Semi-Supervised Learning based on Kernel Regression

Junshi Wang WJSHKU@CONNECT.HKU.HK
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong

Stephen Lee SMSLEE@HKU.HK
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong

Editor: Not Applicable

Abstract

We study a semi-supervised regression setting where responses are observed for some data points (labelled) but missing for others (unlabelled). We propose an estimator inspired by semi-supervised learning and bias reduction. We start with a potentially biased estimator obtained by nonparametrically regressing the labelled dataset. It is used to generate pseudo-labels for unlabelled data points. A self-supervised estimator is then constructed based on this pseudo-labelled dataset. By exploiting the relationships in biases and convergence rates between these two estimators, we construct a linear mixture of them, yielding a hybrid estimator endowed with a better convergence rate. For the sake of formal theoretical investigation, we focus on the Nadaraya–Watson estimator, a basic form of kernel regression. We describe the mechanism underlying our method and establish its asymptotic properties under general conditions. In particular, under proper tuning of hyperparameters, our proposed method is shown to enjoy improved efficiency, especially under cases of multi-dimensional covariates or cases where a large unlabelled dataset is available. We also show that even an unlabelled dataset of a size smaller than that of the labelled dataset can still be utilised effectively to boost performance, a point seldom mentioned in the literature. We conduct simulation studies to demonstrate the superiority of this novel yet simple method, which is shown to make as much as 30% improvement based on relatively small samples.

Keywords: bias reduction; kernel regression, linear mixing, pseudo-labelling; semi-supervised learning.

1 Introduction

Semi-supervised learning (SSL) leverages both labelled and unlabelled datasets to generate better results. It has gained wide popularity in practice (Chapelle et al., 2009), mainly because it lowers the human cost for collecting labelled data and helps boost the performances of various models. Within this large framework lie various distinct techniques such as co-training, feature extraction and boosting, which utilise unlabelled data in different

ways: see van Engelen and Hoos (2020) for a recent review on semi-supervised learning in greater detail.

Many researchers have contributed to the development of semi-supervised learning, shedding light on ways to make full use of a limited supply of labelled data with the assistance of unlabelled data. Directly related to our work is an approach known as pseudo-labelling. Lee et al. (2013) consider training a neural network in a semi-supervised fashion. Their model is trained simultaneously on both labelled and unlabelled data. The labelled data are utilised to predict the classes of unlabelled data. These predictions, termed pseudo-labels, are treated as if they were observed values. This approach can in principle be applied to any neural network model and any training method. It displays excellent performance in experiments on the MNIST dataset.

SSL has also been extensively studied in statistics. Besides classification, its applications have been extended to conventional regression settings. Wasserman and Lafferty (2007) study SSL theoretically under smoothness and manifold assumptions. They adopt a minimax perspective and highlight a few scenarios under which SSL is found to be helpful. However, their kernel regression method, commonly referred to as a distributional approach, requires a strong assumption on the relationship between the regression function and the density function of the covariate. It is not difficult to see that their method amounts to explicit bias reduction under the assumption of a known leading term of the bias. A recent paper by Azriel et al. (2022) provides new food for thought with the proposal of a new framework for SSL in linear regression, under which a superior estimator could be found when unlabelled data are available.

Besides SSL, other techniques for boosting the performance of regression or classification estimators have been studied, among which linear or convex mixtures of estimators have received extensive attention. Schucany et al. (1971) propose a very simple method for bias reduction related to the "jackknife" via a combination of two estimators derived from the same dataset. Yang (2004) considers a rather different but equally interesting situation where mixing could be used to automatically select the best combination. This idea has also been applied to forecasting by Clemen (1989), who shows that a weighted average of crowd forecasts may outperform a specialist. Lee and Soleymani (2015) develop a simple but general framework under which a hybrid estimator can be tuned to optimally combine estimators converging at different rates.

Inspired by both SSL and linear mixing, we propose a new SSL method for nonparametric regression based on kernel estimation. Consider a labelled dataset $(\mathbf{X}, \mathbf{Y})$ of n independent and identically distributed data points, where $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{R}^d$, and $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$, $y_i \in \mathcal{R}$. We assume, for $i = 1, \ldots, n$, that $y_i = m(x_i) + \epsilon_i$, for some unknown function m and a random error ϵ_i with mean 0. We assume further that an unlabelled dataset of k independent observations is available, denoted by $\mathbf{U} = (u_1, u_2, \ldots, u_k)$, where $u_j \in \mathcal{R}^d$ follows the same distribution as does x_i. Our goal is to estimate $m(x)$ for some fixed $x \in \mathcal{R}^d$, without imposing any parametric structure on m. To start with, we construct a potentially biased estimator $\hat{m}(x)$ by nonparametrically regressing $\mathbf{Y}$ on $\mathbf{X}$. We next predict $m(u_j)$ by $\hat{m}(u_j)$, $j = 1, \ldots, k$, which are treated as pseudo-labels. A self-supervised estimator will be constructed based on this pseudo-labelled dataset. Note that the above

two estimators are correlated, have asymptotic biases satisfying a simple relationship and converge at different rates in general. These properties can potentially be exploited to motivate a linear mixture of the two estimators which converges at a faster rate.

In this paper, we choose as our initial nonparametric regression estimator the Nadaraya–Watson (NW) estimator, an elementary form of kernel regression which facilitates succinct theoretical investigation void of overwhelming notational complexities, while generalising readily to local polynomial regression, The detailed mechanism underlying our proposed SSL approach and the choices of hyperparameters including the bandwidths are discussed in Section 2. We also establish there the asymptotic properties which lend support to superiority of the proposed estimator. Specifically, the proposed method succeeds in improving efficiency significantly, especially under multi-dimensional cases, provided that we tune the hyperparameters properly. In general, the bigger the unlabelled sample size, the more significant the improvement will be. However, our theory suggests that even a small unlabelled dataset of size $k < n$ can still boost the performance, a result rarely found in the literature. Section 3 presents simulation studies to illustrate empirical performance of the proposed estimator in different inference contexts. The results show that our proposed linear combination of the supervised and self-supervised estimators evinces as much as 30% improvement under relatively small samples. Especially noteworthy is the need for customising the optimal choice of bandwidths according to the specific inference problem in hand. In particular, the mean squared error of the estimator and the coverage error of a confidence interval built upon which cannot be optimised by the same choice of bandwidths, a phenomenon which has considerable implications for general statistical practice.

2 Methodology

In this section, we introduce the procedures for constructing our proposed estimator, which is composed of a supervised and a self-supervised components. To facilitate statistical analysis and inference based on the estimator, we prove its asymptotic properties and derive optimal choices of tuning parameters involved in its construction.

In what follows we write $a \succ b$ if $a/b \to \infty$ and $a \asymp b$ if a and b are of the same order. For any sufficiently smooth d-variate real-valued function g, denote by $\mathcal{D}_g(x)$ the vector of first-order partial derivatives of g at $x \in \mathcal{R}^d$ and by $\mathcal{H}_g(x)$ the $d \times d$ Hessian matrix of g at $x = (x_1, \ldots, x_d)$, with its (i, j)-th entry equal to $(\partial^2/\partial x_i \partial x_j)g(x)$. Other notations and details of the proofs are delegated to the Appendix.

2.1 Supervised estimator

Assume that

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $x_1, \ldots, x_n, \epsilon_1, \ldots, \epsilon_n$ are independent, with x_i and ϵ_i following the density functions p and p_0, respectively. Following Wand and Jones (1994), we define, for any fixed $x \in \mathcal{R}^d$,

the NW estimator of $m(x)$ to be $\hat{m}(x) = \hat{\alpha}(x)/\hat{p}(x)$, where

$$\hat{\alpha}(x) = n^{-1}\sum_{i=1}^{n} y_i K_h(x - x_i), \quad \hat{p}(x) = n^{-1}\sum_{i=1}^{n} K_h(x - x_i), \quad K_h(x) = h^{-d}K(x/h),$$

for some d-variate kernel function K and some bandwidth $h > 0$.

