香 港 大 學

THE UNIVERSITY OF HONG KONG

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Semi-Supervised Learning based on NW Estimator

Author: Junshi WANG
Project Supervisor: Prof. Stephen M.S. 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 12, 2021

November 11, 2022

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. We intend to show that the choice of (h,g) is of great importance and the decision depends largely on the objective of research.

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.

There are dozens of papers in semi-supervised learning and statistical inference(in our case, kernel regression) with rigorous proof. However, these two subjects which receive enormous attention in respective fields are seldom related. Here we propose a method inspired by the idea of semi supervised learning to construct a hybrid estimator based on NW estimator.

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 $\hat{p}(x)$.

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

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

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

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

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 $x = x_0$ 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 $\mathbf{Var}(z_{in}) \equiv \sigma_{in}^2$ and r-th absolute central moment $\mathbf{E}[|z_{in} - \mathbf{E}(z_{in})|^r] \equiv \rho_{in} < \infty$ 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 $\overline{z_n}$ 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 $\frac{1}{nh_n}\sum_{i=1}^{n} y_i K(\frac{x-x_i}{h_n})$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

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

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

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

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

$$\begin{aligned}&\mathbf{E}(|m(x_i)^q\epsilon^{r-q}K(\frac{x-x_i}{h_n})^r|) \leq \mathbf{E}(|\epsilon|^{r-q})\mathbf{E}(|m(x_i)^qK(\frac{x-x_i}{h_n})^r|) \\ &= \mathbf{E}(|\epsilon|^{r-q})[\int_{\mathbf{C_1}} m(y)^qK(\frac{x-y}{h_n})^r p(y)dy - \int_{\mathbf{C_2}} m(y)^qK(\frac{x-y}{h_n})^r p(y)dy] \\ &= O(h_n)\end{aligned}$$

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

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

Then the result follows. □

2.1.2 Asymptotic Distribution of NW Estimator

After proving normality of $\hat{\alpha}(x)$ and $\hat{p}(x)$, we are able to discuss the distribution of $\hat{m}(x)$. To simplify the expressions, we define random variables A and B.

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

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

Then we can write NW estimator in the following way,

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

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's Theorem** to the equation and focus on the first order terms. For the sake of convenience, define function $g(s,t) = \frac{s}{t}$. Its first derivative is

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

and we can express $\hat{m}(x) - m(x)$ as a linear combination of random variables following asymptotic normal distribution.

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

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 of order h_n^2. When $h_n^2 \succ (nh_n)^{-1/2}$, the leading term is E, and when $h_n^2 \prec (nh_n)^{-1/2}$, the leading term will be $(nh_n)^{-1/2}\frac{A-m(x)B}{p(x)}$.

Therefore $\hat{t}(x)$ 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 if (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 the joint distribution of (A, B) tends to a normal distribution, we only need to show that every linear combination of these two variables is normal. Let $x_1, x_2 \in \mathbf{R}$, and a linear combination X can be formulated as,

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

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 the one examined in Theorem 2.1.2, thus the result follows.

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

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

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

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, $\mathbf{U}$. 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 NW estimator. And under most circumstances, the distribution of u_i is identical to that of x_i.

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

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

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

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

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

First of all, $\hat{q}(x)$ itself is not dependent on labeled data set. Therefore, similar to 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,

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

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

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

2.2.2 Asymptotic independence

Even though the conditional distribution of $\hat{\beta}(x)$ is explicit, its marginal distribution is still very complicated. However, we may try to transform $\hat{\beta}(x)$ into the sum of its expectation(conditional expectation) and a asymptotic normal random variable of order $(mg_m)^{-1/2}$ so as to use **Taylor's Theorem**. Resembling the procedure for NW estimator, introduce the new variable with conditionally asymptotically normal distribution,

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

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

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

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 $\mathbf{X},\mathbf{Y}$ and subsequently $\hat{m}(x)$.

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

To facilitate the calculation, first show the term $C - \hat{m}(x)D$ | X,Y is asymptotically normal of order g_m, where $D = (nh_n)^{1/2}[\hat{q}(x) - \mathbf{E}[\hat{q}(x)]]$. Since its expectation is 0, we only need to consider its asymptotic variance, which is approximately its conditional variance.

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

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

Theorem 2.1.2, we have the following,

Proof.

