Yen-Chi Chen
A Note on Debiased Kernel Density Estimator
November 12, 2016

I find the following paper (CCF) states a very useful result about nonparametric inference:

- Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. "On the effect of bias estimation on coverage accuracy in nonparametric inference." arXiv preprint arXiv:1508.02973 (2015).

They propose to use a debiased kernel in the kernel density estimator (KDE) such that the resulting KDE has a higher order bias $O(h^3)$ and the usual variance $O(\sqrt{\frac{1}{nh^d}})$. Thus, we can perform valid inference directly for p under the optimal smoothing bandwidth $h \sim n^{-\frac{1}{d+4}}$.

A good news is that they only propose a pointwise inference and they estimate the variance using the sample variance. We can generalize all these ideas to a uniform sense and use Chernozhukov's work to perform a valid bootstrap inference. Here is a succint description about their methods.

1 Debiased KDE

Let $X_1, \cdots, X_n$ be IID from an unknown density function p with a compact support $\mathbb{K} \in \mathbb{R}$. For simplicity, we consider $d = 1$ case. We define the naive KDE as

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

where $K(x)$ is the kernel function and $h > 0$ is the smoothing bandwidth.

Now we define the Hessian estimator using another smoothing bandwidth $b > 0$ as

$$\widehat{p}_b^{(2)}(x) = \frac{1}{nb^3}\sum_{i=1}^n K^{(2)}\left(\frac{x - X_i}{b}\right),$$

where $K^{(2)}(x) = \frac{d^2}{dx^2}K(x)$ is the second derivative of the kernel function $K(x)$.

Let $\tau = \frac{h}{b}$. The *debiased KDE* is

$$\begin{aligned}
\widehat{p}_\tau(x) &= \widehat{p}_h(x) - c_K \cdot h^2 \cdot \widehat{p}_b^{(2)}(x) \\
&= \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) - c_K \cdot h^2 \cdot \frac{1}{nb^3}\sum_{i=1}^n K^{(2)}\left(\frac{x-X_i}{b}\right) \\
&= \frac{1}{nh}\sum_{i=1}^n M_\tau\left(\frac{x-X_i}{h}\right),
\end{aligned} \tag{1}$$

where

$$M_\tau(x) = K(x) - c_K \cdot \tau^3 \cdot K^{(2)}(\tau \cdot x), \tag{2}$$

where $c_K = \int x^2 K(x)dx$. The function $M_\tau(x)$ can be viewed as a new kernel function, which we called it *debiased kernel function*. Note that the second quantity $c_K \cdot h^2 \cdot \widehat{p}_b^{(2)}(x)$ is an estimate for the asymptotic bias in the KDE so it is to reduce the bias in the naive KDE.

What is important here is that *we allow* $\tau \in (0, \infty)$ and we still have a valid confidence set.

2 Analysis for the Bias

We first show that under usual assumption, the debiased KDE $\widehat{p}_\tau(x)$ has a bias at the order of $O(h^3)$. To see this, note that

$$\begin{aligned}
\mathbb{E}\left(\widehat{p}_\tau(x)\right) &= \mathbb{E}\left(\widehat{p}_h(x)\right) - c_K \cdot h^2 \cdot \mathbb{E}\left(\widehat{p}_b^{(2)}(x)\right) \\
&= p(x) + c_K \cdot h^2 \cdot p^{(2)}(x) + O(h^3) - c_K \cdot h^2 \cdot p^{(2)}(x)(1 + O(b^2)) \\
&= p(x) + O(h^3) + O(h^2 \cdot b^2).
\end{aligned}$$

Thus, the debiased KDE has the bias at the order of $O(h^3)$.

3 Analysis for the Variance

The actual power of the debiased KDE is in its variance:

$$\begin{aligned}
\mathsf{Var}\left(\widehat{p}_\tau(x)\right) &= \mathsf{Var}\left(\widehat{p}_h(x)\right) + \mathsf{Cov}\left(\widehat{p}_h(x), c_K \cdot h^2 \cdot \widehat{p}_b^{(2)}(x)\right) + \mathsf{Var}\left(c_K \cdot h^2 \cdot \widehat{p}_b^{(2)}(x)\right) \\
&= O\left(\frac{1}{nh}\right) + O\left(\frac{1}{\sqrt{nh}} \cdot \frac{h^2}{\sqrt{nb^3}}\right) + O\left(\frac{h^2}{nb^3}\right) \\
&= O\left(\frac{1}{nh}\right) + O\left(\frac{1}{nh} \cdot \tau^{\frac{3}{2}}\right) + O\left(\frac{1}{nh} \cdot \tau^3\right).
\end{aligned}$$

Thus, as long as $\tau < \infty$, we have the asymptotic variance

$$nh\mathsf{Var}\left(\widehat{p}_\tau(x)\right) = O(1) + O\left(\tau^{\frac{3}{2}}\right) + O\left(\tau^3\right).$$

