known constant depending on the kernel function and ∇^2 is the Laplacian operator.
- We can estimate $\nabla^2 p$ via applying the Laplacian operator to a KDE $\widehat{p}_h$.
- However, such an estimator is inconsistent when we choose $h_{\mathrm{opt}} \asymp n^{-\frac{1}{d+4}}$ because
$$\nabla^2 \widehat{p}_h(x) - \nabla^2 p(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right).$$

The Debiased Method - 1

- The idea of the debiased method is based on the fact that a leading term of $O(h^2)$ is
$$\frac{h^2}{2} C_K \cdot \nabla^2 p(x),$$
where C_K is a known constant depending on the kernel function and ∇^2 is the Laplacian operator.
- We can estimate $\nabla^2 p$ via applying the Laplacian operator to a KDE $\widehat{p}_h$.
- However, such an estimator is inconsistent when we choose $h_{\text{opt}} \asymp n^{-\frac{1}{d+4}}$ because
$$\nabla^2 \widehat{p}_h(x) - \nabla^2 p(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right).$$
- The choice $h = h_{\text{opt}} \asymp n^{-\frac{1}{d+4}}$ implies
$$\nabla^2 \widehat{p}_h(x) - \nabla^2 p(x) = o(1) + O_P(1).$$

The Debiased Method - 2

- To handle this situation, people suggested using two KDE's, one for estimating the density and the other for estimating the bias.

[1]This idea has been used in Calonico et al. (2015) for a pointwise confidence interval.

The Debiased Method - 2

- To handle this situation, people suggested using two KDE's, one for estimating the density and the other for estimating the bias.
- However, actually we ONLY need one KDE.

[1] This idea has been used in Calonico et al. (2015) for a pointwise confidence interval.

The Debiased Method - 2

- To handle this situation, people suggested using two KDE's, one for estimating the density and the other for estimating the bias.
- However, actually we ONLY need one KDE.
- We propose using the same KDE $\widehat{p}_h(x)$ to 'debias' the estimator[1].

[1] This idea has been used in Calonico et al. (2015) for a pointwise confidence interval.

The Debiased Method - 2

- To handle this situation, people suggested using two KDE's, one for estimating the density and the other for estimating the bias.
- However, actually we ONLY need one KDE.
- We propose using the same KDE $\widehat{p}_h(x)$ to 'debias' the estimator[1].
- Namely, we propose to use

$$\widetilde{p}_h(x) = \widehat{p}_h(x) - \frac{h^2}{2} C_K \cdot \nabla^2 \widehat{p}_h(x)$$

  with $h = h_{\text{opt}} \asymp n^{-\frac{1}{d+4}}$.
- The estimator $\widetilde{p}_h(x)$ is called the debiased estimator.

[1]This idea has been used in Calonico et al. (2015) for a pointwise confidence interval.

The Debiased Method + Bootstrap

- To construct a confidence band, we use the bootstrap again but this time we compute the bootstrap debiased estimator

$$\widetilde{p}_h^*(x) = \widehat{p}_h^*(x) - \frac{h^2}{2} C_K \cdot \nabla^2 \widehat{p}_h^*(x)$$

  and evaluate $\|\widetilde{p}_h^* - \widetilde{p}_h\|_\infty$.

- After repeating the bootstrap procedure many times, we compute the EDF $\widetilde{F}_n$ of the realizations of $\|\widetilde{p}_h^* - \widetilde{p}_h\|_\infty$ and obtain the quantile $\widetilde{t}_\alpha^* = \widetilde{F}_n^{-1}(1-\alpha)$.
- The confidence band is

$$\widetilde{L}_\alpha(x) = \widetilde{p}_h(x) - \widetilde{t}_\alpha^*, \quad \widetilde{U}_\alpha(x) = \widetilde{p}_h(x) + \widetilde{t}_\alpha^*.$$

Theory of the Debiased Method

Theorem (Chen 2017)

Assume p belongs to β-Hölder class with $\beta > 2$ and the kernel function satisfies smoothness conditions. When $h \asymp n^{-\frac{1}{d+4}}$,

$$P\left(\widetilde{L}_\alpha(x) \leq p(x) \leq \widetilde{U}_\alpha(x) \ \forall x \in [0,1]^d\right) = 1 - \alpha + o(1).$$

Namely, the debiased estimator leads to an asymptotic simultaneous confidence band under the choice $h \asymp n^{-\frac{1}{d+4}}$.

