香 港 大 學

THE UNIVERSITY OF HONG KONG

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

SEMI-SUPERVISED LEARNING WITH NW-ESTIMATOR

AUTHOR: WANG JUNSHI
PROJECT SUPERVISOR: PROF. STEPHEN LEE

Declaration

I confirm that I have read and understood the University's Academic Integrity Policy. I confirm that I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study.
I confirm that I have not copied material from another source nor committed plagiarism nor fabricated, falsified or embellished data when completing the attached piece of work. I confirm that I have not copied material from another source, nor colluded with any other student in the preparation and production of this work.

Signed: Date: November 15, 2021

NOVEMBER 15, 2021

Contents

List of Figures

List of Tables

1 Background

1.1 Introduction

In this report, the author will discuss how semi-supervised learning (SSL) can be used in kernel regression, particularly Nadaraya-Watson estimator(NW-Estimator). SSL here refers to the statistical approach to leverage both labeled and unlabeled to generate better results in terms of mean square error and other criteria. The method of SSL is powerful in that it not only focuses on predicting the unobserved points, but also lays emphasis on explore unspecified patterns Chapelle et al. (2009). This helps boost the performance of estimators when labeled data are sparse and expensive to collect while unlabeled data can be relatively easily obtained. Under the context of NW-Estimator, the classical estimator and the self-supervised estimator using labeled and unlabeled data will be merged into a hybrid estimator. The asymptotic distribution, mean square error(MSE) and confidence interval(CI) of the hybrid estimator will be calculated to demonstrate the effectiveness of SSL. Finally, simulations will be carried out to visualize the performance of each estimator.

1.2 Literature Review

Many researchers have contributed to the development of semi-supervised learning and shed light on ways to take full advantage of limited labelled data with the assistance of unlabelled data.

Pseudo-Labeling Lee et al. (2013) proposed a convenient way to train neural

network in a semi-supervised fashion. Their model is trained simultaneously with both labeled and unlabeled data. The labeled data are utilized to predict the class of the unlabeled data. These predictions are treated as if they were observed values, named as Pseudo-Labels. In principle, almost all neural network models and training methods are merged into this model. The proposed method generates excellent performance in further experiments on the MNIST dataset.

MixMatch: Proposed by a research team from Google Berthelot et al. (2019), MixMatch is a model that incorporate several mainstream SSL techniques such as Entropy Mininization, Pseudo-labelling and consistency regularisation. With a given sample of labeled and unlabeled data X and U, it first applies data augmentations to both X and U. Next the augmented X will be used to train a classifier and make predictions for augmented U. Then the predictions are averaged across these augmentations and sharpened to give label guesses. Then a convex combination of both labeled data and unlabeled data with guessed label through a special shuffling process called Mixup Zhang et al. (2017) will be constructed resulting in new datasets X′ and U′. The final classifier is trained through minimising the loss function L = LX + LU , which is a linear combination of losses from X ′ and U′ respectively.

2 | Methodology and Calculations

2.1 Supervised Estimator

Definition of NW-Estimator is generally given by the following equations and it can be divided into two parts, namely $\hat{\alpha}(x)$ and $\hat{p}(x)$.

$$NW_{Labeled} = \hat{m}(x) = \frac{\hat{\alpha}(x)}{\hat{p}(x)} \tag{1}$$

$$\hat{\alpha}(x) = \frac{1}{nh_n}\Sigma y_i \cdot K(\frac{x - x_i}{h_n}) \tag{2}$$

$$\hat{p}(x) = \frac{1}{nh_n}\Sigma K(\frac{x - x_i}{h_n}) \tag{3}$$

$$y_i = m(x_i) + \epsilon_i \tag{4}$$

The following assumptions are generally adopted (Wand and Jones (1995)):

(i) m$''$(x) is continuous for all x $\in$ [0, 1].

(ii) The kernel k(x) is symmetric about x = 0 and supported on $[-1, 1]$.

(iii) $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$.

