Efficient estimation for semiparametric models by reproducing kernel Hilbert space *

Masaaki Imaizumi
University of Tokyo

December 9, 2014

Preliminary and Incomplete.

Abstract

A semiparametric model is a class of statistical models, which are characterized by a finite dimensional parameter and an infinite dimensional parameter. Asymptotic variance of estimator of the finite dimensional parameter is minimized when semiparametric efficient estimation is implemented. However, the efficient estimation is not possible for some models. We suggest a general method to carry out the efficient estimation for wide range of semiparametric models. Our method adopt a theory of reproducing kernel Hilbert space. Based on the theory, we represent an operator to a linear space of score function, and it enables us to implement the efficient estimation. We also provide theory of consistency of our method, and some numerical experiments.

*I would like to express my deepest gratitude to Katsumi Shimotsu who provided helpful comments and suggestions. I also owe a very important debt to Yoichi Nishimura, Kengo Kato and Yukitoshi Matsushita whose meticulous comments were an enormous help to me. I would also like to express my gratitude to my family for their moral support and warm encouragements. Finally, I gratefully appreciate the financial support of GSDM project that made it possible to complete our research. The responsibility of any errors is of course mine.

1 Introduction

A semiparametric model is a class of statistical models, which is characterized by a finite dimensional parameter θ and an infinite dimensional parameter η. For most cases, η is a possibly smooth function, and we separately estimate η by some nonparameteric way such as sieve or kernel. The semiparametric class includes numerous statistical models and there are used in vast application fields. For example, partially linear regression model, Cox regression model, single index model and so on. We formulate the object function of the model as $m(Z, \theta, \eta)$ with observation $\{Z_i\}_{i=1}^n$. Our main interest is value of θ, and we treat η as a nuisance parameter. In other words, a final goal of actual analysis is to derive a asymptotic distribution of θ.

Deriving the asymptotic distribution of the semiparametric model has some arbitrariness, and the point alters the asymptotic variance of estimators. *An semiparametric efficient estimation* is an estimation which minimizes the variance of, and its properties are discussed in [6], [20] and [17]. The estimator of θ and η is characterized by optimal conditions, and it adopts derivatives of $m(Z, \theta, \eta)$ with respect to each parameters. Since η has infinite dimension, we have to choose a derivative path function when implement the differentiation, and the selection of the derivative path affects the asymptotic variance. There are many researches to implement the efficient estimation for each models.

There are some general theory and method of semiparametric efficient estimation. [20] provides a *an efficient score function* and its condition

$$(I - \Pi_{\theta,\eta}) \frac{\partial}{\partial \theta} m(Z, \theta, \eta) = 0, \tag{1}$$

where I is an identity operator and $\Pi_{\theta,\eta}$ is a projection operator onto a linear space of score function with respect to η. However, evaluating the operator $\Pi_{\theta,\eta}$ requires an analytical form, and it is difficult for most semiparametric models. Some researches provide a theory and method to implement efficient estimation for general semiparametric models with some restriction. [15] show a theory for semipatametric models when η is estimated by sieve method and estimation should be a maximum likelihood estimation. [14] and [2] is a method for models which can be decomposed to conditional part. [10] provides a efficient estimation for instrumental variables. [3] is a more general method for sieve estimation for η and least square loss with moment condition restriction.

Our research provides a method to evaluate the equation (1) directly. Thus our method can implement semiparametric efficient estimation without additional restriction such as sieve method or requirement of estimation type. When the existence of $\Pi_{\theta,\eta}$ is guaranteed, our method approximate the projection operator by matrix form, and replicate the projection with n observation. When smoothness of the efficient score is satisfied, we can implement the semiparametric efficient estimation. As a result, we can carry out the efficient estimation for huge range of semiparametric models.

Our method depends on a theory of *Reproducing kernel Hilbert space*, henceforth RKHS. This is a method which represent an element of Hilbert space by linear sum of kernel function $k(\cdot, \cdot)$. When k has some properties, the kernel representation has theoretical foundation, and its details are in [5]. In this paper, we represent an operator by kernel

function k, and it enables us to handle operators. Similar techniques are used in [8] and [1], which represents operators by similar kernel methods.