Why the Debiased Method Work? - 1

- Why the debiased method work? Didn't we have an inconsistent bias estimator?

Why the Debiased Method Work? - 1

- Why the debiased method work? Didn't we have an inconsistent bias estimator?
- We indeed do not have a consistent bias estimator but this is fine!

Why the Debiased Method Work? - 1

- Why the debiased method work? Didn't we have an inconsistent bias estimator?
- We indeed do not have a consistent bias estimator but this is fine!
- Recall that when $h \asymp n^{-\frac{1}{d+4}}$,

$$\nabla^2 \widehat{p}_h(x) - \nabla^2 p(x) = \underbrace{o(1)}_{\text{bias}} + \underbrace{O_P(1)}_{\text{stochastic variation}}.$$

Why the Debiased Method Work? - 1

- Why the debiased method work? Didn't we have an inconsistent bias estimator?
- We indeed do not have a consistent bias estimator but this is fine!
- Recall that when $h \asymp n^{-\frac{1}{d+4}}$,

$$\nabla^2 \widehat{p}_h(x) - \nabla^2 p(x) = \underbrace{o(1)}_{\text{bias}} + \underbrace{O_P(1)}_{\text{stochastic variation}}.$$

- Thus, our debiased estimator has three errors:

$$\begin{aligned}\widetilde{p}_h(x) - p(x) &= \widehat{p}_h(x) - \frac{h^2}{2} C_K \nabla \widehat{p}_h(x) - p(x) \\ &= \underbrace{\frac{h^2}{2} C_K \nabla^2 p(x) + o(h^2)}_{\text{bias}} + O_P\left(\sqrt{\frac{1}{nh^d}}\right) - \frac{h^2}{2} C_K \nabla^2 \widehat{p}_h(x)\end{aligned}$$

Why the Debiased Method Work? - 2

- The above equation equals ($h \asymp n^{-\frac{1}{d+4}}$)

$$\begin{aligned}
\widetilde{p}_h(x) - p(x) &= \underbrace{\frac{h^2}{2} C_K \nabla p(x) + o(h^2)}_{\text{bias}} + O_P\left(\sqrt{\frac{1}{nh^d}}\right) - \frac{h^2}{2} C_K \nabla \widehat{p}_h(x) \\
&= o(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) + \frac{h^2}{2} C_K \underbrace{\left(\nabla^2 p(x) - \nabla^2 \widehat{p}_h(x)\right)}_{=o(1)+O_P(1)} \\
&= o(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) + o(h^2) + O_P(h^2) \\
&= o(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) + O_P\left(h^2\right).
\end{aligned}$$

- Both the orange and purple terms are stochastic variation.
- Orange: from estimating the density.
- Purple: from estimating the bias.

Why the Debiased Method Work? - 3

- When $h \asymp n^{-\frac{1}{d+4}}$, the error rate

$$\widetilde{p}_h(x) - p(x) = o(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) + O_P\left(h^2\right)$$
$$= O_P(n^{-\frac{2}{d+4}})$$

  is dominated by the stochastic variation.
- As a result, the bootstrap can capture the errors, leading to an asymptotic valid confidence band.

Why the Debiased Method Work? - 4

- Actually, after closely inspecting the debiased estimator, you can find that

$$\begin{aligned}\widetilde{p}_h(x) &= \widehat{p}_h(x) - \frac{h^2}{2} C_K \cdot \nabla^2 \widehat{p}_h(x) \\ &= \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) - \frac{h^2}{2} C_K \cdot \frac{1}{nh^d} \sum_{i=1}^{n} \nabla^2 K\left(\frac{X_i - x}{h}\right) \\ &= \frac{1}{nh^d} \sum_{i=1}^{n} M\left(\frac{X_i - x}{h}\right),\end{aligned}$$

where

$$M(x) = K(x) - \frac{C_K}{2} \cdot \nabla^2 K(x).$$

- Namely, the debiased estimator is a KDE with kernel function $M(x)$!