(iv) The given point x = x0 must satisfy $h_n < x_0 < 1 - h_n$ for all $n \geq n_0$ where n_0 is a fixed number.

2.1.1 Asymptotic normality of $\hat{\alpha}(x)$ and $\hat{p}(x)$

To demonstrate the asymptotic normality of $\hat{\alpha}(x)$ and $\hat{p}(x)$, one way is to refer to **Lyapunoví Central Limit Theorem for Triangular Arrays**,

Theorem 2.1.1. *if the scalar random variable z_{in} is independently (but not necessarily identically) distributed with variance $Var(z_{in}) \equiv \sigma_{in}^2$ and r-th absolute central moment $E[|z_{in} - E(z_{in})|^r] \equiv \rho_{in} < \infty$ for some r > 2; and if*

$$\frac{(\Sigma_{i=1}^n \rho_{in})^{1/r}}{(\Sigma_{i=1}^n \sigma_{in}^2)^{1/2}} \to 0$$

then standardized $\{z_{in}\}$ will converge to Normal distribution with mean 0 and variance 1.

Based on this theorem, we are able to deduce the condition under which $\hat{\alpha}(x)$ and $\hat{p}(x)$ will be asymptotically normal. The proof for the later one, which is just kernel density estimator, is available in many works Nadaraya (1965), Wand and Jones (1994), Wied and Weißbach (2012).

Theorem 2.1.2. *Let $y_i = m(x_i) + \epsilon_i$ and ϵ_i follows i.i.d normal distribution, then sequences of the form $\frac{1}{nh_n}\Sigma y_i \cdot K(\frac{x-x_i}{h_n})$are asymptotically normal if $nh_n \to \infty$.*

Proof. Here $z_{in} = \frac{1}{h_n} y_i K(\frac{xi-x}{h_n})$, $yi = m(x_i) + \epsilon_i$. We already know the result that

$$\sigma_{in}^2 = Var(z_{in}) = nVar(hat\alpha(x)) = O(h^{-1})$$

For all r larger than or equal to 3, we have the following

$$\begin{aligned}
\rho_{in} &= E[|z_{in} - E(z_{in})|^r] \leq E[(|z_{in}| + |E(z_{in})|)^r] \\
&= \sum_{k=0}^{r} C_k^r E(|z_{in}|^k |E(z_{in})|^{r-k}) \leq \sum_{k=0}^{r} C_k^r E(|z_{in}|^r)^{\frac{k}{r}} E(|z_{in}|^r)^{\frac{r-k}{r}} \\
&= 2^r E(|z_{in}|^r)
\end{aligned}$$

$$\rho_{in}^{\frac{1}{r}} \leq 2E(|\frac{1}{nh_n}y_iK(\frac{x-x_i}{h_n})|^r)^{\frac{1}{r}} = 2\frac{1}{h_n}E(|(m(y)+\epsilon)K(\frac{x-y}{h_n})|^r)^{\frac{1}{r}}$$
$$= 2\frac{1}{h_n}E(\Sigma C_q^r|m(y)^q\epsilon^{r-q}K(\frac{x-y}{h_n})^r|)^{\frac{1}{r}}$$

$$E(|m(y)^q\epsilon^{r-q}K(\frac{x-y}{h_n})^r|) \leq E(\epsilon^{r-q})E(|m(y)^q\epsilon^{r-q}K(\frac{x-y}{h_n})^r|)$$
$$= C_\epsilon[\int_{C1} m(y)^q\epsilon^{r-q}K(\frac{x-y}{h_n})^r p(y)dy - \int_{C2} m(y)^q\epsilon^{r-q}K(\frac{x-y}{h_n})^r p(y)dy]$$
$$= O(h_n)$$