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

For all $r \geq 3$, we have the following

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

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

Hence, $\frac{C - \hat{m}(x)D}{g_m}$ given X and Y is asymptotically normal. Moreover, it is asymptotically independent of $\mathbf{X}, \mathbf{Y}$ because $\hat{m}'(x) \xrightarrow{p} m'(x)$. g Subsequently, we can simplify $\hat{r}(x)$ into a linear combination of random variables using the same method as that of $\hat{m}(x)$,

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

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. (details available in full report).

When n and m are large enough, we can approximate the variance with a simple equation, using formula (10) and (16). $\hat{\sigma_\epsilon^2}$ and $\hat{m}'(x)^2$ can be attained using NW estimator.

$$\begin{aligned}\mathbf{Var}(\hat{y}_c(x) - m(x)) &= \mathbf{Var}((nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} - \frac{h_n^2(mg_m)^{-1/2}}{g_m^2} g_m \frac{E}{H}\frac{C - \hat{m}(x)D}{q(x)g_m}) \\ &\approx \frac{1}{nh_n}\frac{1}{p^2(x)}\mathbf{Var}(A - m(x)B) + \frac{h_n^4}{mg_m^3}\frac{(E/H)^2}{q(x)^2}\mathbf{Var}(\frac{C - \hat{m}(x)D}{g_m}|\mathbf{X},\mathbf{Y}) \\ &= \frac{1}{nh_n}\frac{1}{\hat{p}(x)}\hat{\sigma_\epsilon^2} r(k) + \frac{h_n^4}{mg_m^3}\frac{E^2}{H^2\hat{q}(x)}\hat{m}'(x)^2\sigma_k^2\end{aligned} \tag{18}$$

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. However, bias term is essentially negligible when h_n and g_m are small and chosen with caution so we prefer not to take it into consideration.

$$\begin{aligned}\mathbf{E}(\hat{y}_c(x) - m(x)) = (&\frac{h_n^4 - h_n^2 g_m^2}{2})[\frac{m'(x)p'''(x)}{p(x)} + \frac{m''(x)p''(x)}{p(x)} + \frac{m'''(x)p'(x)}{p(x)} \\ &- \frac{m'(x)p'(x)p''(x)}{p(x)^2}]\mu_k^2\end{aligned} \tag{19}$$

To approximate the precision of hybrid estimator, the above equations are utilized to produce a 95% confidence interval. Here we provide the proposed procedure(Algorithm 1) to attain this confidence interval when $\hat{\sigma_\epsilon^2}$ is unknown.

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

```
Algorithm 1: Confidence Interval for Hybrid Estimator($\sigma_\epsilon^2$ 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 $\hat{p}(x) \approx \hat{q}(x)$ 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 18 and 19
    Test whether $m(x)$ is covered in the confidence interval or not
Output the average length of interval and coverage probability
```

normal approximation given by equation (18) and (19) will be evaluated as well through the length and coverage probability of the confidence interval.

3 | Experiments and Simulations

Compared with NW estimator which have optimal MSE $= O(n^{-4/5})$ when $h_n \propto n^{-1/5}$, MSE of hybrid estimator given by equation 18 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. The setup for experiment is as follows.

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

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

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

$$\textbf{Choice of } h, g: h = 0.1, 0.15, ..., 1, \text{and } g = 0.3, 0.6, ..., 2.4$$

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

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

Table 3.1: Grid Search Results of Mean Square Error

tested.

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 (18), 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. The true coverage

Labeled	h	Width	Coverage	h	Width	Coverage
32	0.10	6.66	0.89	0.30	1.90	0.71
64	0.10	2.54	0.90	0.25	1.39	0.85
128	0.10	1.66	0.86	0.20	1.13	0.83