Why the Debiased Method Work? - 5

- The kernel function

$$M(x) = K(x) - \frac{C_K}{2} \cdot \nabla^2 K(x)$$

is actually a higher order kernel.

Why the Debiased Method Work? - 5

- The kernel function

$$M(x) = K(x) - \frac{C_K}{2} \cdot \nabla^2 K(x)$$

is actually a higher order kernel.

- You can show that if the kernel function $K(x)$ is a γ-th order kernel function, then the corresponding $M(x)$ will be a $(\gamma + 2)$-th order kernel (Calonico et al. 2015, Scott 2015).

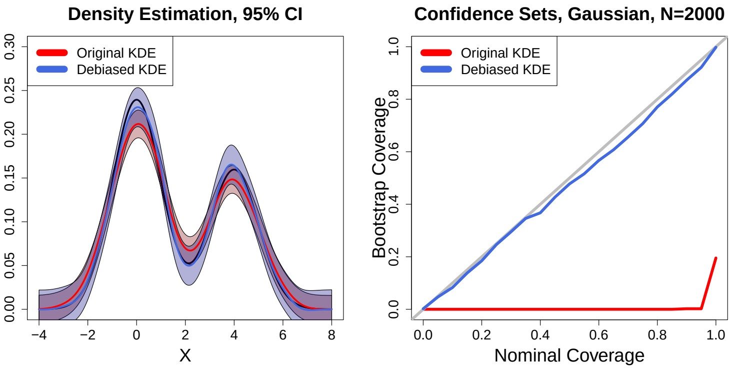
Why the Debiased Method Work? - 5

- The kernel function

$$M(x) = K(x) - \frac{C_K}{2} \cdot \nabla^2 K(x)$$

  is actually a higher order kernel.
- You can show that if the kernel function $K(x)$ is a γ-th order kernel function, then the corresponding $M(x)$ will be a $(\gamma + 2)$-th order kernel (Calonico et al. 2015, Scott 2015).
- Because the debiased estimator $\widetilde{p}_h(x)$ uses a higher order kernel, the bias is moved to the next order, leaving the stochastic variation dominating the error.

Simulation



Density Estimation, 95% CI
Original KDE
Debiased KDE
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−4
−2
0
2
4
6
8
X
Confidence Sets, Gaussian, N=2000
Original KDE
Debiased KDE
Bootstrap Coverage
0.0
0.2
0.4
0.6
0.8
1.0
Nominal Coverage

Conclusion

- We illustrate a bootstrap approach to construct a simultaneous confidence band via a debiased KDE.
- This approach allows us to choose the smoothing bandwidth optimally and still leads to an asymptotic confidence band.
- A similar idea can also be applied to regression problem and local polynomial estimator.
- More details can be found in
    - Chen, Yen-Chi. "Nonparametric Inference via Bootstrapping the Debiased Estimator." arXiv preprint arXiv:1702.07027 (2017).

Thank you!

References

1. Y.-C. Chen. Nonparametric Inference via Bootstrapping the Debiased Estimator. arXiv preprint arXiv:1702.07027, 2017.
2. Y.-C. Chen. A Tutorial on Kernel Density Estimation and Recent Advances. arXiv preprint arXiv:1704.03924, 2017.
3. S. Calonico, M. D. Cattaneo, and M. H. Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. arXiv preprint arXiv:1508.02973, 2015.
4. V. Chernozhukov, D. Chetverikov, and K. Kato. Anti-concentration and honest, adaptive confidence bands. The Annals of Statistics, 42(5):1787–1818, 2014.
5. M. H. Neumann and J. Polzehl. Simultaneous bootstrap confidence bands in nonparametric regression. Journal of Nonparametric Statistics, 9(4):307–333, 1998.
6. D. W. Scott. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, 2015.