香 港 大 學

THE UNIVERSITY OF HONG KONG

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

SEMI-SUPERVISED LEARNING WITH NW-ESTIMATOR

AUTHOR: JUNSHI WANG
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: December 2, 2021

DECEMBER 2, 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 $\hat{p}(x)$.

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

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

$$\hat{p}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x - x_i}{h_n}) \tag{3}$$

$$y_i = m(x_i) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{4}$$

The following assumptions are generally adopted (Wand and Jones, 1994):

(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*

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

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 $\frac{1}{nh_n} \sum_{i=1}^{n} y_i \cdot K(\frac{x-x_i}{h_n})$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

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

□

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.

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

$$\hat{\beta}(x) = \frac{1}{mg_m}\sum_{i=1}^{n} w_iK(\frac{x - u_i}{g_m}) \tag{12}$$

$$\hat{q}(x) = \frac{1}{mg_m} \sum_{i=1}^{n} K(\frac{x - u_i}{g_m}) \tag{13}$$

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

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 (Tang, 2021),

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

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 $\hat{m}(x)$

To facilitate the calculation, first 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.

$$\begin{aligned} Var(C - \hat{m}(x)D|X,Y) &= Var(C|X,Y) + Var(\hat{m}(x)D|X,Y) - 2Cov(C, \hat{m}(x)D) \\ &= \hat{m}'(x)^2 q(x)\sigma_k^2 g_m^2 + O(g_m^4) \end{aligned} \tag{16}$$

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 $\frac{\hat{\beta}(x)-\hat{m}(x)\hat{q}(x)}{g_m}|X,Y$. Going through the similar calculation as 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 $r \geq 3$, 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 $\frac{(\Sigma_{i=1}^m \rho_{im})^{1/r}}{(\Sigma_{i=1}^m \sigma_i^2{}_m)^2} \to 0$ □

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 $\hat{m}'(x) \xrightarrow{p} m'(x)$.

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

$$\begin{aligned}\hat{r}(x) - \hat{m}(x) &= G((mg_m)^{-1/2}C + E(\hat{\beta}(x)|X,Y), (mg_m)^{-1/2}D + E(\hat{q}(x))) - \hat{m}(x) \\ &= H + (mg_m)^{-1/2}g_m[IC + JD + \frac{C - \hat{m}(x)D}{q(x)g_m}] + O_p((mg_m)^{-1}g_m) \qquad (17) \\ &= (mg_m)^{-1/2}\hat{s}(x)\end{aligned}$$

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.

2.3 Hybrid Estimator

Based on $\hat{m}(x)$ and $\hat{r}(x)$, a hybrid estimator $\hat{y}_c(x)$ can be constructed with improved asymptotic properties.

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

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

Combining equation (7) and (17), we can expand $\hat{y_c}(x)$ to be $\hat{y_c}(x)$

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

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

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

Therefore we can find the distribution of $\hat{y_c}(x) - m(x)$ through calculating its variance, which is of order $max\{(nh_n)^{-1/2}, h_n^2(mg_m^3)^{-1/2}\}$. 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 with a simple equation, using formula (10) and (16) in which $\hat{\sigma_\epsilon^2}$ and $\hat{m}'(x)^2$ can be readily attained,

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

$$Var(\hat{y_c}(x) - m(x)) \approx \frac{1}{nh_n}\frac{1}{p(x)}\hat{\sigma_\epsilon^2} r(k) + \frac{h_n^4}{mg_m^3}\frac{E^2}{H^2 q(x)}\hat{m}'(x)^2\sigma_k^2 \tag{20}$$

because $\frac{C - \hat{m}(x)D}{q(x)g_m}$ is asymptotically independent of (X,Y), and therefore of $A - m(x)B$.

As for bias, it is much harder to estimate due to the presence of high order derivatives. Moreover, bias term is essentially negligible when h_n and g_m are small and chosen with caution.

$$E(\hat{y_c}(x) - m(x)) = (\frac{h_n^4 - h_{ng}^2 g_m^2}{2})[\frac{m'(x)p'''(x)}{p(x)} + \frac{m''(x)p''(x)}{p(x)} + \frac{m'''(x)p'(x)}{p(x)} - \frac{m'(x)p'(x)p''(x)}{p(x)^2}]\mu_k^2 \tag{21}$$

To approximate the precision of hybrid estimator, the above equations are utilized to produce a 95% confidence interval. Here we provide the proposed procedure to attain this confidence interval when $\hat{\sigma_\epsilon^2}$ is unknown.

