Nonparametric Estimation: Final Report

Jaehyeong Ahn

December 2020

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be i.i.d copies of the random vector (X, Y) with the joint density $f_{X,Y}(x, y)$. Let $F_{Y|X}(y|x)$ be the conditional distribution function of Y given $X = x$.

1. Construct an estimator of $F_{Y|X}(y|x)$ by using kernel density estimators. Note that $F_{Y|X}(y|x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,t)dt}{f_X(x)}$ where $f_X(x)$ is the marginal density of X.

$$F_{Y|X}(y|x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,t)dt}{f_X(x)} \tag{1}$$

$$= \frac{\int_{-\infty}^{y} n^{-1} \sum_{i=1}^{n} h_x^{-1} K\left(\frac{x-X_i}{h_x}\right) h_y^{-1} K\left(\frac{t-Y_i}{h_y}\right) dt}{n^{-1} \sum_{i=1}^{n} h_x^{-1} K\left(\frac{x-X_i}{h_x}\right)} \tag{2}$$

$$= \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_x}\right) \int_{-\infty}^{y} h_y^{-1} K\left(\frac{t-Y_i}{h_y}\right) dt}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_x}\right)} \tag{3}$$

$$= \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_x}\right) L\left(\frac{y-Y_i}{h_y}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_x}\right)} \tag{4}$$

(1) $\rightarrow$ (2): Estimate $f_{X,Y}(\cdot, \cdot)$ by multivariate kernel density estimation with product kernel and estimate $f_X(\cdot)$ by univariate kernel density estimation (let h_x be a bandwidth of X and h_y be a bandwidth of Y)

(3) $\rightarrow$ (4): Let $L(\frac{y-Y_i}{h_y}) = \int_{-\infty}^{y} h_y^{-1} K(\frac{t-Y_i}{h_y})dt$

2. Construct an estimator of $F_{Y|X}(y|x)$ by using the idea of the local constant estimator. Note that $F_{Y|X}(y|x) = E(I(Y \leq y)|X = x)$.

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) \tag{5}$$
$$= E(I(Y \leq y)|X = x) \tag{6}$$
$$= \frac{n^{-1}\sum_{i=1}^{n} h_x^{-1} K\left(\frac{x-X_i}{h_x}\right) I(Y_i \leq y)}{n^{-1}\sum_{i=1}^{n} h_x^{-1} K\left(\frac{x-X_i}{h_x}\right)} \tag{7}$$
$$= \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_x}\right) I(Y_i \leq y)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_x}\right)} \tag{8}$$

(6) $\rightarrow$ (7): Estimate the conditional expectation by using the form of Nadarya-Watson estimator

3. Design a simulation study of comparing two estimators with following models and report results.

a. $(X, Y) \sim N_2(\mu, \Sigma)$ where $\mu = (0, 0)^T$ and $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

b. $Y = sin(\pi X) + exp(-X^2)\epsilon$ where $\epsilon \sim N(0, 1), X \sim N(0, 1)$, and X and ϵ are independent.

Simulation Design
Let's call the estimator (4) as smoothed conditional distribution estimator ($\hat{F}(y|x)$) and the estimator (8) as unsmoothed conditional distirbution estimator ($\tilde{F}(y|x)$).

Simulation is designed as follow:

- Sample size: 200
- Grid point: seq(min(X),max(X),by=0.01), seq(min(Y),max(Y),by=0.01)
- $K(\cdot)$: A density function of standard normal distribution
- Bandwidth is selected by using a cross-validatoin method for conditional distribution estimation which was introduced in Li et al. [2013]
  - Bandwidth for Unsmoothed estimator is obtained by minimizing follow $USCV(y)$ function:

$$USCV(y, h_x) = \frac{1}{n}\sum_{i=1}^{n}\left(I(Y_i \leq y) - \hat{F}_{-i}(y|X_i)\right)^2 \tag{9}$$
$$USCV(y) = \int USCV(y, h_x)dy \tag{10}$$

– Bandwidth for Smoothed estimator is obtained by minimizing follow $SCV(y)$ function

$$SCV(y, h_x, h_y) = \frac{1}{n}\sum_{i=1}^{n}\Big(I(Y_i \leq y) - \tilde{F}_{-i}(y|X_i)\Big)^2 \tag{11}$$

$$SCV(y) = \int SCV(y, h_x, h_y)dy \tag{12}$$

– Where, $h_x = h_y = seq(0.01, 0.5, 0.01)$

Simulation Result

(a)