The rest of this paper is organized as follows. Section 2 provide a basis theory of semiparametric estimation as preliminary. Section 3 is a main part of this paper, which provide a estimation method for semiparametric efficient estimation and its theoretical aspects. Section 4 is for numerical experiments. Section 5 is for conclusion. Appendix is consisted from proofs of all theorem, lemma and propositions.

2 Semiparametric model and efficient estimation

In this section, we provide a formulation of the semiparametric model as preparation. Mainly we show a definition of semiparametric model, asymptotic normality of the model, and efficient estimation of the model.

2.1 Semiparametric model

Consider a case n i.i.d. observations $\{Z_i\}_{i=1}^n$ are obtained, and $\forall i, Z_i \in \mathcal{Z}$. Suppose there exists a finite dimensional parameter $\theta \in \Theta \subset \mathcal{R}^p$, where Θ is a compact parameter space, and p is a number of dimension. Also assume there is a infinite dimensional parameter $\eta \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. Assume that $m(Z;\theta,\eta) : \mathcal{Z} \times \Theta \times \mathcal{H} \to \mathcal{R}$ is a known loss function and true value of the unique parameters (θ_0, η_0) and estimator of the parameters $(\hat{\theta}, \hat{\eta})$ satisfy following equation

$$
\begin{aligned}
(\theta_0, \eta_0) &= \arg\max E[m(Z;\theta,\eta)], \\
(\hat{\theta}, \hat{\eta}) &= \arg\max \frac{1}{n}\sum_{i=1}^{n} m(Z_i;\theta,\eta).
\end{aligned} \tag{2}
$$

In most cases, we treat η as a nuisance parameter, thus we do not care about value of η but we are interested in estimating θ. From now on, we will discuss about estimation properties of θ.

2.1.1 Examples

We provide some examples of the semiparametric model.

Partially linear model

Let $Z = (Y, X_1, X_2) \in \mathcal{Y} \times \mathcal{X}_2 \times \mathcal{X}_2$, where Y is a respondent variable, and X_1 and X_2 are covariates. We assume that the observation is obeyed from

$$
y_i = x_{1,i}^T \theta + \eta(x_{2,i}) + \epsilon_i,
$$

where $\eta(\cdot)$ is an unknown function, and ϵ_i is a noise variable with finite variance, conditional independent and mean zero. Since $\eta(\cdot)$ is infinite dimensional parameter, and we estimate $\eta(\cdot)$. When $\eta(\cdot)$ is nonlinear, this model can treat a nonlinear relationship between X_2 and Y.

Ordinary, the parameters are estimated by least square method as

$$(\hat{\theta}, \hat{\eta}) = \arg\min \frac{1}{n} \sum_{i=1}^{n} \left[y_i - x_{1,i}^T \theta - \eta(x_{2,i}) \right]^2 . \tag{3}$$

The estimation of $\hat{\eta}(\cdot)$ is implemented by sieve method or kernel method. Many estimation method for this models are suggested, for instance [12].

Copula

Copula is a statistical model for representing a relationship of variables. In this case, we consider the copula model for two variables. Let $Z = (X_1, X_2) \in \mathcal{X} \times \mathcal{X}$. The correlation of X_1 and X_2 is represented by a joint distribution function. Denote $C(\cdot, \cdot; \theta) \to \mathcal{X} \times \mathcal{X} \times \Theta \to [0, 1]$ as a copula function and denote the joint distribution function as

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2); \theta),$$

where $F_j(\cdot)$ is a distribution function of X_j. Some functional form are suggested as the copula function, and one example is Clayton-Cook-Johnson function

$$C(u_1, u_2; \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}. \tag{4}$$

In this model, the distribution function $F_j(\cdot)$ is a infinite parameter.

The parameter of interest θ is estimated by maximum likelihood function as

$$\hat{\theta} = \arg\max \frac{1}{n} \sum_{i=1}^{n} \log c(\hat{F}_1(x_{1,i}), \hat{F}_2(x_{2,i}); \theta). \tag{5}$$

$\hat{F}_j(\cdot)$ is an estimated empirical distribution function, and $c(\cdot, \cdot; \theta)$ is a pdf of the copula function.

2.2 Estimation and its properties

In this subsection, we show the estimation properties of θ, which is the parameter of our interest. Mainly, it includes consistency and asymptotic normality of $\hat{\theta}$. This asymptotic theory of the semiparametric estimator is mainly provided by [9]. Furthermore, these semiparametric asymptotic theory strongly relate to a semiparametric efficiency theory by [6] and [20].