Algorithm 1: Confidence Interval for Hybrid Estimator(σ_ϵ^2 unknown)

Input: h, g, m, n
for *i in 1:rounds* **do**
  Generate data: labeled dataset of size n and unlabeled dataset of size m
  Find h_1^{opt} for kernel density estimator
  Calculate $\hat{p}(x) = \hat{q}(x)$ using h_1^{opt}
  Find h_2^{opt} for NW estimator through leave-one-out cross validation
  Calculate $\hat{\sigma_\epsilon^2}$ and $\hat{m}'(x)$ using h_2^{opt}
  Find 95% confidence interval based on Equation 20 and 21
  Test whether $m(x)$ is covered in the confidence interval or not
Output the average length of interval and coverage probability

In the next section, the author will demonstrate how the proposed hybrid estimator performs under difference pairs of (h, g) in terms of mean square error. Its normal approximation given by Equation 20 and 21 will be evaluated as well through the length and coverage probability of the confidence interval.

3 | Experiments and Simulations

Compared 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 20 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 Main features of the proposed hybrid estimator

Objective of the first experiment is to demonstrate how hybrid estimator behaves under different choice of parameters.

$$\textbf{Statistical Model} : m(x) = x^2, X_i \overset{iid}{\sim} N(0, \sigma_x^2), \sigma_x = 1, \epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2), \sigma_\epsilon = 1$$

$$\textbf{Choice of size} : n = 32, 64 \text{ or } 128, m = n^{10/19}$$

$$\textbf{Testing data point} : x = 1.5, m(x) = 1.5^2$$

To estimate the true MSE at each choice of (h, g), repeated calculations have been carried out for both NW estimator and the proposed hybrid estimator and Monte Carlo error has been estimated (Koehler et al., 2009). Subsequently, grid search for optimal (h, g) is done to find the smallest MSE. Then the pair of (h, g) with smallest MSE is used to construct confidence intervals and the coverage probability is tested.

Labeled	Unlabeled	h	MSE $\pm$ sd.	h	g	MSE $\pm$ sd.
32	6	0.30	0.447 $\pm$0.006	0.30	1.5	0.440 $\pm$0.005
64	9	0.25	0.228 $\pm$0.004	0.30	0.9	0.220 $\pm$0.003
128	13	0.20	0.121 $\pm$0.002	0.25	0.9	0.112 $\pm$0.002

Table 3.1: Grid Search Results

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. When $n = 32$, the difference is relatively small. In fact, the performances of two estimators at $n = 32$ are indistinguishable due to MC error. However, with the increase of data points, the benefit of using hybrid estimator become obvious.

Details of how MSE changes with h are depicted in figure 3.1a, 3.1b and 3.2a. It seems that the optimal choice of h for hybrid estimator is in very close to that of NW estimator, but resulting in a smaller minimal MSE. This agrees with the 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 20, we can then construct CI based on several reasonable assumptions. Firstly, since p(x) and q(x) are generally identical in most cases, we may assume that E and H are virtually the same. Cases where the variance of error σ is be known and unknown are explored separately. In general, the estimation of $Var(\hat{y_c}(x))$ is as follows.

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

With this estimation, 0.95 level confidence level can be attained. We set $h = 0.1, 0.15, ..., 1$ and $g = 0.3, 0.6, ..., 2.4$ and loop through all the combinations.