Actually, if we derive the variance in details, we have

$$\sigma_\tau^2(x) = nh\mathsf{Var}\left(\widehat{p}_\tau(x)\right) = \sigma_1^2(x) + \tau^{\frac{3}{2}} \cdot \sigma_{12}^2(x) + \tau^3\sigma_2^2(x) + o(1),$$

where

$$\begin{aligned}
\sigma_1^2(x) &= \frac{1}{h}\mathsf{Var}\left(K\left(\frac{x-X_i}{h}\right)\right) \\
\sigma_{12}^2(x) &= \frac{c_K}{\sqrt{hb}}\mathsf{Cov}\left(K\left(\frac{x-X_i}{h}\right), K^{(2)}\left(\frac{x-X_i}{b}\right)\right) \\
\sigma_2^2(x) &= \frac{1}{b}\mathsf{Var}\left(K^{(2)}\left(\frac{x-X_i}{b}\right)\right).
\end{aligned}$$

Thus, when $\tau < \infty$, the variance is at rate $O(\frac{1}{nh})$, which is the same as the naive KDE! This implies that if we choose $h \sim b \sim h^{-1/5}$, the debiased KDE has bias $O(h^3) = O(n^{-3/5})$ and stochastic variation $O_P\left(\sqrt{\frac{1}{nh}}\right) = O_P(n^{-2/5})$, so the stochastic part dominates the bias, meaning that as long as we can estimate the variance well, we have a valid confidence interval.

So what happens here? An observation is that when $b \sim h^{-1/5}$, the Hessian estimator $\widehat{p}_b^{(2)}(x)$ is not consistent for $p^{(2)}(x)$ because the variance does not converges. However, the bias does converge. Thus, asymptotically $\widehat{p}_b^{(2)}(x)$ is centered around $p^{(2)}(x)$ with a non-vanishing limiting distribution (Gaussian).

Now because we multiply the second derivative estimator (debiased part) by h^2, the asymptotic distribution of $\widehat{p}_b^{(2)}(x) - p^{(2)}(x)$ converges at rate $O(h^2)$. Therefore, the debiased KDE is still consistent even if we do not consistently estimate the second derivative. The non-vanishing of the bias in the second derivative estimator contributes to the asymptotic variance of the debiased KDE.

4 Inference using the Debiased KDE

In the CCF paper, the propose to use a sample variance estimate $\widehat{\sigma}_\tau^2(x)$ for the asymptotic variance $\sigma_\tau^2(x)$, which has the property

$$\frac{\widehat{\sigma}_\tau^2(x)}{\sigma_\tau^2(x)} \xrightarrow{P} 1$$

and further leads to a pointwise confidence set.

We can improve their result using the bootstrap and L_∞ metric. Recall from (1),

$$\begin{aligned}\widehat{p}_\tau(x) &= \frac{1}{nh}\sum_{i=1}^n M_\tau\left(\frac{x-X_i}{h}\right)\\ &= \frac{1}{h}\int M_\tau\left(\frac{x-y}{h}\right)d\widehat{\mathbb{P}}_n(y).\end{aligned}$$

The bias analysis implies

$$\mathbb{E}\left(\widehat{p}_\tau(x)\right) = \frac{1}{h}\int M_\tau\left(\frac{x-y}{h}\right)d\mathbb{P}(y) = p(x) + O(h^3).$$

Using the notation of empirical process and define $f_x(y) = \frac{1}{\sqrt{h}}M_\tau\left(\frac{x-y}{h}\right)$,

$$\widehat{p}_\tau(x) - p(x) = \frac{1}{\sqrt{h}}\left(\widehat{\mathbb{P}}_n(f_x) - \mathbb{P}(f_x)\right) + O(h^3).$$

Thus,

$$\sqrt{nh}\left(\widehat{p}_\tau(x) - p(x)\right) = \mathbb{G}_n(f_x) + O(\sqrt{nh^7}) = \mathbb{G}_n(f_x) + o(1).$$

Now define the function class

$$\mathcal{F}_\tau = \{f_x(y) : x \in \mathbb{K}\}.$$

Note that as long as we assume VC-type class for the kernel function K and its second derivative $K^{(2)}$, F_τ will also be a VC-type class. Thus, by Chernozhukov's approach, the L_∞-norm $\sup_{x\in\mathbb{K}}\cdot\sqrt{nh}\,\|\widehat{p}_\tau(x) - p(x)\|$ converges to the supremum of a Gaussian process. Namely, there exists a Gaussian process $\mathbb{B}$ such that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{nh}\,\|\widehat{p}_\tau - p\|_\infty \leq t\right) - \mathbb{P}\left(\sup_{f\in\mathcal{F}_M}\|\mathbb{B}(f)\| \leq t\right)\right| = O\left(\left(\frac{\log^7 n}{nh}\right)^{1/6}\right).$$

Moreover, we can use the bootstrap to derive the uniform confidence set for $p(x)$.

Note that although I derived all the above results using $d = 1$, it is easy to generalize it to multivariate case. The only difference is that the function $M_\tau(x)$ will be

$$M_\tau(x) = K(x) - c_K \cdot h^2 \cdot \nabla^2 K(\tau \cdot x).$$

Based on the debiased KDE, most of our methods, including inferences for level sets, ridges, cluster trees, persistent diagrams,...etc can all be improved. We no longer have to focus on a smoothed surrogate or use undersmoothing to handle the bias.