Therefore, $\rho_{in} \leq O(h_n^{-r+1})$ and if $nh_n \to \infty$, then $\frac{(\Sigma_{i=1}^n \rho_{in})^{1/r}}{(\Sigma_{i=1}^n \sigma_{in}^2)^2} \to 0$, since

$$\frac{(\Sigma_{i=1}^n \rho_{in})^{1/r}}{(\Sigma_{i=1}^n \sigma_{in}^2)^2} \leq \frac{C_1 n^{1/r} h_n^{-1+1/r}}{C_2 n^{1/2} h_n^{-1/2}}$$
$$= Cn^{-1/2+1/r}h_n^{-1/2+1/r}$$
$$= C(nh_n)^{-1/2+1/r} \to 0$$

□

2.1.2 Asymptotic Distribution of NW-Estimator

After proving normality of $\hat{\alpha}(x)$ and $\hat{p}(x)$, we will able to discuss the distribution of $\hat{m}(x)$.

$$A = (nh_n)^{1/2}[\hat{\alpha}(x) - E[\hat{\alpha}(x)]] \tag{5}$$

$$B = (nh_n)^{1/2}[\hat{p}(x) - E[\hat{p}(x)]] \tag{6}$$

Thus we can write NW-Estimator in the following way,

$$\begin{aligned}\hat{m}(x) - m(x) &= \frac{\hat{\alpha}(x)}{\hat{p}(x)} - m(x) \\ &= \frac{(nh_n)^{-1/2})A + E[\hat{\alpha}(x)]}{(nh_n^{-1/2})B + E[\hat{p}(x)]} - m(x) \\ &= \frac{(nh_n^{-1/2})(A - Bm(x)) + E[\hat{\alpha}(x) - m(x)\hat{p}(x)]}{(nh_n^{-1/2})B + E[\hat{p}(x)]}\end{aligned}$$

when $nh_n^2 \to \infty, A \xrightarrow{d} N(0, \sigma_a^2), B \xrightarrow{d} N(0, \sigma_b^2)$. Therefore $(nh_n)^{-1/2}A$ and $(nh_n)^{-1/2}B \xrightarrow{p} 0$ and we can apply Taylor expansion to the equation.

$$\nabla G(E(\hat{\alpha}(x)), E(\hat{p}(x)) = \begin{bmatrix} \frac{1}{E(\hat{p}(x))} \\ -\frac{E(\hat{\alpha}(x))}{E(\hat{p}(x))^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{p(x)} + O(h_n^2) \\ -\frac{m(x)}{p(x)} + O(h_n^2) \end{bmatrix}$$

$$\begin{aligned}\hat{m}(x) - m(x) &= G((nh_n)^{-1/2}A + E(\hat{\alpha}(x)), (nh_n)^{-1/2}B + E(\hat{p}(x))) - m(x) \\ &= E + (nh_n)^{-1/2}FA + (nh_n)^{-1/2}GB + (nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} + O_p((nh_n)^{-1})\end{aligned} \tag{7}$$

Where $E = [h_n^2 m'(x)p'(x)\mu_2(k) + \frac{h_n^2}{2}m''(x)p(x)\mu_2(k)]\frac{1}{p(x)} + O(h_n^4)$. F and G are both $O(h_n^2)$.

Combining all information together, $\hat{t}(x)$ as defined below, with leading term $\frac{A - m(x)B}{p(x)}$ may have an asymptotically normal distribution if $h_n^2 \cdot (nh_n)^{1/2}$ is bounded and (A,B) follows a bivariate normal distribution asymptotically.

$$\hat{t}(x) = (nh_n)^{1/2}[\hat{m}(x) - m(x)] \tag{8}$$

2.1.3 Joint Distribution of $\hat{\alpha}(x)$ and $\hat{p}(x)$

To prove their joint distribution will tend to a normal distribution, we only need to show that every linear combination of these two variables is normal.

$$X = x_1\hat{\alpha}(x) + x_2\hat{p}(x) = \frac{1}{nh_n}\Sigma(x_1y_i + x_2)K(\frac{x - x_i}{h_n}) = \frac{1}{nh_n}\Sigma(t(xi) + \eta_i)K(\frac{x - x_i}{h_n})$$

where $t(x) = x_1m(x) + x_2$ and $\eta_i = x_1\epsilon_i$. It is not hard to observe that X shares the same form as in Theorem 2.1.2, thus the result follows.