Labeled	h	Width	Coverage	h	Width	Coverage
32	0.10	6.662	0.89	0.30	1.895	0.71
64	0.10	2.548	0.90	0.25	1.391	0.85
128	0.10	1.664	0.86	0.20	1.126	0.83

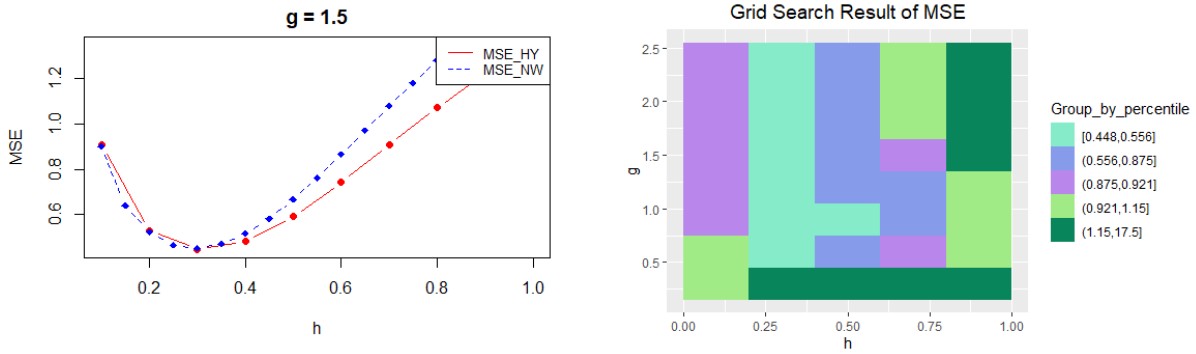
Table 3.2: Coverage probability and width of interval based on NW estimator

Labeled	Unlabeled	h	g	Width	Coverage	h	g	Width	Coverage
32	6	0.10	1.5	2.890	0.90	0.30	1.5	1.670	0.83
64	9	0.10	2.4	2.084	0.91	0.30	0.9	1.244	0.83
128	13	0.10	1.8	1.486	0.94	0.25	0.9	0.963	0.83

Table 3.3: Coverage probability and width of interval based on Hybrid estimator(error known)

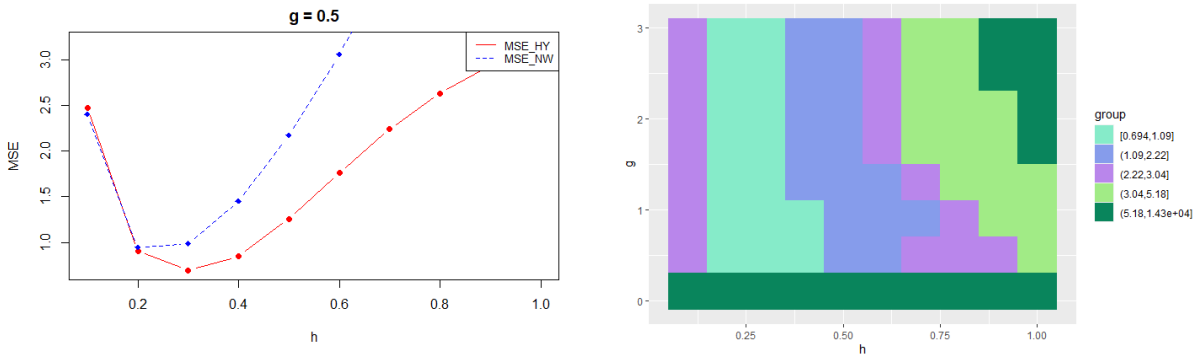
The true coverage probability is estimated with Algorithm 1(algorithms are basically the same for NW estimator and the case when σ_ϵ is known).

In Table 3.2, 3.3 and 3.4, some typical pairs of (h, g) are chosen for comparison. In each table, on the left are choices of h that have the best coverage probability and on the right are those with minimal mean square error. There are several comments to be made for the tables. First of all, smaller h and g tend to yield better coverage(details can be found in heatmaps below) and hybrid estimator commonly performs better because it has less coverage error and shorter width. Secondly, the confidence interval based on (h, g) that minimizes mean square error seems to underestimate the variance of hybrid estimator, resulting in short width and lower coverage probability. And the coverage does not increase with n but stays at around 0.83 This suggests we should use normal approximation with extra caution(some solutions will be proposed in **??**). Carefully examining the heatmap (Figure 3.2b, 3.3a and 3.3b) depicting changes of error(difference between 0.95 and coverage probability) when h and g changes. One main observation from the table is that the error of confidence interval continues to decrease as n increases. It seems that when h and g are both large, the performance of proposed confidence interval is less satisfactory, which is expected and stress the effect of bandwidth.