We first state two lemmas useful for our theoretical investigation.

Lemma 1 *(Multivariate Taylor's Theorem). Let g be a smooth d-variate real-valued function and α_n be a $d \times 1$ vector with its components all tending to zero as $n \to \infty$. Then, assuming that all entries of $\mathcal{H}_g(x)$ are continuous in a neighbourhood of a fixed $x \in \mathcal{R}^d$, we have*

$$g(x + \alpha_n) = g(x) + \alpha_n^T \mathcal{D}_g(x) + 2^{-1}\alpha_n^T \mathcal{H}_g(x)\alpha_n + o(\alpha_n^T \alpha_n).$$

Lemma 2 *(Lyapunov's Central Limit Theorem for Triangular Arrays). For each $n \geq 1$, assume that the scalar random variables $(z_{1n}, \ldots, z_{nn})$ are independently (but not necessarily identically) distributed with variance $\mathbf{Var}(z_{in}) = \sigma_{in}^2$ and r^{th} absolute central moment $\mathbf{E}[|z_{in} - \mathbf{E}(z_{in})|^r] = \rho_{in} < \infty$ for some $r > 2$ such that*

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

Then $(\sum_{i=1}^n \sigma_{in}^2)^{-1/2} \sum_{i=1}^n \{z_{in} - \mathbf{E}(z_{in})\}$ is asymptotically standard normal.

We hereafter make the following regularity assumptions.

(A1) Each entry of $\mathcal{H}_m(.)$ is piecewise continuous and square integrable.

(A2) The bandwidth $h = h_n$ satisfies $(nh^d)^{-1} + h \to 0$ as $n \to \infty$.

(A3) K is a d-variate kernel function which is bounded, compactly supported, symmetric about 0 and satisfies, for $\gamma \geq 1$,

$$\int K(z)^\gamma dz = \mu_{0,\gamma}(K), \quad \int zz^T K(z)^\gamma dz = \mu_{2,\gamma}(K)I, \quad \mu_{0,1}(K) = 1,$$

where $\mu_{0,\gamma}(K), \mu_{2,\gamma}(K)$ are finite positive constants and I denotes the $d \times d$ identity matrix.

(A4) p_0 has mean 0 and a finite s^{th} moment for some $s \geq 3$.

The proof of asymptotic normality of $\hat{p}(x)$ is standard and available in many works (Nadaraya, 1965; Wand and Jones, 1994; Wied and Weißbach, 2012). The following lemma establishes asymptotic normality of $\hat{\alpha}(x)$.

Lemma 3 $(nh^d)^{1/2}\{\hat{\alpha}(x) - \mathbf{E}[\hat{\alpha}(x)]\}$ *is asymptotically normal with mean 0 and a finite positive variance.*

Proof Let $z_{in} = h^{-d} y_i K((x - X_i)/h)$. Then we have

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

and, for any $r \geq 3$,

$$\rho_{in} = \mathbf{E}[|z_{in} - \mathbf{E}(z_{in})|^r] \leq \mathbf{E}[(|z_{in}| + |\mathbf{E}(z_{in})|)^r] = \sum_{k=0}^{r} C_k^r \mathbf{E}(|z_{in}|^k |\mathbf{E}(z_{in})|^{r-k})$$
$$\leq \sum_{k=0}^{r} C_k^r \mathbf{E}(|z_{in}|^r)^{k/r} \mathbf{E}(|z_{in}|^r)^{(r-k)/r} = 2^r \mathbf{E}(|z_{in}|^r)$$
$$\leq 2^r h^{-rd} \sum_{q=0}^{r} C_q^r \mathbf{E}(\epsilon_i^{r-q}) \mathbf{E}\big(|m(X_i)^q K((x - X_i)/h)^r|\big) = O(h^{-(r-1)d}).$$

It follows by the condition $nh^d \to \infty$ that $(\sum_{i=1}^n \rho_{in})^{1/r} (\sum_{i=1}^n \sigma_{in}^2)^{-1/2} \to 0$. The result then follows from Lemma 2. ■

The following theorem establishes asymptotic normality of $\hat{m}(x)$, with σ_ϵ^2 denoting the variance of ϵ_i.

Theorem 4 *The supervised NW estimator $\hat{m}(x)$ admits the stochastic representation*

$$\hat{m}(x) = m(x) + h^2 E(x) + (nh^d)^{-1/2} p(x)^{-1} \{A(x) - m(x)B(x)\} + O_p(h^4 + (nh^d)^{-1}), \quad (2)$$

where

$$A(x) = (nh^d)^{1/2} \{\hat{\alpha}(x) - \mathbf{E}[\hat{\alpha}(x)]\}, \quad B(x) = (nh^d)^{1/2} \{\hat{p}(x) - \mathbf{E}[\hat{p}(x)]\},$$
$$E(x) = \mathcal{D}_m(x)^T \mathcal{D}_p(x) + 2^{-1} tr(\mathcal{H}_m(x)) p(x) \mu_{2,1}(K),$$

and $(nh^d)^{1/2}\{\hat{m}(x) - m(x) - h^2 E(x)\}$ is asymptotically normal with mean 0 and variance $\sigma_\epsilon^2 p(x)^{-1} \mu_{0,2}(K)$.

Proof The stochastic representation follows by Taylor expanding

$$\hat{m}(x) = \frac{(nh^d)^{-1/2} A(x) + \mathbf{E}[\hat{\alpha}(x)]}{(nh^d)^{-1/2} B(x) + \mathbf{E}[\hat{p}(x)]} \quad (3)$$

and noting that $A(x), B(x) \asymp 1$,

$$\mathbf{E}[\hat{\alpha}(x)] = m(x)p(x) + h^2 \mu_{2,1}(K) \big\{2^{-1} m(x) tr(\mathcal{H}_p(x)) + \mathcal{D}_m(x)^T \mathcal{D}_p(x) + 2^{-1} p(x) tr(\mathcal{H}_m(x))\big\}$$
$$+ O(h^4) = m(x)p(x) + O(h^2),$$
$$\mathbf{E}[\hat{p}(x)] = p(x) + h^2 \mu_{2,1}(K) 2^{-1} tr(\mathcal{H}_p(x)) + O(h^4) = p(x) + O(h^2).$$

It suffices to prove that $A(x) - m(x)B(x)$ is asymptotically normal. Note that

$$\hat{\alpha}(x) - m(x)\hat{p}(x) = (nh^d)^{-1} \sum_{i=1}^{n} \{m(x_i) - m(x) + \epsilon_i\} K((x - x_i)/h),$$

which shares the same form as $\hat{\alpha}(x)$ and is therefore asymptotically normal according to Lemma 3. The asymptotic variance can be derived using standard arguments based on Taylor expansions. ■

It is clear that minimisation of $\hat{m}(x) - m(x)$ yields an optimal error of order $n^{-2/(d+4)}$, attained by setting the bandwidth $h \asymp n^{-1/(d+4)}$.

2.2 Self-supervised estimator

We now move on to establish the arguments for the estimator derived from the unlabelled dataset $\mathbf{U}$, using notations similar to those for the supervised NW estimator. Analogous to (A2), we assume the following condition on the bandwidth g with respect to the unlabelled sample size k.

(A5) The bandwidth $g = g_n$ satisfies $(kg^d)^{-1} + g \to 0$ as $k \to \infty$.

