Kernel density estimation

- Suppose that $X_1, \ldots, X_n$ are IID data with Lebesgue density f. The kernel density estimator of f using kernel function k and bandwidth h is given by
$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right). \tag{1}$$

- Kernel function. A kernel function k usually satisfies the usual constraints:

  (a) $k \geq 0$.

  (b) $\int_{-\infty}^{\infty} k(s)ds = 1$.

  (c) $\int_{-\infty}^{\infty} sk(s)ds = 0$.

  (d) $\int_{-\infty}^{\infty} s^2k(s)ds < \infty$.

- Mean and variance of $\hat{f}(x_0)$. Suppose that $h \to 0$ as $n \to \infty$, $\int_{-\infty}^{\infty} k^2(s)ds < \infty$, and f'' is continuous at x_0. Suppose that $f > 0$ on (a, b) and $f = 0$ outside (a, b). Then it can be shown that
$$\begin{aligned} E(\hat{f}(x_0)) &= E\left((nh)^{-1}\sum_{i=1}^{n} k((x_0 - X_i)/h)\right) \\ &= f(x_0)\int_{(x_0-b)/h}^{(x_0-a)/h} k(u)du - hf'(x_0)\int_{(x_0-b)/h}^{(x_0-a)/h} uk(u)du \\ &\quad + \frac{f''(x_0)h^2}{2}\int_{(x_0-b)/h}^{(x_0-a)/h} u^2k(u)du + o(h^2) \end{aligned} \tag{2}$$
and
$$\begin{aligned} Var(\hat{f}(x_0)) &= \frac{1}{nh^2}\left[E\left(k^2((x_0 - X_1)/h)\right)\right] - \frac{1}{n}E\left(h^{-1}k((x_0 - X_1)/h)\right) \\ &= \frac{1}{nh}\left[f(x_0)\int_{(x_0-b)/h}^{(x_0-a)/h} k^2(u)du + o(1)\right] + O\left(\frac{1}{n}\right). \end{aligned}$$
When $a < x_0 < b$, $(x_0 - a)/h \to \infty$ and $(x_0 - b)/h \to -\infty$,
$$E(\hat{f}(x_0)) \to f(x_0)$$
as $h \to 0$. However, if $x_0 \approx a$ or $x_0 \approx b$, the bias of $\hat{f}(x_0)$ can be very large.

- Example 1. Compute the kernel density estimator based on 5000 observations from $Uniform(0, 1)$.

```
fhat0 <- function(x, x0, h, k){  return(mean( k((x0-x)/h)/h )) }
get_fhat <- function(x,h, k=dnorm){
 f <- function(x0){ return( fhat0(x,x0,h, k) ) }
 f1 <- Vectorize(f); return(f1)
}
```

```
set.seed(1)
x <- runif(5000)
fhat <- get_fhat(x, 0.07)
curve(fhat,0,1)
curve(dunif,0,1,add=T,col=2)
```

- Boundary bias correction. Suppose that $f > 0$ on (a, b) and $f = 0$ outside (a, b). Consider replacing the kernel k in (2) by k_1, where

$$k_1(u) = Ak(u) + Buk(u) \tag{3}$$

for $-\infty < u < \infty$ and A and B are two constants such that

$$\int_{(x_0-b)/h}^{(x_0-a)/h} k_1(u)du = 1 \tag{4}$$

and

$$\int_{(x_0-b)/h}^{(x_0-a)/h} uk_1(u)du = 0. \tag{5}$$

For $i = 0$, 1, 2, let

$$g_i(s,t) = \int_s^t u^i k(u)du$$

$$a_i(x_0) = g_i\left(\frac{x_0 - b}{h}, \frac{x_0 - a}{h}\right).$$

Then (4) and (5) can be written as

$$\begin{cases} a_0(x_0)A + a_1(x_0)B = 1 \\ a_1(x_0)A + a_2(x_0)B = 0 \end{cases}$$

Solving for A, B and plug the results in (3), then we have

$$k_1(u) = \frac{a_2(x_0)k(u) - a_1(x_0)uk(u)}{a_0(x_0)a_2(x_0) - a_1^2(x_0)}$$

for $u \in (-\infty, \infty)$. We can then estimate $f(x_0)$ using

$$\hat{f}_L(x_0) = \frac{1}{nh}\sum_{i=1}^{n} k_1\left(\frac{x_0 - X_i}{h}\right). \tag{6}$$

- Let ϕ be the $N(0, 1)$ PDF (`dnorm`) and Φ be the $N(0, 1)$ CDF (`pnorm`). Then for $k = \phi$,

$$g_0(s,t) = \Phi(t) - \Phi(s),$$

$$g_1(s,t) = -\phi(t) + \phi(s),$$

and

$$g_2(s,t) = -t\phi(t) + s\phi(s) + \Phi(t) - \Phi(s)$$

- The idea for the above correction can be found in a PDF file by Tine Buch-Kromann. Title: Simple boundary correction for kernel density estimation. Link:

`https://www.semanticscholar.org/paper/Simple-boundary-correction-for-kernel-density-Buch-Kromann/b2b73f1a526a5d8064cecc61473c20bec6644942`

- Bandwidth selection. We use leave-one-out cross-validation to choose h for a given kernel k. Two types of cross-validation are considered:
  - Least square cross-validation;
  - Likelihood cross-validation.
- Leave-one-out least square cross-validation. Let $\hat{f}_{-i,h}$ be the kernel estimator for f with bandwidth h based on $X_1, \ldots, X_n$ with X_i removed and $\hat{f}_h$ be the kernel estimator for f with bandwidth h based on $X_1, \ldots, X_n$. Let

$$LSCV(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i,h}(X_i).$$

  Leave-one-out least square cross-validation: choose the bandwidth h so that $LSCV(h)$ is minimized.
- Leave-one-out likelihood cross-validation. Let $\hat{f}_{-i,h}$ be the kernel estimator for f with bandwidth h based on $X_1, \ldots, X_n$ with X_i removed. Let

$$LikCV(h) = \sum_{i=1}^{n} \log \hat{f}_{-i,h}(X_i).$$

  Leave-one-out likelihood cross-validation: choose the bandwidth h so that $LikCV(h)$ is maximized.
  - Suppose that f and g are positive probability density functions. Then

$$\int \log\left(\frac{f(x)}{g(x)}\right) f(x)dx \geq 0,$$

    and equality holds when $f = g$ almost everywhere.
- More information about least square cross-validation and likelihood cross-validation can be found in [1] and [2].
- Exercise 1.
  (a) Write an R function that computes the kernel density estimator of f in (6) with given data, kernel and bandwidth.
  (b) Suppose that $n = 5000$. Compute the IMSE of the kernel estimator in (6) based on simulated data $X_1, \ldots, X_n$ from $Uniform(0, 1)$. The kernel function k used for computing k_1 in (6) is the $N(0, 1)$ PDF and the bandwidth $h = 0.08$. The IMSE is computed based on 200 simulation runs.
  (c) Compute the IMSE of the kernel estimator in (1) based on simulated data $X_1, \ldots, X_n$ in Part (b). The kernel function k used in (1) is the $N(0, 1)$ PDF and the bandwidth $h = 0.08$. Compare the IMSE with the IMSE in Part (b).

Exercise 2.

(a) Suppose that $n = 100$. Compute the IMSE of the kernel estimator in (1) based on simulated data $X_1, \ldots, X_n$ from $N(0,1)$. The kernel function k is the $N(0,1)$ PDF and the bandwidth h is selected by leave-one-out least square cross-validation. The IMSE is computed based on 200 simulation runs. The range of h is $[1/n, 0.5]$.

(b) Do Part (a) again with least square cross-validation replaced by likelihood cross-validation (using the same data). Compare the IMSE with the IMSE from Part (a).

Exercise 3. Suppose that $n = 5000$. Generate IID data $X_1, \ldots, X_n$ from the exponential distribution with mean 1.

(a) Estimate the density of X_i using the $\hat{f}$ in (1) with k being the $N(0,1)$ PDF and $h = 0.08$. Approximate the bias $E(\hat{f}(0)) - f(0)$ based on 200 simulation runs.

(b) Propose a kernel density estimator $\hat{f}$ so that the boundary bias at 0 can be corrected. Use $h = 0.08$. Compute the IMSE based on 200 runs.

- Multivariate kernel density estimation. Suppose that $X_1, \ldots, X_n$ are IID data with Lebesgue density f on R^d. The kernel density estimator of f using kernel function k and bandwidth h is given by

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right).$$

- Kernel function on R^d. A kernel function k on R^d usually satisfies the usual constraints:

  (a) $k \geq 0$.
  (b) $\int k(s)ds = 1$.
  (c) $\int s_j k(s_1, \ldots, s_d) d(s_1, \ldots, s_d) = 0$ for $j = 1, \ldots, d$.
  (d) $\int \|s\|^2 k(s)ds < \infty$, where $\|(s_1, \ldots, s_d)\|^2 = \sum_{j=1}^{d} s_j^2$.

- Example of a kernel function on R^d. Let $k_1, \ldots, k_d$ be d univariate kernel functions. Define

$$k(x_1, \ldots, x_d) = k_1(x_1) \cdots k_d(x_d) \tag{7}$$

  for $(x_1, \ldots, x_d) \in R^d$. Then k is a kernel function on R^d. A kernel k of the form in (7) is called a product kernel.

References

[1] J. S. Horne and E. O. Garton, Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in kernel home-range analysis, The Journal of Wildlife Management, 70 (2006), pp. 641–648.

[2] B. W. Silverman, Density estimation for statistics and data analysis, Chapman & Hall Ltd, London; New York, 1986.