Indeed, we can easily calculate $Cov(A, B)$ and $Var(A - m(x)B)$ for later usage as below,

$$\begin{aligned} Cov(A, B) = [m(x)p(x)r(k) + \frac{1}{2}m(x)p''(x)\sigma_k^2h_n^2 + m'(x)p'(x)\sigma_k^2h_n^2 \\ + m''(x)p(x)\sigma_k^2h_n^2 + O(h_n^4)] - h_nE(\hat{\alpha}(x))E(\hat{p}(x)) \end{aligned} \tag{9}$$

$$\begin{aligned} Var(A - m(x)B) &= Var(A) + Var(m(x)B) - 2Cov(A, m(x)B) \\ &= \sigma_\epsilon^2 p(x)r(k) + O(h_n^2) \end{aligned} \tag{10}$$

Here $r(k) = \int K^2(z)dz$ and $\sigma_k^2 = \int K^2(z)z^2dz$.

2.2 Self-Supervised Estimator

We now move on to establish the arguments for the estimator using unlabeled data. The variables are denoted in a similar way as the case of NW-Estimator. Note that the value of w_i completely rely on the prediction of previous NE-Estimator. And under most circumstances, the distribution of u_i is identical to that of x_i.

$$NW_{Unlabled} = \hat{r}(x) = \frac{\hat{\beta}(x)}{\hat{q}(x)} \tag{11}$$

$$\hat{\beta}(x) = \frac{1}{mg_m}\Sigma w_iK(\frac{x - u_i}{g_m}) \tag{12}$$

$$\hat{q}(x) = \frac{1}{mg_m}\Sigma K(\frac{x-u_i}{g_m}) \tag{13}$$

$$w_i = \hat{m}(x_i) \tag{14}$$

2.2.1 Asymptotic normality of $\hat{\beta}(x)$ and $\hat{q}(x)$ conditioned on (X,Y)

First of all, $\hat{q}(x)$ itself is actually not dependent on labeled dataset. Therefore, similar the previous derivation for $\hat{p}(x)$, it is asymptotically normal and independent of $\hat{\alpha}(x)$ and $\hat{p}(x)$.

Regarding $\hat{\beta}(x)$, we shall first examine its conditional distribution on $\hat{\alpha}(x)$ and $\hat{p}(x)$, then figure out their joint distribution. As a matter of fact, it can be shown that $\hat{m}(x), \hat{m}'(x), \hat{m}''(x)$ have small error terms,

$$\begin{aligned}
\hat{m}(x) &= m(x) + O_p(h_n^2 + \frac{1}{\sqrt{nh_n}}) \\
\hat{m}'(x) &= m'(x) + O_p(h_n^2 + \frac{1}{\sqrt{nh_n^3}}) \\
\hat{m}''(x) &= m''(x) + O_p(h_n^2 + \frac{1}{\sqrt{nh_n^5}})
\end{aligned}$$

Hence conditional expectation and variance can be written as,

$$E(\hat{\beta}(x)|X,Y) = q(x)\hat{m}(x) + O(g_m^2) + O_p(g_m^2 h_n^2 + \frac{g_m^2}{\sqrt{nh_n^5}} + g_m^4)$$

$$Var(\hat{\beta}(x)|X,Y) = \frac{1}{mg_m}[\hat{m}^2(x)q(x)\mu_2(k) + O_p(g_m^2)] - \frac{1}{m}E^2(\hat{\beta(x)}|X,Y)$$

Quoting the result for the asymptotically normal $\hat{\alpha}(x)$, we shall easily see that $\hat{\beta}(x)|X,Y$ is also asymptotically normal.