Define the self-supervised NW estimator by $\hat{r}(x) = \hat{\beta}(x)/\hat{q}(x)$, where

$$\hat{\beta}(x) = (kg^d)^{-1} \sum_{i=1}^{k} \hat{m}(u_i) K((x - u_i)/g), \quad \hat{q}(x) = (kg^d)^{-1} \sum_{i=1}^{k} K((x - u_i)/g).$$

Define

$$\beta(x) = (kg^d)^{-1} \sum_{i=1}^{k} m(u_i) K((x - u_i)/g),$$
$$C(x) = (kg^d)^{1/2}\{\beta(x) - \mathbf{E}(\beta(x))\}, \quad D(x) = (kg^d)^{1/2}\{\hat{q}(x) - \mathbf{E}(\hat{q}(x))\}.$$

Like $(A(x), B(x))$, asymptotic normality holds for $(C(x), D(x))$, which does not depend on the labelled dataset $(\mathbf{X}, \mathbf{Y})$ and is therefore independent of $(A(x), B(x))$.

Following similar arguments for proving Lemma 3, we have, as $k \to \infty$,

$$g^{-2}\mathbf{Var}(C(x) - m(x)D(x)) \to \mu_{2,2}(K) p(x) \mathcal{D}_m(x)^T \mathcal{D}_m(x) \text{ in probability.} \tag{4}$$

Thus, $g^{-1}\{C(x) - m(x)D(x)\}$ has an asymptotically normal distribution with mean 0 and variance $\mu_{2,2}(K) p(x) \mathcal{D}_m(x)^T \mathcal{D}_m(x)$, independent of $(\mathbf{X}, \mathbf{Y})$. Similar to Theorem 4, we may now provide a stochastic representation of $\hat{r}(x)$,

$$\begin{aligned}\hat{r}(x) = m(x) &+ (h^2 + g^2)E(x) + (kg^d)^{-1/2} p(x)^{-1}\{C(x) - m(x)D(x)\} \\ &+ (nh^d)^{-1/2} p(x)^{-1} \int K(z)\{A(x - gz) - m(x - gz)B(x - gz)\}\, dz\{1 + o_p(1)\} \\ &+ O_p(h^4 + g^4 + g(kg^d)^{-1}),\end{aligned} \tag{5}$$

with $\max\{g/h, 1\}^{d/2} \int K(z)\{A(x - gz) - m(x - gz)B(x - gz)\, dz$ converging in distribution to a normal variable with mean 0 and variance

$$\sigma_\epsilon^2 p(x) \int \Big(\int K(u + z \min\{h, g\}/\max\{h, g\}) K(z)\, dz \Big)^2 du,$$

independent of $g^{-1}\{C(x) - m(x)D(x)\}$. Note that the above asymptotic variance reduces to $\sigma_\epsilon^2 p(x) \mu_{0,2}(K)$ if either $h \prec g$ or $g \prec h$.

2.3 Hybrid estimator

Consider a hybrid estimator based on a linear combination of $\hat{m}(x)$ and $\hat{r}(x)$, defined to be

$$\hat{y}_c(x) = \lambda \hat{m}(x) + (1 - \lambda)\hat{r}(x),$$

for some constant $\lambda \in \mathcal{R}$ to be specified. Combining (2) and (5), $\hat{y}_c(x)$ admits an expansion

$$\begin{aligned}\hat{y}_c(x) = m(x) &+ \{h^2 + (1-\lambda)g^2\}E(x) + (1-\lambda)(kg^d)^{-1/2}p(x)^{-1}\{C(x) - m(x)D(x)\} \\ &+ \lambda(nh^d)^{-1/2}p(x)^{-1}\{A(x) - m(x)B(x)\} \\ &+ (1-\lambda)(nh^d)^{-1/2}p(x)^{-1}\int K(z)\{A(x-gz) - m(x-gz)B(x-gz)\}\,dz\{1+o_p(1)\} \\ &+ \lambda O_p(h^4 + (nh^d)^{-1}) + (1-\lambda)O_p(h^4 + g^4 + g(kg^d)^{-1}).\end{aligned}$$

To establish the asymptotic properties of $\hat{y}_c(x)$, we provide first an expansion for the covariance between the two correlated stochastic components in the expansion for $\hat{y}_c(x)$:

$$\begin{aligned}&\mathbf{Cov}\Big(A(x) - m(x)B(x),\ \int K(z)\{A(x-gz) - m(x-gz)B(x-gz)\}\,dz\Big) \\ &= \min\{h/g, 1\}^d \sigma_\epsilon^2 p(x)\int K(u)K(v)K\Big(\frac{\min\{h,g\}}{\max\{h,g\}}\,u + \frac{h}{\max\{h,g\}}\,v\Big)\,du\,dv\{1+o(1)\}.\end{aligned}$$

We set $\lambda = 1 + (h/g)^2$ to eliminate the bias term $\{h^2 + (1-\lambda)g^2\}E$. Then the expression of $\hat{y}_c(x)$ is further reduced to

$$\begin{aligned}\hat{y}_c(x) = m(x) &- (h/g)^2(kg^d)^{-1/2}p(x)^{-1}\{C(x) - m(x)D(x)\} \\ &+ \{1 + (h/g)^2\}(nh^d)^{-1/2}p(x)^{-1}\{A(x) - m(x)B(x)\} \\ &- (h/g)^2(nh^d)^{-1/2}p(x)^{-1}\int K(z)\{A(x-gz) - m(x-gz)B(x-gz)\}\,dz\{1+o_p(1)\} \\ &+ O_p(h^4 + (nh^d)^{-1} + h^6g^{-2} + (h/g)^2(nh^d)^{-1} + (hg)^2 + (h/g)^2g(kg^d)^{-1}).\end{aligned}$$

Minimisation of the order of the estimation error

$$\hat{y}_c(x) - m(x) \asymp (h^2/g)(kg^d)^{-1/2} + (h/g)^2(nh^d)^{-1/2} + (nh^d)^{-1/2} + (hg)^2 + h^6g^{-2}$$

enables us to derive the optimal orders of the bandwidths (h, g) and the asymptotic distribution of $\hat{y}_c(x)$ constructed using the optimal bandwidths.

1. For $k \prec n^{(d+6)/(d+8)}$, $\hat{y}_c(x) - m(x)$ has minimum order $n^{-2/(d+4)}k^{-2d/\{(d+4)(d+6)\}}$, attained by setting, for some fixed constants $C_h, C_g > 0$,

   $$h = C_h n^{-1/(d+4)}k^{4/\{(d+4)(d+6)\}}, \quad g = C_g k^{-1/(d+6)},$$

   based on which $(nh^d)^{1/2}\{\hat{y}_c(x) - m(x)\}$ has an asymptotically normal distribution with mean $\asymp 1$ and variance

   $$p(x)^{-1}\big\{\sigma_\epsilon^2\mu_{0,2}(K) + C_h^{d+4}C_g^{-d-2}\mu_{2,2}(K)\mathcal{D}_m(x)^T\mathcal{D}_m(x)\big\}.$$

2. For $k \asymp n^{(d+6)/(d+8)}$ or $k \succ n^{(d+6)/(d+8)}$, $\hat{y}_c(x) - m(x)$ has minimum order $n^{-4/(d+8)}$, attained by setting, for some fixed constant $\varrho > 0$,

   $$h = \varrho g \asymp n^{-1/(d+8)},$$

in which case $(nh^d)^{1/2}\{\hat{y}_c(x) - m(x)\}$ has an asymptotically normal distribution with mean $\asymp 1$ and variance

$$(1+\varrho^2)^2\sigma_\epsilon^2 p(x)^{-1}\mu_{0,2}(K)$$
$$+ \varrho^2 \min\{\varrho,1\}^d \sigma_\epsilon^2 p(x)^{-1}\Big[\varrho^2 \int \Big(\int K(u + z\min\{\varrho,1\}/\max\{\varrho,1\})K(z)\,dz\Big)^2 du.$$
$$- 2(1+\varrho^2)\int K(u)K(v)K\Big(\frac{\min\{\varrho,1\}}{\max\{\varrho,1\}}\,u + \frac{\varrho}{\max\{\varrho,1\}}\,v\Big)\,du\,dv\Big].$$

In the special case where $\varrho = 1$, the above asymptotic variance reduces to

$$\sigma_\epsilon^2 p(x)^{-1}\int \Big\{2K(u) - \int K(u+v)K(v)\,dv\Big\}^2 du.$$

In both cases (1) and (2) above, under-smoothing the optimal h to a slightly smaller order renders the bias asymptotically negligible and $(nh^d)^{1/2}\{\hat{y}_c(x) - m(x)\}$ asymptotically normal with mean 0 and variance $\sigma_\epsilon^2 p(x)^{-1}\mu_{0,2}(K)$. Inference about $m(x)$ may then be conducted by means of normal approximation, with the variance estimated by plugging in $\hat{p}(x)$ and $n^{-1}\sum_{i=1}^n\{y_i - \hat{y}_c(x_i)\}^2$ to replace $p(x)$ and σ_ϵ^2, respectively, in the expression for the asymptotic variance.