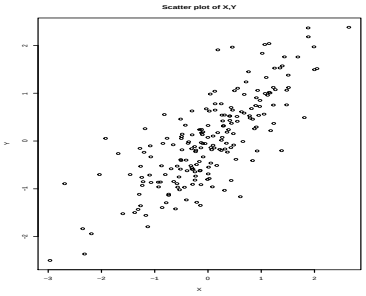
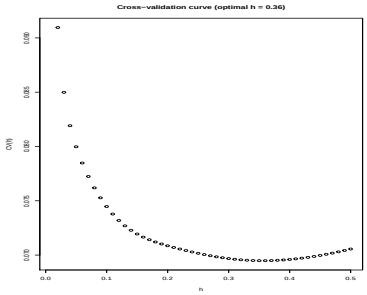
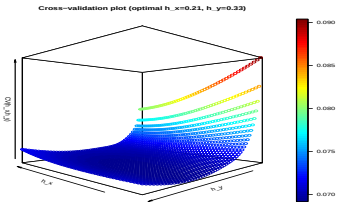


Figure 1: Scatter plot of X, Y

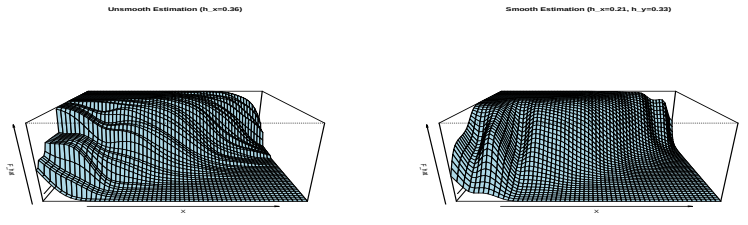


(a) Cross-validation curve of unsmoothed estimator



(b) Cross-validation curve of smoothed estimator

Figure 2: Cross-validation curves of unsmoothed and smoothed estimators. In (a), the curve has minimum value when $h_x = 0.36$. In (b), the function has minimum value when $h_x = 0.21, h_y = 0.33$



(a) Estimated conditional distribution function by unsmoothed estimator with $h_x = 0.36$

(b) Estimated conditional distribution function by smoothed estimator with $h_x = 0.21, h_y = 0.33$

Figure 3: Estimated conditional distribution function by unsmoothed and smoothed estimator with selected bandwidth

In Figure 3, we can observe that unsmoothed conditional distribution estimator is more discrete than smoothed conditional distribution estimator or smoothed conditional distribution estimator is smoother than unsmoothed conditional distribution estimator.

(b)

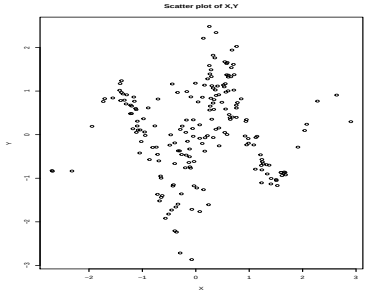
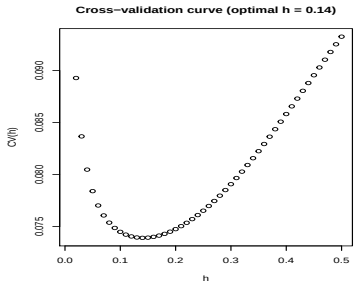
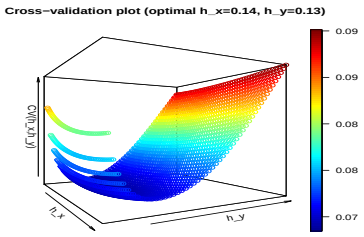


Figure 4: Scatter plot of X, Y

In Figure 6, we can observe that unsmoothed conditional distribution estimator is more discrete than smoothed conditional distribution estimator or smoothed conditional distribution estimator is smoother than unsmoothed conditional distribution estimator.

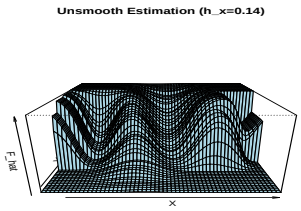


(a) Cross-validation curve of unsmoothed estimator

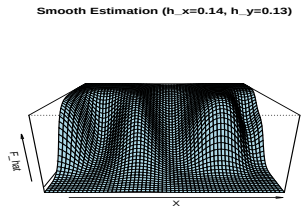


(b) Cross-validation curve of smoothed estimator

Figure 5: Cross-validation curves of unsmoothed and smoothed estimators. In (a), the curve has minimum value when $h_x = 0.14$. In (b), the function has minimum value when $h_x = 0.14, h_y = 0.13$



(a) Estimated conditional distribution function by unsmoothed estimator with $h_x = 0.14$



(b) Estimated conditional distribution function by smoothed estimator with $h_x = 0.14, h_y = 0.13$

Figure 6: Estimated conditional distribution function by unsmoothed and smoothed estimator with selected bandwidth

References

Qi Li, Juan Lin, and Jeffrey S Racine. Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*, 31(1):57–65, 2013.