2.2.2 Asymptotic normality of $\hat{\beta}(x)$ conditioned on $\hat{m}(x)$

First, introduce a new variable with conditionally asymptotically normal distribution,

$$C = (mg_m)^{1/2}[\hat{\beta}(x) - E[\hat{\beta}(x)|X, Y]] \tag{15}$$

Based on the conditional distribution of $\hat{\beta}(x)|X, Y$, we are able to derive $F(C|X, Y)$,

$$F(C|X, Y) = P(C \leq c|X, Y) = \Phi(c/\sigma_c)$$

where,

$$\sigma_c^2 = m^2(x)q(x)\mu_2(k) + O_p(h_n^2 + \frac{1}{\sqrt{nh_n}} + g_m + g_m^2)$$

Therefore the asymptotic distribution of C is actually independent of X,Y and $\hat{m}(x)$.

2.2.3 Asymptotic normality of $\hat{r}(x)$ conditioned on $\hat{m}(x)$

Given $\hat{m}(x)$, we can derive the conditional normality of $\hat{r}(x)$ using the same method as that of $\hat{m}(x)$,

$$\begin{aligned}
\hat{r}(x) - \hat{m}(x) &= G((mg_m)^{-1/2}C + E(\hat{\beta(x)}|X, Y), (mg_m)^{-1/2}D + E(\hat{q(x)})) - \hat{m}(x) \\
&= H + (mg_m)^{-1/2}[IC + JD + \frac{C - \hat{m}(x)D}{q(x)}] + HigherOrderTerms \\
&= (mg_m)^{-1/2}\hat{s}(x)
\end{aligned} \tag{16}$$

Where $H = [g_m^2 m'(x)q'(x)\mu_2(k) + \frac{g_m^2}{2}m''(x)q(x)\mu_2(k)]\frac{1}{q(x)} + O_p(\frac{g_m^2}{\sqrt{nh_n^5}} + g_m^4)$. I and J are both of order g_m^2.

Next, I will show the term $C - \hat{m}(x)D$ is asymptotically normal of order g_m. Since its expectation is 0, we only need to consider its asymptotic variance, which is

approximately its conditional variance.

$$\begin{aligned} Var(C - \hat{m}(x)D|X,Y) &= Var(C|X,Y) + Var(\hat{m}(x)D|X,Y) - 2Cov(C,\hat{m}(x)D) \\ &= \hat{m}'(x)^2 q(x)\sigma_k^2 g_m^2 + O(g_m^4) \end{aligned} \tag{17}$$

The distribution of $\frac{C-\hat{m}(x)D}{g_m}|X,Y$ remain to be examined. This can be done by applying Theorem 2.1.1 to $\frac{\hat{\beta}(x)-\hat{m}(x)\hat{q}(x)}{g_m}|X,Y$. Follow the similar calculation of Theorem 2.1.2, we have the following

Proof.

$$\sigma_{im}^2 = Var(z_{im}) = \hat{m}'(x)^2 q(x)\sigma_k^2 g_m^{-1} + o(g_m^{-1})$$

For all r larger than or equal to 3, we have the following

$$\begin{aligned} \rho_{im}^{\frac{1}{r}} &\leq 2E(|\frac{1}{g_m^2}(\hat{m}(y) - \hat{m}(x))K(\frac{y-x}{g_m})|^r)^{\frac{1}{r}} \\ &= 2\frac{1}{g_m^2}(\int(\hat{m}(x) + \hat{m}'(x)zg_m - \hat{m}(x))^r K(z)^r p(z) g_m dz)^{\frac{1}{r}} \\ &= 2\frac{1}{g_m}(\int(\hat{m}'(x)z)^r K(z)^r p(z) g_m dz)^{\frac{1}{r}} \end{aligned}$$

Therefore, $\rho_{im} \leq O(g_m^{-r+1})$ and if $mg_m \to \infty$, then $\frac{(\Sigma_{i=1}^m \rho_{im})^{1/r}}{(\Sigma_{i=1}^m \sigma_{i\,m}^2)^2} \to 0$ □

Hence, $\frac{C-\hat{m}(x)D}{g_m}$ given X and Y is asymptotically normal. Moreover, it is asymptotically independent of X,Y because $\hat{m}'(x) \to []pm'(x)$.