A comparison of the optimal root mean squared error (RMSE) of $\hat{y}_c(x)$ with that of the supervised estimator $\hat{m}(x)$ reveals the level of improvement afforded by our hybrid construction. It follows from the previous results that

$$\frac{\inf_{h,g>0}\mathrm{RMSE}(\hat{y}_c(x))}{\inf_{h>0}\mathrm{RMSE}(\hat{m}(x))} \asymp \begin{cases} k^{-2d/\{(d+4)(d+6)\}}, & k \prec n^{(d+6)/(d+8)}, \\ n^{-2d/\{(d+4)(d+8)\}}, & k \asymp n^{(d+6)/(d+8)} \text{ or } k \succ n^{(d+6)/(d+8)}. \end{cases}$$

We see that $\hat{y}_c(x)$ significantly improves upon $\hat{m}(x)$ in having an optimal RMSE of a smaller order. For a fixed supervised sample size n, increasing the unsupervised sample size k accelerates the reduction in the optimal RMSE until k reaches the order $n^{(d+6)/(d+8)}$. The reduction remains stable as k increases beyond the latter order.

2.4 Coverage Error

Besides mean square error, coverage error as an important indicator in statistical inference is of interest. For simplicity, we limit the focus to $d = 1$. Let r_n denote the normalizing constant. We have $\hat{y_c}(x)$ as an asymptotic normal random variable which can be used to obtain the confidence interval for a point estimation.

$$Z_n = r_n(\hat{y_c}(x) - m(x)) \overset{approx}{\sim} \mathbf{N}(b, \sigma^2)$$

In general, $r_n^2 = MSE = O(h_n^8 + \frac{1}{nh_n} + h_n^4 g_m^4 + \frac{h_n^4}{mg_m^3})$. But the concept of confidence interval only makes sense when $\sigma^2 > 0$. As we can see, bias term is very complicated so hard to estimate. Therefore, we propose a naive estimator $\hat{b} = 0$.

To start with, we focus on the case where $b = o(1)$. As we can observe from the expression of MSE, $b = o(1)$ if $\frac{h_n^4}{mg_m^3} \succ h_n^8 + \frac{1}{nh_n} + h_n^4 g_m^4$. In this case, we let $r_n^2 = \frac{mg_m^3}{h_n^4}$.

We can estimate the coverage error by the Berry-Esseen theorem directly via conditioning on (X, Y). Considering writing the self-supervised estimator in terms of a triangular array $\{\xi_{j,m}\}_{j=1}^{n}$:

$$\begin{aligned}\frac{C-\hat{m}(x)D}{q(x)g_m} &= (mg_m^3q^2(x))^{-1/2}\sum_{j=1}^{m}[\hat{m}(u_j)K(\frac{x-u_j}{g_m}) - \hat{m}(x)K(\frac{x-u_j}{g_m}) \\ &+ \mathbf{E}(\hat{m}(x)K(\frac{x-u_j}{g_m})) - \mathbf{E}(\hat{m}(u_j)K(\frac{x-u_j}{g_m})|X,Y)] \\ &= \frac{1}{m}\sum_{j=1}^{m}\xi_{j,m}\end{aligned}$$

It is possible to show that

$$\begin{aligned}\mathbf{E}[|\xi_{j,m}|^2] &= \Theta(\hat{m}'(x)^2 m g_m^2 \frac{\sigma_k^2}{q(x)}) \\ \mathbf{E}[|\xi_{j,m}|^3] &= O(\hat{m}'(x)^3 m^{3/2} g_m^{5/2})\end{aligned}$$

Then apply the Berry Esseen's Theorem (details available in Appendix E.1).

$$\mathbf{P}\{\frac{r_n(\hat{y}_c - m(x)) - \hat{b}}{\hat{\sigma^2}} \leq z_\alpha\} = \Phi(z_\alpha) + O(\delta + \delta_b + \delta_\sigma + \delta_{CD} + \frac{(mg_m^3)^{1/2}}{(nh_n^5)^{1/2}} + m^{-1/2}g_m^{-1/2})$$

Typically $\delta_p = h_n^2 + (nh_n)^{-1/2}$, $\delta_q = g_m^2 + (mg_m)^{-1/2}$ and δ_ϵ should be greater than $n^{-1/2}$. Finally, the coverage error is

$$Error_{coverage} = O((mg_m^3)^{1/2}(h_n^2 + g_m^2 + (nh_n^5)^{-1/2}) + (mg_m)^{-1/2} + g_m)$$

In general, the true bias does not decay to zero $b = \Theta(1)$, it is fairly difficult to estimate its value and result in $\delta_b = 1$. Hence this case is trivial.

$$\mathbf{P}\{\frac{r_n(\hat{y}_c - m(x)) - \hat{b}}{\hat{\sigma^2}} \leq z_\alpha\} - \Phi(z_\alpha) \approx \Phi(z_\alpha + C) - \Phi(z_\alpha)$$

where C is some non-trivial constant. All in all, the coverage error $Error_{coverage}$ can be $o(1)$ only if $\frac{h^4}{mg_m^3} \succ h_n^8 + \frac{1}{nh_n} + h_n^4 g_m^4$.

$$Error_{coverage} = O((mg_m^3)^{1/2}(h_n^2 + g_m^2 + (nh_n^5)^{-1/2}) + (mg_m)^{-1/2} + g_m)$$

Recall that the mean square error(MSE) can be expressed as,

$$MSE = O(h_n^8 + \frac{1}{nh_n} + h_n^4 g_m^4 + \frac{h_n^4}{mg_m^3})$$

Therefore, it's not hard to see the coverage error will always be $\Theta(1)$ when MSE is minimized. This implies that it is impossible to achieve the best coverage error and mean square error at the same time. However, some compromises could be made to strike a balance.

If we optimize the coverage error given by Equation 1.5 regardless how large MSE is going to be, when the condition in 1.2.1 is fulfilled, namely $b = o(1)$, we have the following result.

If $m^{-1/6} \succ n^{-1/9}$, the minimax choice should be

$$Error_{coverage} = O(m^{-1/3}), MSE = O(m^{-1/6}), h = n^{-1/9}, g = m^{-1/3}$$

If $m^{-1/6} \prec n^{-1/9}$, it should be,

$$Error_{coverage} = O(m^{-1/4}n^{-1/18}), MSE = O(m^{1/2}n^{-7/9}), h = n^{-1/9}, g = m^{-1/2}n^{1/9}$$

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 (15) and (C.1) will be evaluated as well through the length and coverage probability of the confidence interval.