In this part, we show a theory of asymptotic normality for estimating θ. These theories is provided by [9]. Firstly, we define derivatives as follow,

$$m_1(Z_i, \theta, \eta) = \frac{\partial}{\partial \theta} m(Z_i, \theta, \eta),$$
$$m_2(Z_i, \theta, \eta)[a] = \frac{\partial}{\partial \eta} m(Z_i, \theta, \eta)[a].$$

The first term $m_1(Z_i, \theta, \eta)$ is an ordinary partial derivative and it is a vector with dimension p. The second term $m_2(Z_i, \theta, \eta)$ is a functional partial derivative with respect to the infinite

dimensional parameter η. To implement the partial derivative, a derivative path function a is required. Thus, it is defined as

$$\frac{\partial}{\partial \eta} m(Z_i, \theta, \eta)[a] = \lim_{t \to 0} \frac{m(Z_i, \theta, \eta + ta) - m(Z_i, \theta, \eta)}{t}.$$

In [20], it is expressed as "one dimensional submodel". The choice of b is arbitrary, and we will discuss the choice later.

Higher order derivatives are defined similarly as

$$\begin{aligned}
m_{11}(Z_i, \theta, \eta) &= \frac{\partial}{\partial \theta} m_1(Z_i, \theta, \eta), \\
m_{21}(Z_i, \theta, \eta)[a] &= \frac{\partial}{\partial \theta} m_2(Z_i, \theta, \eta)[a], \\
m_{12}(Z_i, \theta, \eta)[a] &= \frac{\partial}{\partial \eta} m_1(Z_i, \theta, \eta)[a], \\
m_{22}(Z_i, \theta, \eta)[a_1][a_2] &= \frac{\partial}{\partial \eta} m_2(Z_i, \theta, \eta)[a_1][a_2],
\end{aligned}$$

where a_1 and a_2 are some derivative path functions. We also define a set of derivative path functions $A = (a_1, \ldots, a_p) \in \mathcal{H}^p$, and denote

$$m_2(Z_i, \theta, \eta)[A] = (m_2(Z_i, \theta, \eta)[a_1], \ldots, m_2(Z_i, \theta, \eta)[a_p])^T,$$

as a vector of the derivatives. The higher order derivatives are constructed in the same way.

For general semiparametric M-estimators, the estimator is obtained by minimizing the empirical loss (2). Denote a score function as

$$\tilde{m}(Z, \theta, \eta)[a] := m_1(Z, \theta, \eta) - m_2(Z, \theta, \eta)[a],$$

with some derivative path functions a. Thus, an optimal condition of the minimizing problem is written as

$$\tilde{m}(Z, \theta, \eta)[A] = 0,$$

with some A. In some models, the estimator $\hat{\theta}$ and $\hat{\eta}$ is calculated by two separated equations $\frac{1}{n}\sum_{i=1}^n m_1(Z_i, \theta, \eta) = 0$ and $\frac{1}{n}\sum_{i=1}^n m_2(Z_i, \theta, \eta)[A] = 0$. In semiparametric field, there is another way by constructing estimation for θ and η separately. In a such case, we require the consistency and the entropy condition on the each estimation equation, and obtain same result for asymptotic normality.

Using the second order derivatives, a Hesse matrix $H_{\theta,\eta}[A_1, A_2]$ for obtaining variance of the asymptotic distribution is defined as

$$H_{\theta,\eta}[A_1, A_2] = E\left[m_{11}(Z, \theta, \eta) + m_{12}(Z, \theta, \eta)[A_2] + m_{21}(Z, \theta, \eta)[A_1] + m_{22}(Z, \theta, \eta)[A_1][A_2]\right].$$

According to the previous notation, we show the conditions for asymptotic normality.