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

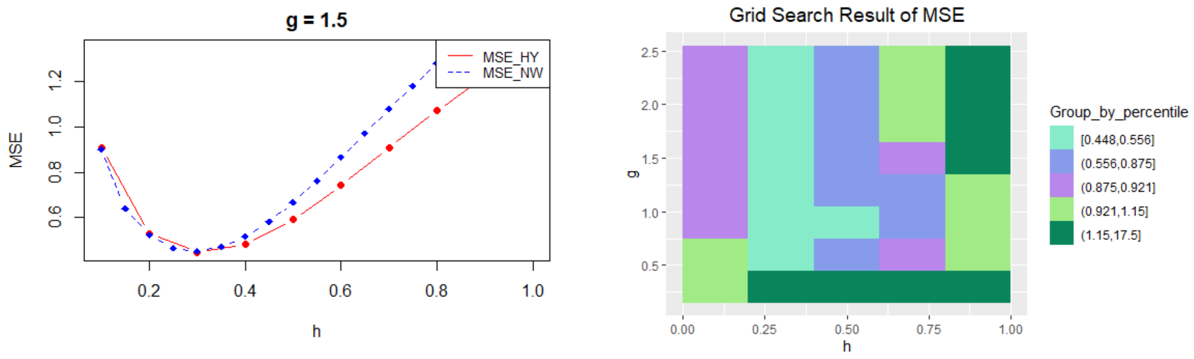
Labeled	Unlabeled	h	g	Width	Coverage	h	g	Width	Coverage
32	6	0.10	1.5	2.89	0.90	0.30	1.5	1.67	0.83
64	9	0.10	2.4	2.08	0.91	0.30	0.9	1.24	0.83
128	13	0.10	1.8	1.49	0.94	0.25	0.9	0.96	0.83

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

probability is estimated with Algorithm 2(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 Section 3.2). 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.

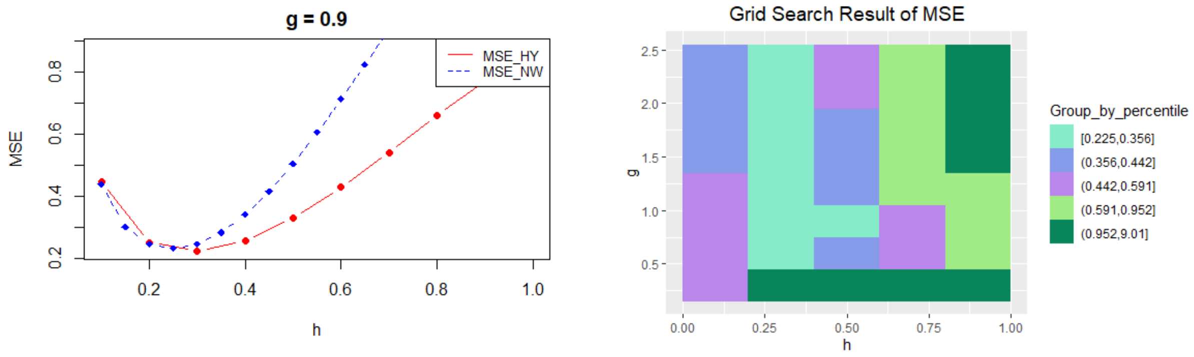
Figure 3.1: Mean Square Error: n = 32



(a) MSE with g fixed at 1.5

(b) Heatmap depicting MSE by percentile

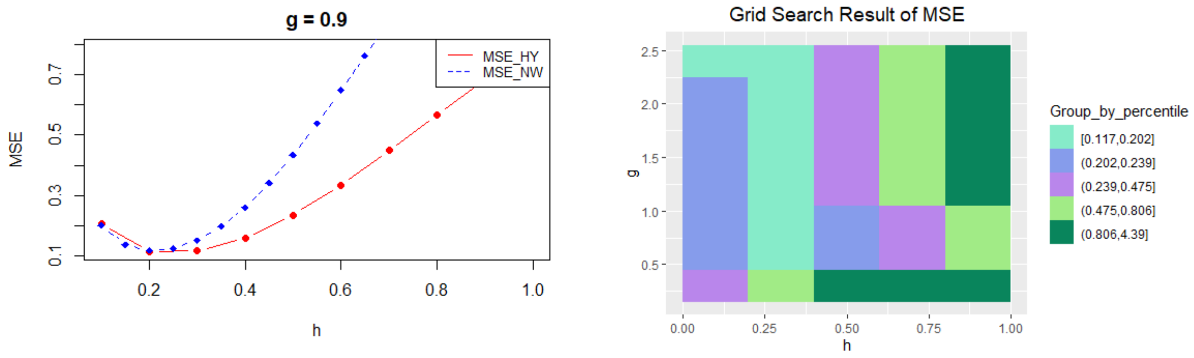
Figure 3.2: Mean Square Error: n = 64



(a) MSE with g fixed at 0.9

(b) Heatmap depicting MSE by percentile

Figure 3.3: Mean Square Error: n = 128



(a) MSE with g fixed at 1.5

(b) Heatmap depicting MSE by percentile

Labeled	Unlabeled	h	g	Width	Coverage	h	g	Width	Coverage
32	6	0.10	0.9	5.01	0.92	0.30	1.5	1.65	0.73
64	9	0.50	0.3	4.51	0.95	0.30	0.9	1.24	0.81
128	13	0.10	2.1	1.57	0.95	0.25	0.9	0.95	0.84