3 Experiment and Simulation

Compared with NW estimator which have optimal MSE $= O(n^{-4/5})$ when $h_n \propto n^{-1/5}$, optimal MSE of hybrid estimator given by equation 15 will always have higher convergence rate. Its confidence interval based on the normal distribution approximation, however, is not that clear. 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.

$$\begin{aligned}
\textbf{Statistical Model} &: m(x) = x^2, X_i \overset{iid}{\sim} \mathbf{N}(0, \sigma_x^2), \sigma_x = 1, \epsilon_i \overset{iid}{\sim} \mathbf{N}(0, \sigma_\epsilon^2), \sigma_\epsilon = 1 \\
\textbf{Choice of size} &: n = 32, 64 \text{ or } 128, m = n^{10/19} \\
\textbf{Testing data point} &: x = 1.5, m(x) = 1.5^2 \\
\textbf{Choice of } \mathbf{h}, \mathbf{g} &: h = 0.1, 0.15, ..., 1, \text{and } g = 0.3, 0.6, ..., 2.4
\end{aligned}$$

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.

3.1.1 Mean Square Error

The results of grid search is summarized in Table 1. Although the optimal MSE varies across choices of n, we can see that hybrid estimator performs better than NW estimator.

n	m	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 1: Grid Search Results of Mean Square Error. With each pair of (n, m), we find the optimal bandwidth for NW(columns 3-4) and HY(columns 5-7) estimator. Standard deviations of MSE are approximated via Monte Carlo error.

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 1, 2 and 3. On the left, we fix the optimal g found through grid search and plot the MSE against h to contrast two estimators better. For every h we selected, the hybrid estimator (blue curve) performs better than or just as well with the other, although with small sample size, the improvement is not very impressive when both are well tuned. Note that we selected less h for hybrid estimator because it's very computationally heavy task. It reveals that the optimal choice of h for HY estimator is larger but very close to that of NW estimator, and results in a smaller minimal MSE. This agrees with the formula and expectation. On the right, we display the heatmap which shows vividly how MSE of hybrid estimator changes with h and g. It is clear that choice of h is more impactful to MSE while g serves more as a step of fine tuning. This suggests some practical methods when searching for optimal h and g.

3.1.2 Confidence Interval

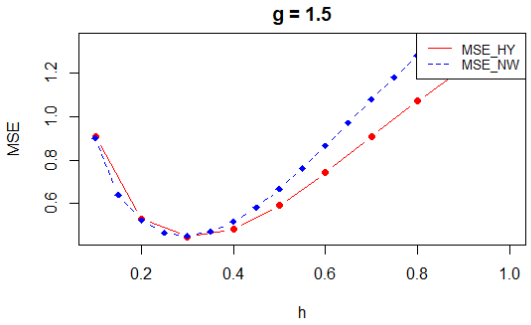
Besides MSE which is a common criterion for point estimation, confidence interval is a crucial subject of interest in statistical inference as well. Making use of equation (15), we can then construct CI based on several reasonable assumptions. Firstly, since p(x) and q(x) are generally identical in most 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. With this estimation, we could attain construct 0.95 level confidence level for the point of interest.

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

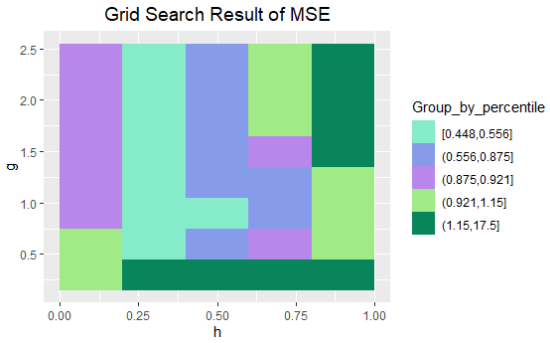
In Table 2, 3 and 4, some typical pairs of (h, g) are chosen for comparison. In each table, on the left are choices of h that have the best coverage probability and on the right are those with minimal mean square error.

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

Figure 1: Mean Square Error: n = 32

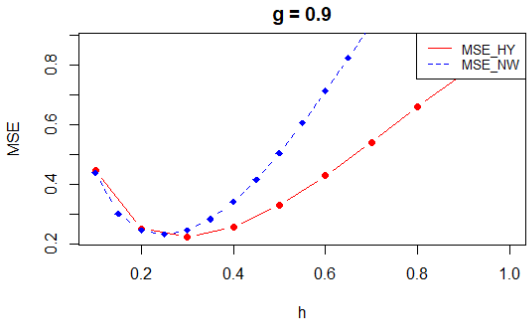


(a) MSE with g fixed at 1.5

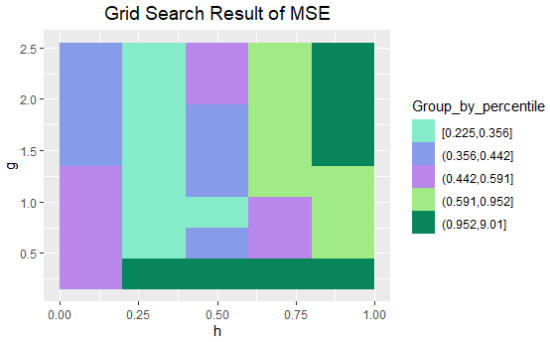


(b) Heatmap depicting MSE by percentile

Figure 2: Mean Square Error: n = 64

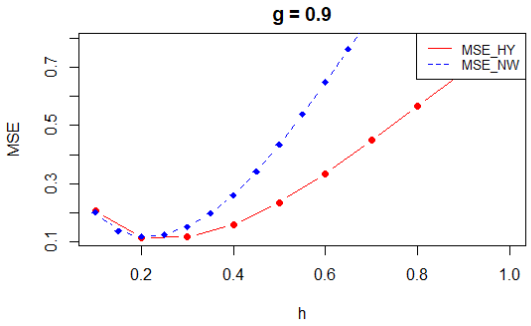


(a) MSE with g fixed at 0.9

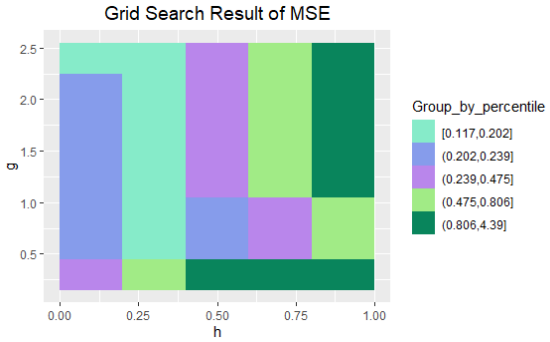


(b) Heatmap depicting MSE by percentile

Figure 3: Mean Square Error: n = 128



(a) MSE with g fixed at 1.5



(b) Heatmap depicting MSE by percentile

n	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 2: Coverage probability and width of interval based on NW estimator. Coverage here refers to the coverage probability. The true coverage is set by default to 0.95. The hs on the right are optimal, the same as in Table 1. That on the left are the optimal hs for coverage.

n	m	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: Coverage probability and width of interval based on Hybrid estimator with error variance σ_ϵ^2 known. The true coverage is set by default to 0.95. The hs and gs on the right are optimal for MSE, the same as in Table 1. That on the left are the optimal choice for coverage.