Assumption 1. *Following four conditions are satisfied.*

i.Consistency

$$||\hat{\theta} - \theta_0|| = op(1),$$
$$||\hat{\eta} - \eta_0|| = Op(n^{-c_1}),$$

where c_1 is a positive constant.

ii.Finite variance *For all $A_1, A_2 \in \mathcal{H}^p$, the Hessian $H_{\theta_0,\eta_0}[A_1, A_2]$ and $E[\tilde{m}(Z, \theta_0, \eta_0)\tilde{m}(Z, \theta_0, \eta_0)^T]$ are invertible and their determinants are finite.*

iii.Entropy condition *Let $\mathcal{M}_n = \{(\theta, \eta) : ||\theta - \theta_0|| = op(1), ||\eta - \eta_0|| = Op(n^{-c_1})\}$ and $\mathcal{F}_n = \{m(Z, \theta, \eta) : Z \in \mathcal{Z}, (\theta, \eta) \in \mathcal{M}_n\}$. Then*

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}_n, ||\cdot||_2)}d\epsilon < \infty,$$

where $N_{[]}(\cdot, \mathcal{F}, ||\cdot||)$ is a bracketing numbers.

iv.Smoothness *Let $c_2 > \max\{1, \frac{1}{2c_1}\}$ and δ_n is a sequence of positive constant converges to 0. For all $(\theta, \eta) \in \mathcal{M}_n$ and $A \in \mathcal{H}^p$,*

$$|E[\{\tilde{m}(Z, \theta, \eta) - \tilde{m}(Z, \theta_0, \eta_0)\} - \{m_{11}(Z, \theta, \eta) + m_{21}(Z, \theta, \eta)[A]\}\delta_\theta -$$
$$\{m_{12}(Z, \theta, \eta)[\delta_\eta/||\delta_\eta||] + m_{22}(Z, \theta, \eta)[A][\delta_\eta/||\delta_\eta||]\}||\delta_\eta||\delta_\theta]| = o(||\delta_\theta||) + O(||\delta_\eta||^{c_2}),$$

where $\delta_\theta = \theta - \theta_0, \delta_\eta = \eta - \eta_0$.

Assumption 1-i requires consistency of the estimators. The sufficient condition to obtain the consistency is well discussed in [19] and [9]. The convergence rate of η differ in each models. 1-iii yields that the loss function $m(Z, \theta, \eta)$ satisfies stochastic equicontinuous. A class of functions which satisfies the entropy condition is called Donsker, and the loss function belongs to the Donsker class has asymptotic equicontinuity. Local smoothness condition is required by assumption 1-iv. It requires that the loss function has higher smoothness in a small ball with the true value. The condition is often required in semiparametric literature.

By the conditions, the following theorem for asymptotic normality is obtained.

Theorem 1. *Consider the estimator in 2. If the assumption 1 is satisfied, then $\forall A_1$ and A_2,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, H^{-1}_{0,A_1,A_2}\Sigma_{A_1}H^{-1}_{0,A_1,A_2}),$$

where

$$H_{0,A_1,A_2} = H_{\theta_0,\eta_0}[A_1, A_2], \Sigma_{A_1} = E\left[\tilde{m}[A_1](Z, \theta_0, \eta_0)\tilde{m}[A_1](Z, \theta_0, \eta_0)^T\right]. \tag{6}$$

This theorem is based on a theory by [9] and proof is in Appendix.

2.3 Efficient derivative path and score

In previous part, the choice of a is arbitrary. A proper choice of a makes the variance in previous theorem 1 smaller. In [20], it is shown that there exists a score function $m_2(Z_i, \theta, \eta)[a^*]$ under some conditions, and it minimizes the asymptotic distribution. Consider a set of nuisance score functions $\mathcal{N} = \{m_2(Z_i, \theta, \eta)[a] : a \in \mathcal{H}\}$, and also consider a space $\overline{lin}(\mathcal{N})$, named a nuisance tangent space. To implement the efficient estimation, the score function $m_1(Z, \theta, \eta)$ should be orthogonal to the nuisance tangent space. To derive the score function with a^*, it is necessary to obtain a projection mapping to the nuisance tangent space.

Let $A^* = (a_1^*, \ldots, a_p^*)$ be a set of derivative path which minimizes the asymptotic variance (6). We will show how to obtain A^* in a following part, and its detail is described by [20]. In following part, we consider $p = 1$ case and mainly consider a^*. In multivariate case, we can derive A^* in same way for each element of θ.

To obtain the projection mapping to the nuisance tangent space, we denote that

$$B_{\theta,\eta}[f] = m_2(Z_i, \theta, \eta)[f],$$

where $B : \mathcal{H} \to \overline{lin}(\mathcal{N})$ is an operator. It is a mapping which takes a derivative path function f as argument and returns derivative function $m_2(Z_i, \theta, \eta)[f]$. To proceed the efficient estimation, a following assumption is required.