2.3 Hybrid Estimator

We now proceed to the ultimate results about hybrid estimator, which is defined as,

$$\hat{y_c}(x) = \lambda\hat{m}(x) + (1-\lambda)\hat{r}(x) \tag{18}$$

Combining equation (7) and (16), we can expand $\hat{y_c}(x)$ to be $\hat{y_c}(x)$

$$\hat{y_c}(x) - m(x) = h_n^2 E + (nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} + (1-\lambda)[g_m^2 H + -\frac{(mg_m)^{-1/2}}{g_m}\frac{C - \hat{m}(x)D}{q(x)g_m}] + o_p((nh_n)^{-1/2} + \frac{h_n^2(mg_m)^{-1/2}}{g_m})$$

Where $E = [m'(x)p'(x)\mu_2(k) + \frac{1}{2}m''(x)p(x)\mu_2(k)]\frac{1}{p(x)} + O(h_n^2)$, and $H = [m'(x)q'(x)\mu_2(k) + \frac{1}{2}m''(x)q(x)\mu_2(k)]\frac{1}{q(x)} + O_p(\frac{1}{\sqrt{nh_n^5}} + g_m^2)$ Since A,B,C,D all follows asymptotic normal distribution with mean zero, to minimize the bias of $\hat{y_c}(x)$, we should choose $1 - \lambda = -\frac{h_n^2 E}{g_m^2 H}$.

Then the expression of $\hat{y_c}(x)$ is further reduced to

$$\hat{y_c}(x) - m(x) = (nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} - \frac{h_n^2(mg_m)^{-1/2}}{g_m}\frac{E}{H}\frac{C - \hat{m}(x)D}{q(x)g_m} + o_p((nh_n)^{-1/2} + \frac{h_n^2(mg_m)^{-1/2}}{g_m})$$

Therefore we can find the distribution of $\hat{y_c}(x) - m(x)$ through calculating its variance, which is of order $max\{(nh_n)^{-1/2}, (mg_m)^{-1/2}\frac{h_n^2}{g_m^2}g_m\}$. Hence the order of variance is smallest when $mg_m^3 \propto nh_n^5$.

When n and m are large enough, we can approximate the variance(equivalent to mse here because the bias is minimised) with a simple equation, using results (10) and (17),

$$\begin{aligned} Var(\hat{y_c}(x) - m(x)) &= Var((nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} - \frac{h_n^2(mg_m)^{-1/2}}{g_m^2}g_m\frac{E}{H}\frac{C - \hat{m}(x)D}{q(x)g_m}) \\ &\approx \frac{1}{nh_n}\frac{1}{p^2(x)}Var(A - m(x)B) + \frac{h_n^4(mg_m)^{-1}}{g_m^4}g_m^2\frac{E^2}{H^2q(x)^2}Var(\frac{C - \hat{m}(x)D}{g_m}) \\ &= \frac{1}{nh_n}\frac{1}{p(x)}\sigma_\epsilon^2 r(k) + \frac{h_n^4}{mg_m^3}\frac{E^2}{H^2q(x)}\hat{m}'(x)^2\sigma_k^2 \end{aligned} \tag{19}$$

because $\frac{C-\hat{m}(x)D}{q(x)g_m}$ is asymptotically independent of (X,Y), and therefore of $A - m(x)B$. In the next section, we will demonstrate how this approximation performs under difference circumstances.

3 Experiments and Simulations

Comparing with traditional NW-Estimator which have optimal MSE $O(n^{-4/5})$ when $h_n \propto n^{-1/5}$, mse of hybrid estimator given by equation 19 will have always have higher convergence rate. Its confidence interval based on the normal distribution approximation, however, can not be directly compared through just the formula. Two experiments under different statistical settings are conducted to demonstrate the ideas.