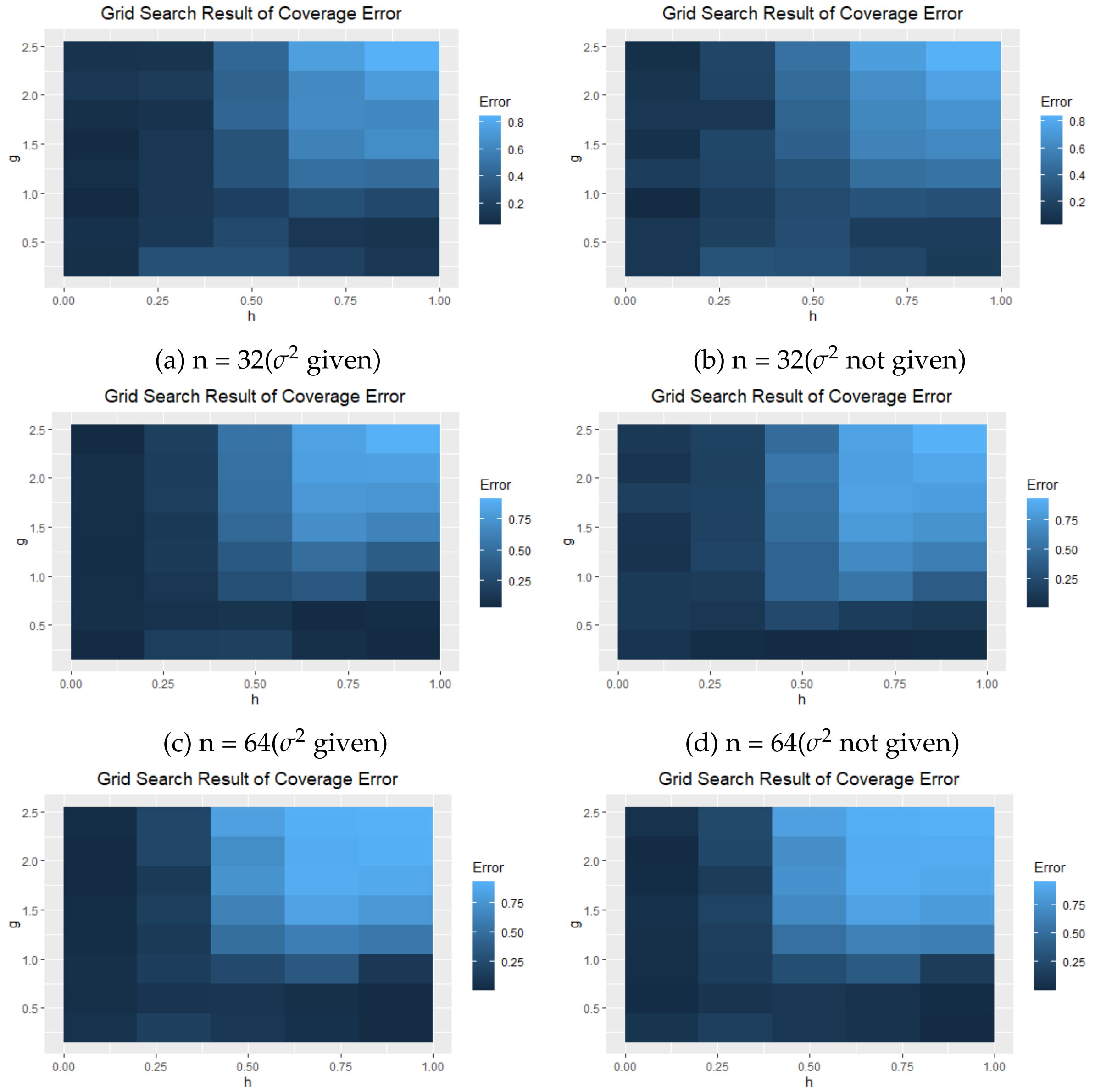
Table 3.4: Coverage probability and width of interval based on Hybrid estimator(error unknown)