Assumption 2. *Following three conditions are satisfied.*
i.Bounded and linearity *$B_{\theta,\eta}$ is bounded linear operator.*
ii.Invertible *$B^*_{\theta,\eta} B_{\theta,\eta}$ is continuously invertible.*
iii. *For all a, $E[m_{12}(X, \theta, \eta)[a] - m_{22}(X, \theta, \eta)[a][a^*]] = 0$.*

where $B^*_{\theta,\eta}$ is an adjoint operator of $B_{\theta,\eta}$ and $\mathcal{R}(\cdot)$ represents a range.

By the assumption, we can denote a projection operator $\Pi : \mathcal{H} \to \overline{lin}(\mathcal{N})$ as

$$\Pi_{\theta,\eta} = B_{\theta,\eta}(B^*_{\theta,\eta} B_{\theta,\eta})^{-1} B^*_{\theta,\eta}. \tag{7}$$

According to the projection operator, we can obtain the score function with a^* by following calculation as

$$\tilde{m}(Z, \theta, \eta)[a^*] = (I - \Pi_{\theta,\eta}) m_1(Z, \theta, \eta), \tag{8}$$

where I is an identity operator. By this operation, we can evaluate the score function in the orthogonal complement space of the nuisance tangent space. Generally, $\tilde{m}(Z, \theta, \eta)[a^*]$ is called as efficient score function. Similarly, the optimal derivative path function a^* is represented as

$$a^* = (B^*_{\theta,\eta} B_{\theta,\eta})^{-1} B^*_{\theta,\eta}.$$

Finally, the efficient estimation is implemented by a following optimal condition

$$\frac{1}{n} \sum_{i=1}^{n} (I - \Pi_{\theta,\eta}) m_1(Z, \theta, \eta) = 0. \tag{9}$$

When the a^* is obtained, we denote the Hesse matrix and asymptotic variance by using a^*. Let us denote

$$H^*_{\theta,\eta} = E\left[m_1 1(Z,\theta_0,\eta_0) + m_{22}(X,\theta_0,\eta_0)[a^*][a^*]\right], \tag{10}$$
$$\Sigma^* = E[\tilde{m}(Z,\theta_0,\eta_0)[a^*]\tilde{m}(Z,\theta_0,\eta_0)[a^*]^T]. \tag{11}$$

Then, asymptotic variance of the estimator of θ becomes $H_*^{-1}\Sigma^* H_*^{-1}$ by theorem 1. Obtaining the efficient estimation and estimate the efficient asymptotic variance are the purpose of this paper.

To evaluate the efficient score, we have to obtain a value of $\Pi_{\theta,\eta} m_1(Z,\theta,\eta)$. However, it is difficult to obtain analytical form of the projection operator $\Pi_{\theta,\eta}$. We provide the models as examples to show how to evaluate the asymptotic distribution in each cases.

2.3.1 Examples (continued)

Partially linear model

According to the target function (3), the derivative functions are written as

$$m_1(Z,\theta,\eta) = -X_{1,i}(Y_i - X_{1,i}\theta - g(X_{2,i}))$$
$$m_2(Z,\theta,\eta)[a_1] = -a_1(X_{2,i})(Y_i - X_{1,i}\theta - g(X_{2,i})),$$

and the optimal condition requires that the derivatives are equal to zero. Higher order derivatives are written as

$$m_{11}(Z,\theta,\eta) = X_{1,i}^2$$
$$m_{12}(Z,\theta,\eta)[a_2] = X_{1,i}a_2(X_{2,i})$$
$$m_{21}(Z,\theta,\eta) = X_{1,i}a_1(X_{2,i})$$
$$m_{22}(Z,\theta,\eta)[a_1][a_2] = a_1(X_{2,i})a_2(X_{2,i}).$$

To implement the efficient estimation of the partially linear model, some methods are suggested. For instance, Robinson estimator by [12] can reach the asymptotic efficient variance when ϵ_i is identical. When ϵ_i is not identical, the estimator is not asymptotically efficient. When the $g(\cdot)$ is estimated by sieve, a method by [3] provide an efficient estimation.