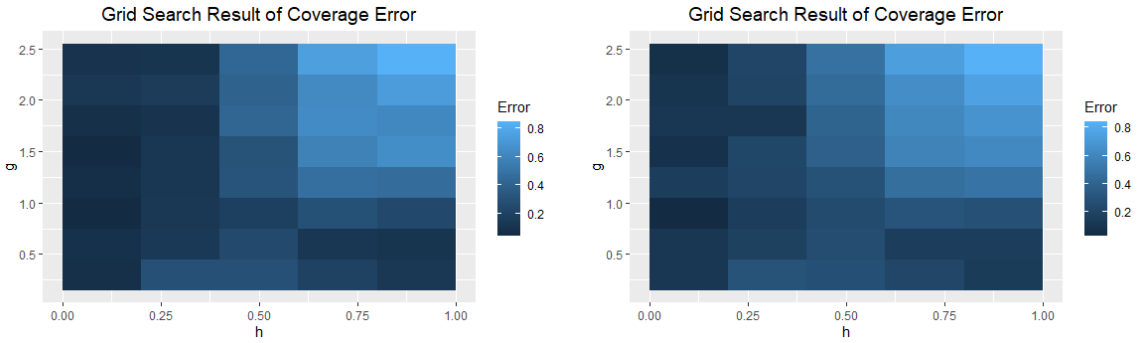
the variance of hybrid estimator, resulting in short width and lower coverage probability. And the coverage does not increase with n but stays at around 0.83. This suggests we should use normal approximation with extra caution. Carefully examining the heatmap (Figure 4b, 4d and 4f) depicting how error(difference between 0.95 and coverage probability) changes with h and g. 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 reiterates the importance of bandwidth.

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, just as discussed in section 2.4. Based on Figure 1b, 2b, 3b and 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 whose research focus is on confidence interval are suggested to construct one using other pairs of (h, g).

3.2 Performance of Hybrid Estimator with multiple dimensions

In this section, we demonstrate how our method could be beneficial with higher dimension data through simulation results. Compared to one-dimensional setting, the situation is slightly more complicated. First, we need to choose a particular kernel. According to Wand (1994; Kernel Smoothing), there are two common techniques for generating multi-

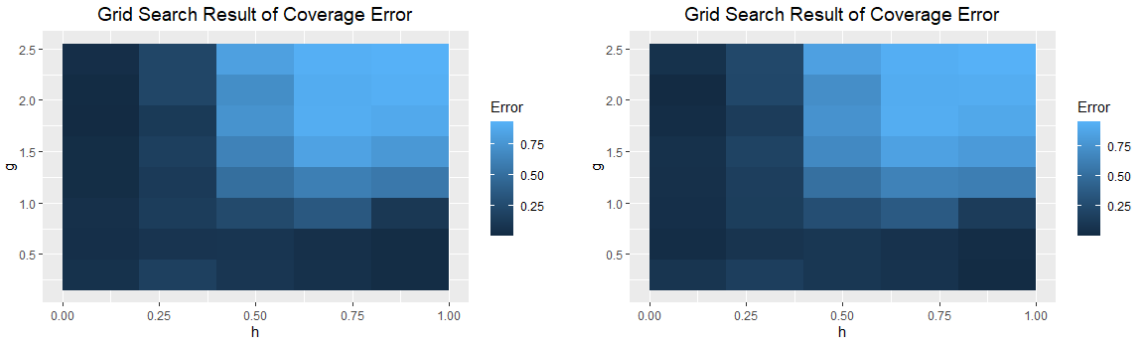


(a) n = 32(σ^2 given)

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

n	m	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 4: Coverage probability and width of interval based on Hybrid estimator with error variance σ_ϵ^2 unknown. The true coverage is set by default to 0.95. The hs and gs on the right are optimal for MSE, the same as in Table 1. That on the left are the optimal choice for coverage.

variate kernels from a symmetric univariate kernel k, which are product kernel and radially symmetric kernel. Here we choose product kernel as an example

$$K^D(x) = \prod_{i=1}^{d} k(x_i)$$

, where k is the uniform kernel.

Here we have chosen the following parameters,

$$\textbf{Statistical Model} : m(x) = \prod^{d} x_i, X_i \overset{iid}{\sim} \mathbf{N}(0, \sigma_x^2), \sigma_x = 2, \epsilon_i \overset{iid}{\sim} \mathbf{N}(0, \sigma_\epsilon^2), \sigma_\epsilon = 1$$
$$\textbf{Choice of size} : n = 32, 64, 128 \text{ or } 256, m = n^{15/19}$$
$$\textbf{Testing data point} : x = (1,2), (1,2,3), (-1,1,2,3) \text{ respectively for } d = 2, 3, 4$$

In Table 5 and Table 6, we present the grid search result for the optimal (h, g). We can observe a clear decreasing trend as n increases and an increasing one as d increases. These fits our expectation intuitively. If the dimension increases, then given the same sample size, the target function is in general harder to estimate and wider bandwidth is needed to give a rough understanding. Besides just the bandwidth, the ratio of optimal MSE of hybrid estimator divided by that of NW estimator is also included. In general, the improvement is quite significant, except for very small sample size. It is noticeable that the hybrid estimator gives more than 30 percent efficiency gain when $n = 256, d = 4$. The last point to note is that, interestingly, the ratio becomes smaller with the same (n, m) as dimension increases from 2 to 4. This corresponds to the theoretical result well.

4 Conclusion

We proposed a simple new method and demonstrated that an unknown function m could be better estimated with this framework. With a suboptimal estimator $\hat{m}(x)$ and extra unlabelled data u_j, the first trick is to construct a self-supervised estimator. This estimator is motivated by pseudo-labelling but fundamentally different. We aim to tackle a complicated regression problem but not a classification problem. We want to derive the statistical properties such as mean square error and asymptotic distribution of such estimator. Therefore, we have chosen to regress completely on the unlabelled data which may

	Optimal h: NW			Optimal h: HY		
n	$d=2$	$d=3$	$d=4$	$d=2$	$d=3$	$d=4$
32	1.4	2.0	1.5	1.5	2.2	1.5
64	1.2	1.5	1.4	1.3	1.9	1.8
128	1.0	1.5	1.3	1.1	1.6	1.6
256	0.9	1.2	1.1	1.0	1.4	1.6

Table 5: Optimal Bandwidth for NW and HY estimator. NW means Nadaraya estiamtor which is supervised. HY stands for the proposed hybrid estiamtor. This table contrasts the value of h for different estimator. Clearly, that of hybrid estimator is larger.

n	Optimal h: HY			g			Ratio		
	$d=2$	$d=3$	$d=4$	$d=2$	$d=3$	$d=4$	$d=2$	$d=3$	$d=4$
32	1.5	2.2	1.5	7	6	8	1	1	1
64	1.3	1.9	1.8	4	3	1.6	0.96	0.91	0.96
128	1.1	1.6	1.6	3	3	0.9	0.91	0.84	0.79
256	1.0	1.4	1.6	3.0	2.5	0.7	0.82	0.77	0.68

Table 6: Optimal Bandwidth and Efficiency Gain via Hybrid Estimator. Ratio refers to the ratio between MSE_HY and MSE_NW, describing the efficiency of hybrid estimator. As sample size and dimension grows, efficiency increases gradually.

sound counterintuitive. However, the advantage is that we could observe the contribution of unlabelled data more directly and clearly. Mixing unlabelled data with labelled data would introduce greater technicality. By the method of conditioning, one successfully reveals the relationship between supervised and self-supervised estimator.

The second trick is bias reduction in the form of $\lambda\hat{m}(x) + (1-\lambda)\hat{r}(x)$. We note that $\lambda = 1 + \frac{E \cdot h_n^2}{F \cdot g_m^2}$ is normally greater than 1. This means self-supervised estimator reinforced the bias existing in supervised estimator. We took advantage of this difference in bias to construct a less biased estimator.

From our simulation result, a few interesting discoveries should be noted. Conventional SSL as in van Engelen and Hoos (2020) tend to use an unlabelled dataset that is much bigger than the labelled one. We have shown that this is not necessary. Even with a size $m \prec n$, the proposed estimator can have a reasonably faster convergence rate. More importantly, we pointed out that the contribution of unlabelled dataset is limited. As m grows beyond $n^{\frac{d+6}{d+8}}$, the minimax rate of convergence will be independent of m. This is intuitive as marginal gain from unlabelled data will be converge to zero. It is instructive for practitioners to weigh the gain from unlabelled data and loss from computation time. Furthermore, we pointed out how a single bandwidth fail to carry out two jobs at the same time, a point receiving little attention. It is indeed quite common for one to use a hyperparameter selected for minimizing mean square error on confidence interval. But it is clear in our study that this should not be taken for granted.