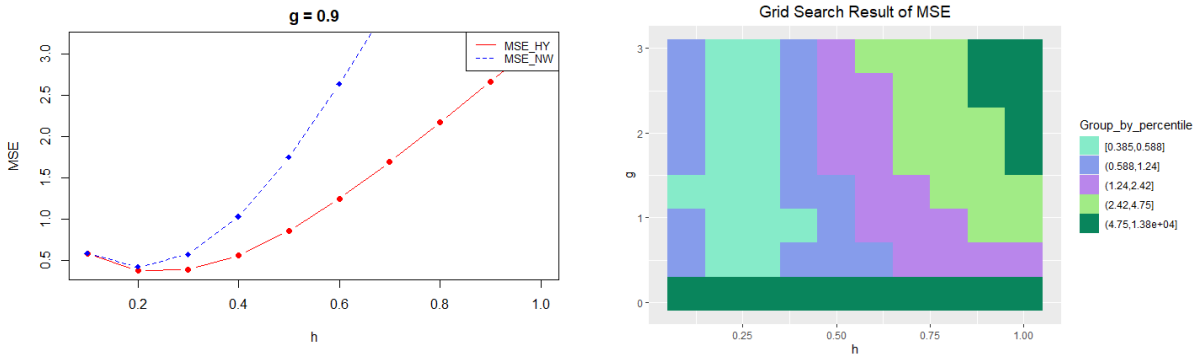
(a) MSE with g fixed at 1.5 (b) Heatmap depicting MSE by percentile

Figure 3.1: Mean Square Error: n = 32



(a) MSE with g fixed at 0.9 (b) Heatmap depicting MSE by percentile

Figure 3.2: Mean Square Error: n = 64



(a) MSE with g fixed at 1.5 (b) Heatmap depicting MSE by percentile

Figure 3.3: Mean Square Error: n = 128

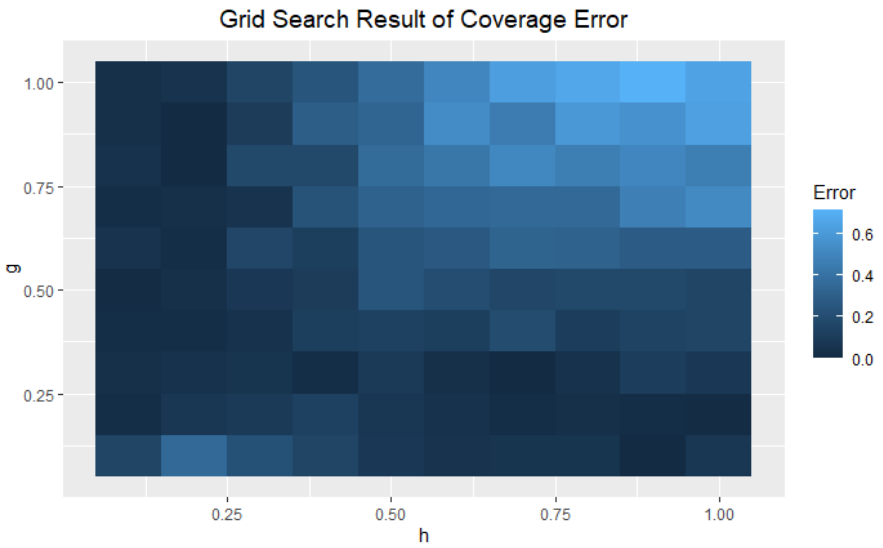
Labeled	Unlabeled	h	g	Width	Coverage	h	g	Width	Coverage
32	6	0.10	0.9	5.013	0.92	0.30	1.5	1.647	0.73
64	9	0.50	0.3	4.514	0.95	0.30	0.9	1.240	0.81
128	13	0.10	2.1	1.571	0.95	0.25	0.9	0.946	0.84

Table 3.4: Coverage probability and width of interval based on Hybrid estimator(error unknown)