Copula

We consider the target function (5) with copula function (5). We also denote copula density function as $c(X_1, X_2, \theta)$, first derivatives as $c_1(X_1, X_2, \theta)$ and $c_2(X_1, X_2, \theta)[a]$, and higher derivatives as following previous notation. The derivative functions of the target function are written as

$$m_1(Z,\theta,\eta) = \frac{c_1(X_1,X_2,\theta)}{c(X_1,X_2,\theta)}$$
$$m_2(Z,\theta,\eta)[a] = \frac{c_2(X_1,X_2,\theta)[a]}{c(X_1,X_2,\theta)}.$$

Higher order derivatives are written as

$$
\begin{aligned}
m_{11}(Z,\theta,\eta) &= \tilde{c}_{X,\theta}(c(X_1,X_2,\theta)c_{11}(X_1,X_2,\theta) - c_1(X_1,X_2,\theta)^2) \\
m_{12}(Z,\theta,\eta)[a_2] &= \tilde{c}_{X,\theta}(c(X_1,X_2,\theta)c_{12}(X_1,X_2,\theta)[a_2] - c_1(X_1,X_2,\theta)c_2(X_1,X_2,\theta)[a_2]) \\
m_{21}(Z,\theta,\eta)[a_1] &= \tilde{c}_{X,\theta}(c(X_1,X_2,\theta)c_{21}(X_1,X_2,\theta)[a_1] - c_1(X_1,X_2,\theta)c_2(X_1,X_2,\theta)[a_1]) \\
m_{22}(Z,\theta,\eta)[a_1][a_2] &= \tilde{c}_{X,\theta}(c(X_1,X_2,\theta)c_{22}(X_1,X_2,\theta)[a_1][a_2] - c_2(X_1,X_2,\theta)[a_1]c_2(X_1,X_2,\theta)[a_2]),
\end{aligned}
$$

where $\tilde{c}_{X,\theta} = c(X_1,X_2,\theta)^{-2}$.

[18] discussed that there is no method to implement efficient estimation for copula model with some copula function.

2.3.2 Relation to efficient variance bound

When the estimation is maximum likelihood estimation, in other words $m(Z,\theta,\eta) = \log(p(Z,\theta,\eta))$, the asymptotic variance from the efficient score estimation corresponds to a semiparametric efficient variance bound. It is greatly detailed in [20].

3 Method to differentiate path

In this section, we show our suggesting method for estimating the asymptotic distribution. To obtain the distribution, we have to evaluate the efficient score (8) by estimating the projection operator (7), and solve the optimal condition (9). However, evaluating the score is difficult since calculating an operator without analytical form is hard. The reproducing kernel Hilbert space (RKHS) enables us to evaluate it. We show its detail as follows.

3.1 Kernel representation

The RKHS method is a way of representing a component of Hilbert space by kernel function. Consider a positive semi definite kernel function $k(z,z') : \mathcal{Z}\times\mathcal{Z} \to \mathcal{R}$, which satisfies $\forall z,z', k(z,z') \geq 0, k(z,z') = k(z',z)$. Gaussian kernel $k(z,z') = (2\pi)^{-1/2}\exp(-\frac{||z-z'||^2}{2})$ is often used.

Denote $\mathcal{H}_k$ is a reproducing kernel Hilbert space with kernel k, which equips an inner product $\langle\cdot,\cdot\rangle$. By the inner product, it is known that there exist a unique $\mathcal{H}_k$, where $\forall x, k(x,y) \in \mathcal{H}_k$ and $lin\{k(\cdot,x_i)\}$ is dense in the $\mathcal{H}_k$. As a result, $\forall f \in \mathcal{H}_k$ satisfies following properties, reproductivity $f(x) = \langle f, k(\cdot,x)\rangle$ and kernel representation $f(x) = \sum_i \alpha_i k(x,x_i)$ with some weight $\{\alpha_i\}$. These theories are described in [5].

Our method evaluate the projection operator $\Pi_{\theta,\eta}$, we estimate the derivative operator $B_{\theta,\eta}$ by RKHS method. To implement the estimation, we provide some theorems. Assume that $B_{\theta,\eta} \in \mathcal{L}_0(\mathcal{H}_k,\mathcal{H}_k)$, where $\mathcal{L}_0(\mathcal{X},\mathcal{Y})$ is a set of compact operator from $\mathcal{X}$ to $\mathcal{Y}$.