3.1 Experiment 1

Statistical Model: $m(x) = x^2$, $X_i \overset{iid}{\sim} N(0, \sigma_x^2), \sigma_x = 5$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \sigma = 1$. The experiment is carried out for n = 32, 64, 128 and $m = n^{10/19}$. Our test data point of x = 2, m(x) = 4. To estimate the mse more accurately, resampling method, i.e. bootstrap, is used. Subsequently, grid search for optimal (h,g) is done to find the smallest mse. Then this finding is used to construct confidence intervals.

3.1.1 Mean Square Error

The results of grid search is summarized in Table 3.1. Although the optimal mse varies across choices of n, we can see that hybrid estimator performs better than NW estimator.

Details of how mse changes with h are depicted in figure 3.1a, 3.1b and 3.1c. Clearly, the optimal choice of h for hybrid estimator is in general larger than that of

n	h(nw)	mse(nw)	h(hybrid)	g(hybrid)	mse(hybrid)
32	0.30	0.4736	0.3	6.0	0.4710
64	0.10	0.3246	0.1	0.3	0.3024
128	0.20	0.8298	0.25	0.5	0.7395

Table 3.1: Grid Search Results

n	h_{sd}	g_{sd}	Coverage	h_{ci}	g_{ci}	Coverage
32	0.1	1.0	0.95	0.3	6.0	0.72
64	0.3	0.1	0.95	0.1	0.3	0.87
128	0.6	0.3	0.95	0.25	0.5	0.64

Table 3.2: Comparison between choice of h and g

NW estimator, resulting in a smaller minimal mse. This fits our formula and expectation.

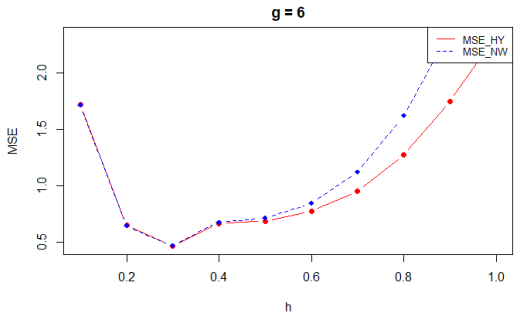
3.1.2 Confidence Interval

Besides mse which is a common criterion for point estimation, confidence interval is a crucial subject of interest in statistical inference as well. Making use of equation 19, we can then construct CI based on several reasonable assumptions. Firstly, since p(x) and q(x) are generally identical in most cases, we can see that E and H are virtually the same. Secondly, for the sake of convenience, we assume that the variance of error σ is known. Therefore, our estimation of $Var(\hat{y_c}(x))$ is as follows.

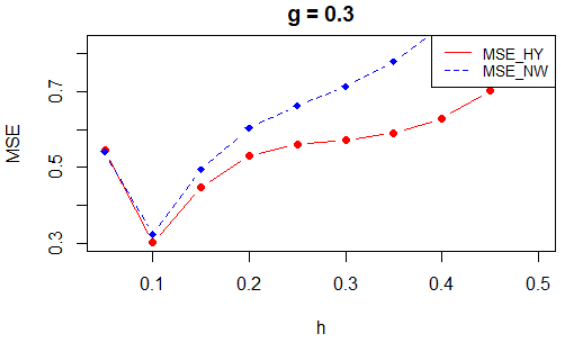
$$\hat{Var}(\hat{y_c}(x)) = \frac{1}{nh_n}\frac{1}{\hat{p}(x)}\sigma_\epsilon^2 r(k) + \frac{h_n^4}{mg_m^3}\frac{1}{\hat{q}(x)}\hat{m}'(x)^2\sigma_k^2$$

With this estimation, 0.95 level confidence level can be attained. We set $h = 0.1, 0.2, ..., 1$ and $g = 0.1, 0.2, ..., 6$ and loop through all the combinations. The true coverage probability is estimated with 100 rounds of repetition.