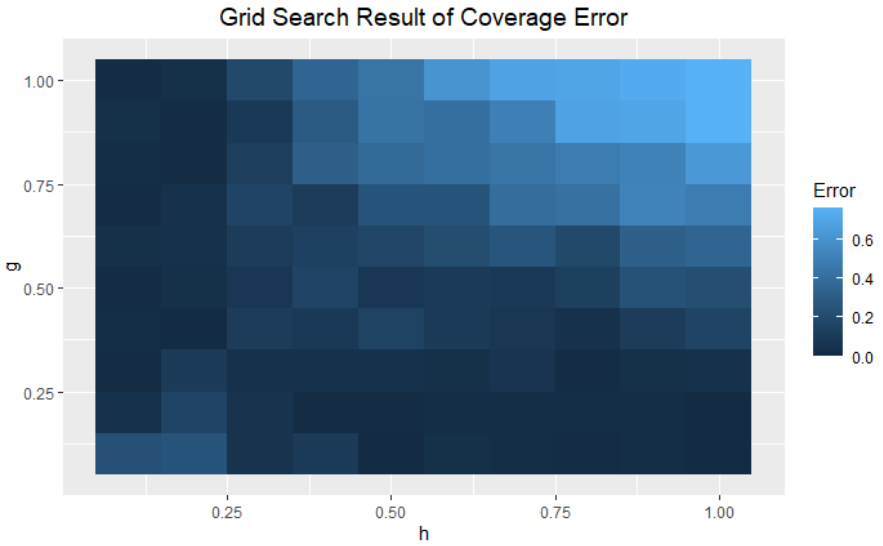
3.1.3 Bandwidth selection

As already discussed, (h, g) influences the performance of hybrid estimator dramatically. Here we further explain its importance. In fact the optimal choice of bandwidth depends on the objective of operation, i.e. whether mean square error or confidence interval is at concern. Based on Figure 3.1b, 3.2b, 3.3b and 3.4, conclusion can be drawn that the pair of (h, g) that provides the smallest mean square error doesn't necessarily secure the best performance in terms of coverage probability. Those who lay emphasis on confidence interval are suggested to construct one using other pairs of (h, g).

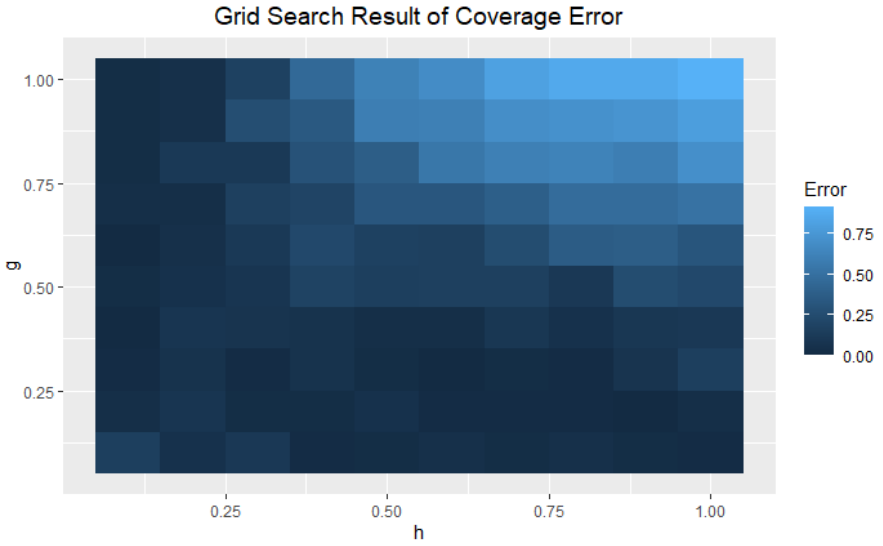
3.2 Experiment 2



(a) n = 32



(b) n = 64



(c) n = 128

Figure 3.4: Coverage Probability and the choice of (h, g)

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 pair of h and g that generates best *mse* does not lead us directly to the optimal confidence. Nevertheless, with smallest length, the confidence interval construct using hybrid estimator still gives around 90% of true coverage and the coverage becomes better as n and m increase.

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.

E. Koehler, E. Brown, and S. J.-P. Haneuse. On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162, 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.

Y. Tang. Semi-supervised learning for non-parametric regression: A technical report. 2021.

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.