While this work is limited to consider only basic kernel regression, a more general framework was introduced by Lee and Soleymani (2015). We have demonstrated a simple but efficient method for improving performance of estimator. This method in practise might face problems like bandwidth selection and is yet to be studied in detail. One possible solution and direction of research is to choose the best bandwidth with bootstrap. It could be easily extended to higher dimensions and applied to other regression such as local polynomial regression as well.

Acknowledgments and Disclosure of Funding

All acknowledgements go at the end of the paper before appendices and references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found on the JMLR website.

Appendix A. Supervised

A.1 Notation

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

$$B = (nh_n)^{1/2}[\hat{p}(x) - \mathbf{E}[\hat{p}(x)]] \tag{7}$$

$$E = [m'(x)p'(x)\mu_2(k) + \frac{1}{2}m''(x)p(x)\mu_2(k)]\frac{1}{p(x)} + O(h_n^2) \tag{8}$$

A.2 Taylor Expansion

The first derivative of g 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

$$\hat{m}(x) - m(x) = Eh_n^2 + (nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} + O_p((nh_n)^{-1} + (nh_n)^{-1/2}h_n^2) \tag{9}$$

Next, we present $\mathbf{Cov}(A, B)$ and $\mathbf{Var}(A - m(x)B)$,

$$\begin{aligned} \mathbf{Cov}(A, B) = [m(x)p(x)r(k) + \frac{1}{2}m(x)p''(x)\sigma_k^2 h_n^2 + m'(x)p'(x)\sigma_k^2 h_n^2 \\ + m''(x)p(x)\sigma_k^2 h_n^2 + O(h_n^4)] - h_n \mathbf{E}(\hat{\alpha}(x))\mathbf{E}(\hat{p}(x)) \end{aligned} \tag{10}$$

$$\begin{aligned} \mathbf{Var}(A - m(x)B) &= \mathbf{Var}(A) + \mathbf{Var}(m(x)B) - 2\mathbf{Cov}(A, m(x)B) \\ &= \sigma_\epsilon^2 p(x)r(k) + O(h_n^2) \end{aligned} \tag{11}$$

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

Appendix B. Self-Supervised

B.1 Notation

$$C = (mg_m)^{1/2}[\hat{\beta}(x) - \mathbf{E}[\hat{\beta}(x)|\mathbf{X}, \mathbf{Y}]] \tag{12}$$

$$D = (mg_m)^{1/2}[\hat{q}(x) - \mathbf{E}[\hat{q}(x))]] \tag{13}$$

$$F = [m'(x)q'(x)\mu_2(k) + \frac{1}{2}m''(x)q(x)\mu_2(k)]\frac{1}{q(x)} + O_p(\frac{1}{\sqrt{nh_n^5}} + g_m^2) \tag{14}$$

B.2 Asymptotic independence

$$\lim_{n\to\infty} F(C|\mathbf{X}, \mathbf{Y}) \to \Phi(c/\sigma_c')$$

$$\lim_{n\to\infty, m\to\infty} F(C|\mathbf{X}, \mathbf{Y}) \to \Phi(c/\sigma_\infty)$$

$$\sigma_c^{'2} = m^2(x)q(x)\mu_2(k) + O_p(g_m)$$

$$\sigma_\infty^2 = m^2(x)q(x)\mu_2(k)$$

B.3 Theorem3

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 $\frac{(\Sigma_{i=1}^m \rho_{im})^{1/r}}{(\Sigma_{i=1}^m \sigma_{im}^2)^2} \to 0$ and the result follows.

B.4

Here are calculations of $\hat{r}(x)$

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

where I and J are both constants of order 1.

Appendix C. Hybrid

C.1 Expectation and Variance of hybrid estimator

$$\begin{aligned}\mathbf{Var}(\hat{y_c}(x) - m(x)) &= \mathbf{Var}((nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} - \frac{h_n^2(mg_m)^{-1/2}}{g_m^2}g_m\frac{E}{H}\frac{C - \hat{m}(x)D}{q(x)g_m}) \\ &\approx \frac{1}{nh_n}\frac{1}{p^2(x)}\mathbf{Var}(A - m(x)B) + \frac{h_n^4}{mg_m^3}\frac{(E/H)^2}{q(x)^2}\mathbf{Var}(\frac{C - \hat{m}(x)D}{g_m}|\mathbf{X}, \mathbf{Y}) \\ &= \frac{1}{nh_n}\frac{1}{\hat{p}(x)}\hat{\sigma_\epsilon^2}r(k) + \frac{h_n^4}{mg_m^3}\frac{E^2}{H^2\hat{q}(x)}\hat{m}'(x)^2\sigma_k^2\end{aligned} \tag{15}$$

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

Appendix D. General Condition

D.1 Notation

$$
\begin{aligned}
A &= (n|H|^{1/2})^{1/2}[\hat{\alpha}(x) - \mathbf{E}[\hat{\alpha}(x)]] \\
B &= (n|H|^{1/2})^{1/2}[\hat{p}(x) - \mathbf{E}[\hat{p}(x)]] \\
C &= (m|G|^{1/2})^{1/2}[\hat{\beta}(x) - E(\hat{\beta}(x)|X,Y)] \\
D &= (m|G|^{1/2})^{1/2}[\hat{q}(x) - E(\hat{q}(x)|X,Y)] = (m|G|^{1/2})^{1/2}[\hat{q}(x) - E(\hat{q}(x))] \\
H &= (h_{(1)},...,h_{(d)}) = (h_n,...,h_n) \\
G &= (g_{(1)},...,g_{(d)}) = (g_m,...,g_m)
\end{aligned}
$$

D.2 Moments

$$
\begin{aligned}
E[\hat{\alpha}(x)] &= m(x)p(x) + \{\frac{1}{2}m(x)tr(H \cdot H_p(x)) + tr(H \cdot D_m(x) \cdot D_p(x)) \\
&\quad + \frac{1}{2}p(x)tr(H \cdot H_m(x))\}\mu_2(k) + o(tr(H)) = m(x)p(x) + O(tr(H)) \\
E[\hat{p}(x)] &= p(x) + \frac{1}{2}tr(H \cdot H_p(x))\mu_2(k) + o(tr(H)) = p(x) + O(tr(H)) \\
Var[\hat{\alpha}(x)] &= Var[\hat{p}(x)] = O(n^{-1}|H|^{-1/2})
\end{aligned}
$$

D.3 Derivative Estimation

D.4 Expression and Optimization of Hybrid Estimator

Moving on to the hybrid estimator, we have the following result.

$$
\begin{aligned}
\hat{y}_c(x) &= \lambda\hat{m}(x) + (1-\lambda)\hat{r}(x) \\
\hat{m}(x) - m(x) &= Eh_n^2 + (nh_n^d)^{-1/2}\frac{A - m(x)B}{p(x)} + O_p(h_n^4 + \frac{1}{nh_n^d}) \\
\hat{r}(x) - \hat{m}(x) &= Fg_m^2 + (mg_m^d)^{-1/2}g_m\frac{C - \hat{m}(x)D}{q(x)g_m} + O_p(g_m^4 + g_m^2h_n^2 + \frac{g_m}{mg_m^d} + \frac{g_m^2}{\sqrt{nh_n^{d+4}}})
\end{aligned}
$$

where, E, F are the following coefficients,

$$
\begin{aligned}
E &= tr(D_m(x)D_p(x) + \frac{1}{2}H_m(x)p(x)\mu_2(k)) \\
F &= tr(D_m(x)D_q(x) + \frac{1}{2}H_m(x)q(x)\mu_2(k))
\end{aligned}
$$

$$
\begin{aligned}
\hat{y_c}(x) &= \lambda\hat{m}(x) + (1-\lambda)\hat{r}(x) \\
&= (1-\lambda)[\hat{r}(x) - \hat{m}(x)] + [\hat{m}(x) - m(x)] + m(x) \\
&= m(x) + [Eh_n^2 + (nh_n^d)^{-1/2}\frac{A - m(x)B}{p(x)} + O_p(h_n^4 + \frac{1}{nh_n^d})] \\
&+ (1-\lambda)[Fg_m^2 + (mg_m^d)^{-1/2}g_m\frac{C - \hat{m}(x)D}{q(x)g_m} + O_p(g_m^4 + g_m^2h_n^2 + \frac{g_m}{mg_m^d} + \frac{g_m^2}{\sqrt{nh_n^{d+4}}})]
\end{aligned}
$$

The variance could be estimated with a slightly modified formula,

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

where $|z|$ is the norm of vector.