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 3.5, 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 Preliminary solutions for underlying issues

As already mentioned briefly, there are several disturbing problems spotted in the simulation. On one hand, grid search present us with certain insights about the theoretical properties of hybrid estimator compared with NW estimator. Nevertheless, practical issues await to be tackled. The most pressing one is how to find the suitable (h,g). It is unusual to be given free access to a sufficient dataset in real life. When it comes to confidence interval, some statistical techniques may be adopted to approximate the coverage error. Here we test the effect of leave-one-out cross validation. On the other hand, evidence suggests that the proposed estimator as well as NW estimator converges to normal distribution quite slow unlike the usual rate $n^{-1/2}$ resulted from **Central Limit Theorem**. This is also supported by the stagnant improvement of coverage probability. We will attempt bootstrap as an



(a) n = 32(σ^2 given)

(b) n = 32(σ^2 not given)

(c) n = 64(σ^2 given)

(d) n = 64(σ^2 not given)

(e) n = 128(σ^2 given)

(f) n = 128(σ^2 not given)

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

alternative for confidence interval.

3.2.1 Bandwidth selection: Leave-One-Out Cross Validation

To take full advantage of the superiority of hybrid estimator, we use cross validation to choose (h,g) that provides the smallest cross validation coverage error. In fact, the best scenario is that we can test the confidence interval on points that are close enough to test data point(i.e. $x = 1.5$). However, due to the constraint of sample size, this will make our conclusion unreliable. Therefore, we choose to include all labeled data and use leave-one-out cross validation.

Algorithm 2: Bandwidth Selection for Hybrid Estimator with cross-validation

```
Input: $train_{label}$ of size $n = 64$, $train_{un}$ of size $m = 9$ and pairs of $(h,g)$ for
       selection
Estimate $\sigma_{\epsilon}^2$
for each $(h,g)$ do
    for each data point $i$ in $train_{label}$ do
        $test \leftarrow train_{label}^{(i)}$
        $train_{label\_loo} \leftarrow train_{label}^{(-i)}$
        Find 95% confidence interval based on equation (18) and (19) for test
    Calculate coverage probability
Output $(h,g)$ with least coverage error
```

The repeated simulation is shown in Figure 3.5. Its coverage probability is 0.94 which is quite satisfactory. However, confidence interval with small coverage error but larger width tend to be chosen(compare Table 3.3 with 3.4). Within 100 simulations, $(1.0,0.3)$, $(0.9,0.3)$ and $(0.8,0.3)$ are chosen for more than 70 times, each of them have very long width(and relatively large bias as shown in Figure 3.5). This suggests that bandwidth for confidence interval should be adaptive and only proximate points should be used in cross validation.

3.2.2 Confidence Interval: Bootstrap

If the normal approximation is sufficient then for most (h,g), the 95% confidence interval should have coverage probability close to 0.95. However, this is obviously