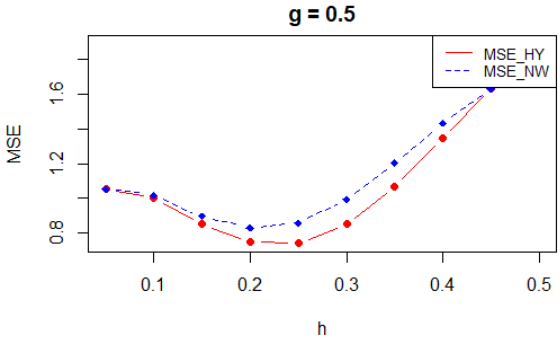
In Table 3.2, h_{sd} and g_{sd} stand for the choice of h and g according to mse while h_{ci} and g_{ci} mean the h and g that provide us with the best coverage probability which is exactly 0.95(note that usually more than one pair of h and g can give this result).



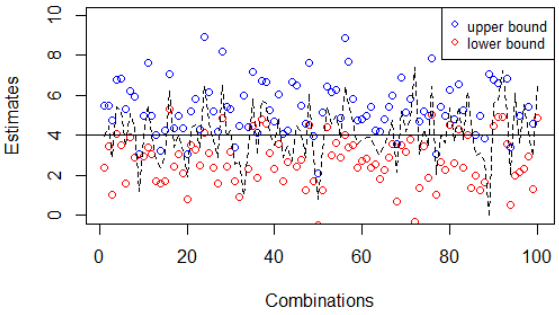
(a) n = 32



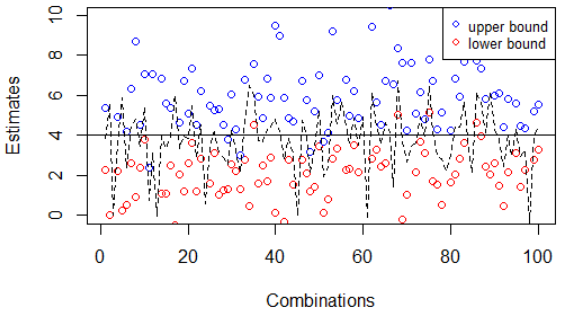
(b) n = 64



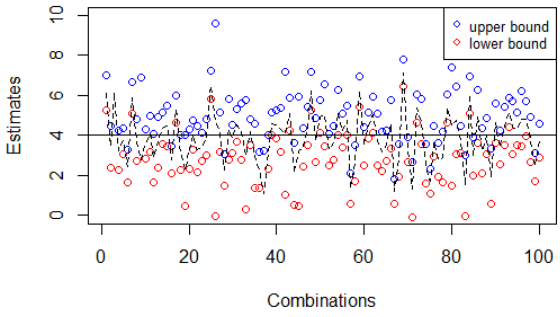
(c) n = 128



(d) n = 32



(e) n = 64



(f) n = 128

Figure 3.1: Mean Square Error and Coverage of CI

Carefully examining Figure 3.1d, 3.1e and 3.1f, it seems that when n is large, formula 3.1.2 tend to underestimate the true variance.

3.2 Experiment 2

4 | Conclusion and Discussion

To summarize, the application of semi-supervised learning in Nadaraya-Watson estimator to leverage unlabeled data is explored in this report. Asymptotic distribution of the hybrid estimator is derived and its expectation and variance are expressed explicitly. These results are subsequently tested in experiments and put into the approximation of confidence interval using hybrid estimator. The main findings are related to the choice of h and g. Firstly, bandwidth selection holds the key to optimal estimation in the proposed hybrid estimator as it is in kernel regression. Secondly, the choice of (h,g) varies according to the purpose of estimation. The best choice if one focus more on mean square error differs greatly from the alternative if one want to achieve the suitable coverage probability.

Acknowledgement

Bibliography

D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

E. Nadaraya. On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1):186–190, 1965.

M. P. Wand and M. C. Jones. *Kernel smoothing*. CRC press, 1994.

D. Wied and R. Weißbach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53(1):1–21, 2012.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.