Appendix E. Coverage Error

In this section, we discuss theorems and calculations used to approximate the coverage error.

E.1 Berry–Esseen Theorem

Applying the Berry–Esseen theorem, we have:

$$\begin{aligned}\mathbf{P}\left\{\frac{r_n(\hat{y}_c - m(x)) - \hat{b}}{\hat{\sigma}^2} \le z_\alpha\right\} &= \mathbf{P}\{r_n(\hat{y}_c - m(x)) \le z_\alpha\hat{\sigma}^2 + \hat{b}\} \\ &= \mathbf{EP}\left\{\frac{1}{\sigma_{CD}^2}\left[-\frac{C-\hat{m}(x)D}{q(x)g_m}\right] \le [z_\alpha + O_p(\tfrac{(mg_m^3)^{1/2}}{(nh_n^5)^{1/2}})]\tfrac{\sigma^2}{\sigma_{CD}^2}\middle| X, Y\right\} \\ &\quad + O(\delta_c + \delta_b + \delta_\sigma) \\ &= \Phi(z_\alpha) + O\Big(\delta_c + \delta_b + \delta_\sigma + \delta_{CD} + \tfrac{(mg_m^3)^{1/2}}{(nh_n^5)^{1/2}} + m^{-1/2}g_m^{-1/2}\Big).\end{aligned}$$

where the error terms are of the following orders:

$$\begin{aligned}\delta_b &= O\Big(h_n^2 m^{1/2} g_m^{3/2} + n^{-1/2}h_n^{-5/2}m^{1/2}g_m^{3/2} + m^{1/2}g_m^{7/2} + m^{-1/2}g_m^{1/2}\Big), \\ \delta_{CD} &\sim \frac{mg_m^3}{nh_n^5}, \qquad \delta_c = r_n\left(h_n^4 + \frac{1}{\sqrt{nh_n}} + h_n^2 g_m^2 + \frac{h_n^2}{(mg_m)^{1/2}} + \frac{h_n^2}{mg_m^2}\right), \\ \delta_\sigma &= h_n^2 + (nh_n^3)^{-1/2} + \frac{mg_m^3}{nh_n^3} + \delta_q + \frac{mg_m^3}{nh_n^5}(\delta_\epsilon + \delta_p).\end{aligned}$$

E.2 Convergence rates

Precise expression of variance of hybrid estimator. For precision concerns, we express variance in a different way here.

$$\begin{aligned}Var(\hat{y_c}(x)) &= Var\{(nh_n)^{-1/2}\frac{A-m(x)B}{p(x)} - [O_p(\frac{1}{\sqrt{nh_n}} + h_n^2 g_m^2) + \frac{h_n^2(mg_m)^{-1/2}}{g_m}\frac{C-\hat{m}(x)D}{q(x)g_m}]\} \\ &= Var\{(nh_n)^{-1/2}\frac{A-m(x)B}{p(x)} - \frac{h_n^2(mg_m)^{-1/2}}{g_m}\frac{C-\hat{m}(x)D}{q(x)g_m}\} + Var(O_p(\frac{1}{\sqrt{nh_n}} + h_n^2 g_m^2)) \\ &\quad - 2Cov\{(nh_n)^{-1/2}\frac{A-m(x)B}{p(x)} - \frac{h_n^2(mg_m)^{-1/2}}{g_m}\frac{C-\hat{m}(x)D}{q(x)g_m}], O_p(\frac{1}{\sqrt{nh_n}} + h_n^2 g_m^2)\} \\ &= E(Var(\frac{h_n^2(mg_m)^{-1/2}}{g_m}\frac{C-\hat{m}(x)D}{q(x)g_m}|X,Y)) + Var((nh_n)^{-1/2}\frac{A-m(x)B}{p(x)}) \\ &\quad + O(\frac{1}{nh_n} + h_n^4 g_m^4 + \frac{h_n^2(mg_m)^{-1/2}}{g_m}\frac{1}{\sqrt{nh_n}} + \frac{h_n^2(mg_m)^{-1/2}}{g_m}h_n^2 g_m^2) \\ &= E[(\hat{m}'(x)^2)\frac{\sigma_k^2}{q(x)}\frac{h_n^4}{mg_m^3} + O(\frac{h^4}{mg_m})] + [\frac{1}{nh_n}\sigma_\epsilon^2\frac{r(k)}{p(x)} + O(\frac{h_n}{n})]\end{aligned}$$

Thus variance of the normalized estimator and its approximation is,

$$\sigma^2 = E[\hat{m}'(x)^2 \frac{\sigma_k^2}{q(x)} + O(g_m^2)] + [\frac{mg_m^3}{nh_n^5}\sigma_\epsilon^2 \frac{r(k)}{p(x)} + O(\frac{mg_m^3}{nh_n^3})]$$

$$\hat{\sigma^2} = \hat{m}'(x)^2 \frac{\sigma_k^2}{\hat{q}(x)} + \frac{mg_m^3}{nh_n^5}\hat{\sigma_\epsilon^2}\frac{r(k)}{\hat{p}(x)}$$

Finally, to accomodate the errors in calculations, we use $\delta_i, i = c, b, CD, \sigma, \sigma_\epsilon, p, q$ to denote them. In particular, $\hat{i}(x) = i(x) + O_p(\delta_i), i = p, q$ and $\hat{\sigma_\epsilon^2} = \sigma_\epsilon^2 + O_p(\delta_\epsilon), \hat{\sigma^2} = \sigma^2 + O(\delta_\sigma)$. And

$$\delta_{CD} = \sigma^2 - Var[\frac{C - \hat{m}(x)D}{q(x)g_m}]$$

$$\delta_c = r_n(\hat{y}_c - m(x) - (nh_n)^{-1/2}\frac{A - m(x)B}{p(x)} - \frac{h_n^2}{(mg_m^3)^{1/2}}\frac{C - \hat{m}(x)D}{q(x)g_m})$$

References

David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, Oct 2022. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2021.1915320.

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

Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989. ISSN 0169-2070.

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

Dong-Hyun 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.

Stephen S. M. Lee and Mehdi Soleymani. A simple formula for mixing estimators with different convergence rates. *Journal of the American Statistical Association*, 110(512): 1463–1478, Oct 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.960966.

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

W. R. Schucany, H. L. Gray, and D. B. Owen. On bias reduction in estimation. *Journal of the American Statistical Association*, 66(335):524–533, 1971. ISSN 0162-1459. doi: 10.2307/2283519.

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

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, Feb 2020. ISSN 1573-0565. doi: 10.1007/s10994-019-05855-6.

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

Larry Wasserman and John Lafferty. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL `https://papers.nips.cc/paper_files/paper/2007/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html`.

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

Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10 (1):25–47, Feb 2004. ISSN 1350-7265. doi: 10.3150/bj/1077544602.