Figure 3.5: Estimation and Confidence Interval: n = 64

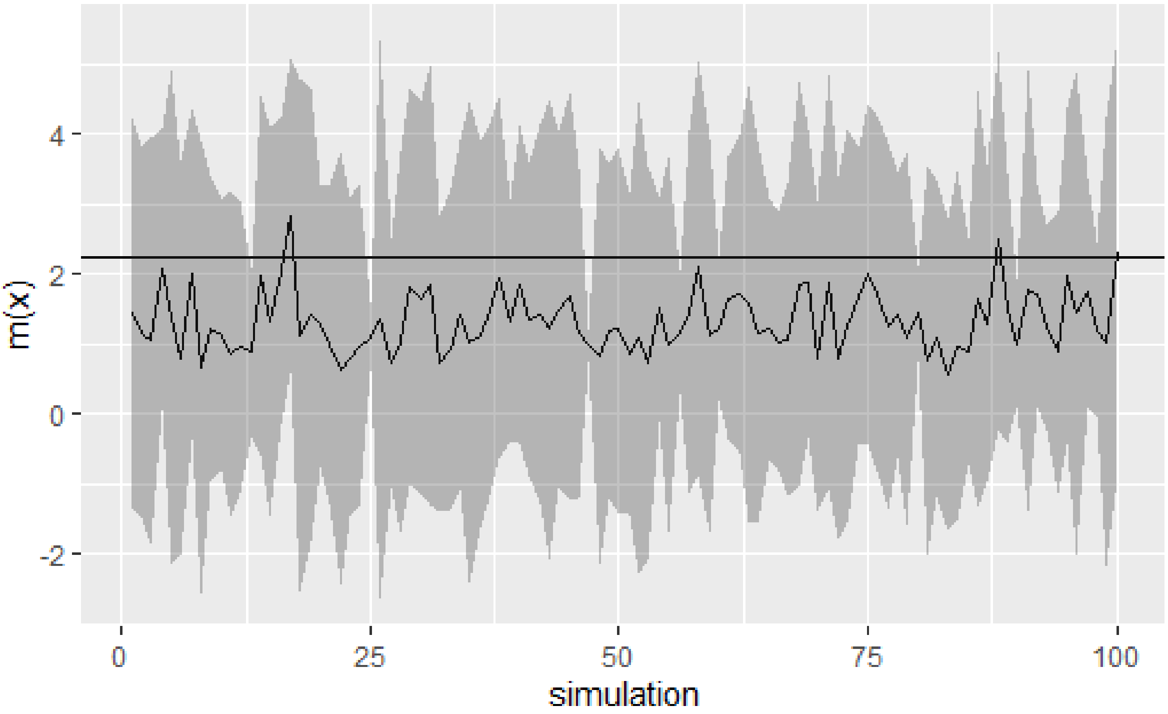


Table 3.5: Coverage probability and width of interval based on LOOCV(error unknown)

Labeled	Unlabeled	h	g	Width	Coverage	Times chosen
64	9	1.0	0.3	16.75	0.92	42
64	9	0.9	0.3	14.79	0.97	15
64	9	0.8	0.3	10.74	0.93	13

Table 3.6: Comparison between method of normal approximation(left) and bootstrap(right) for NW estimator

Labeled	h	Width	Coverage	h	Width	Coverage
32	0.10	6.66	0.89	0.35	1.53	0.79
64	0.10	2.54	0.90	0.35	1.16	0.85
128	0.10	1.66	0.86	0.30	0.87	0.87

not the case for the majority of (h, g). Therefore, an alternative based on bootstrap is worth trying. Residual bootstrap has been used widely in many model-based scenarios and proved to be useful Freedman (1981), Politis (2014). Here we demonstrate it only for NW estimator. In this way we can obtain many replicates of $\hat{m}(x)^* - \hat{m}(x)$ (for a test point x), where $\hat{m}(x)^*$ is derived from a bootstrap sample. These replicates can be used to approximate the MSE because $\sqrt{n}[\hat{m}(x)^* - \hat{m}(x)]$ approximates the distribution of $\sqrt{n}[\hat{m}(x) - m(x)]$. In Table 3.6, the statistics for optimal (h, g) using normal approximation(Table 3.2) are listed on the left, while those with least coverage error using bootstrap are on the right. We observe that bootstrap yields increasingly good coverage with shorter width. However, one disadvantage of bootstrap is that it is computationally demanding.

4 | Conclusion and Discussion

To summarize, the application of semi-supervised learning based on 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. When chosen wisely, hybrid estimator outperforms NW estimator significantly. 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. Meanwhile, underlying issues such as bandwidth selection and normal approximation in the proposed estimator are explained briefly.

Acknowledgement

This report has been done under the supervision of Prof. Stephen M.S. LEE, from whom the idea of hybrid estimator originated. I am very grateful for his support and assistance throughout the semester. I also want to express my gratitude to Department of Statistics and Actuarial Science for providing me with the opportunity

of conducting this directed study.

Appendix

Please refer to the following github link to review the code output shown in this project: Rcode for STAT3799 (https://github.com/wjshku/STAT3799)

Bibliography

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

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

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228.

Koehler, E., Brown, E., and Haneuse, S. J.-P. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162.

Lee, D.-H. et al. (2013). 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.

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

Politis, D. N. (2014). Bootstrap confidence intervals in nonparametric regression without an additive model. In *Topics in Nonparametric Statistics*, pages 271–282. Springer.

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

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

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

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