We represent an operator in $\mathcal{L}_0(\mathcal{X},\mathcal{Y})$ by the kernel $k(\cdot,\cdot)$. For all $f\ in\mathcal{H}_k$, it is writte as

$$
B_{\theta,\eta}[f](x) \approx \sum_{i=1}^{\infty} w_i f(x_i)k(x_i,x), \tag{12}
$$

where $\{w_i\}$ is a sequence of weights. It seems to be a spectral decomposition of operators, however we accept another way to justify this representation.

We also mention a representer theorem described in [13]. This theorem guarantees that sum of n kernels is sufficient to optimize an empirical minimization problem with n observations. In other words, only n bases can optimize the n observation optimization problem. This theorem justifying the representation of function by n kernels, and we apply the theorem for the operator.

Based on the discussion, we show a theory that n kernels is sufficient to estimate an operator with empirical minimization problem.

Proposition 1. *Consider a case n observations is obtained and minimization problem with a loss function $l(\cdot) : \mathcal{Z}^n \times \mathcal{L}_0(\mathcal{H}_k, \mathcal{H}_k) \to \mathcal{R}$ and a penalty function $\Omega : [0, \infty) \to \mathcal{R}$,*

$$\min_{B \in \mathcal{L}_0(\mathcal{H}_k, \mathcal{H}_k)} l(\{Z_i\}_{i=1}^n) + \Omega(||B||).$$

Then, the minimization of the problem is represented as

$$B[f](x) = \sum_{i=1}^{n} w_i f(x_i) k(x_i, x).$$

Proof is in Appendix. Finally, only calculating n components $\{w_i\}_{i=1}^n$ enabled us to evaluate the operator $B_{\theta,\eta}$, with n observations and arbitrary f.

3.2 Operator estimation

In this subsection, we show how to estimate the operator $B_{\theta,\eta}$ by kernel, and represent the projection operator $\Pi_{\theta,\eta}$. We have three steps; (i)obtain a functional derivative coefficients $\{m_2(Z_i, \theta, \eta)[f]\}_{i=1}^n$ for n sample with some function $f \in \mathcal{H}_k$, (ii)estimate the weight $\{w_i\}_{i=1}^n$ with the derivative coefficients as training data, (iii)construct $\hat{\Pi}_{\theta,\eta}$ by using the operator $\hat{B}_{\theta,\eta}$ from estimated $\{w_i\}_{i=1}^n$. Then, we can implement the efficient estimation problem.

Firstly, we actually calculate the functional derivative coefficient $m_2(Z, \theta, \eta)[f]$ with observations $\{Z_i\}_{i=1}^n$. In this part, the derivative path function $f \in \mathcal{H}_k$ is an arbitrary measurable function. Then we define the derivative coefficients as

$$B[f](Z_i) = \lim_{t \to 0} \frac{m(Z_i, \theta, \eta + tf) - m(Z_i, \theta, \eta)}{t},$$

and we get $\{B_{\theta,\eta}[f](Z_i)\}_{i=1}^n$.

Secondly, we estimate $B_{\theta,\eta}$ by following penalized least square estimation as

$$\min_{\hat{B} \in \mathcal{L}_0(\mathcal{H}_k, \mathcal{H}_k)} \sum_{i=1}^{n} (B[f](Z_i) - \hat{B}[f](Z_i))^2 + \lambda_n \Omega(B), \tag{13}$$

where λ_n is a penalty coefficient and $\Omega(\cdot)$ is a penalty function. The penalty term is to avoid overfitting of this estimation problem. By the proposition 1, this optimization problem is rewritten as

$$\min_{\{w\}_{j=1}^n} \sum_{i=1}^n (B[f](Z_i) - \sum_{j=1}^n w_i f(Z_j) K(Z_j, Z_i))^2 + \lambda_n \Omega(B).$$

We will rewrite this problem by matrix representation. Denote each matrix as

$$Y = \begin{pmatrix} B[f](X_1) \\ \vdots \\ B[f](X_n) \end{pmatrix} K = \begin{pmatrix} k(Z_1, Z_1) & \dots & k(Z_1, Z_n) \\ \vdots & \ddots & \vdots \\ k(Z_n, Z_1) & \dots & k(Z_n, Z_n) \end{pmatrix}$$
$$F = \begin{pmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{pmatrix} W = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

Then, we rewrite the problem (13). We also let the penalty function $\Omega(B) = ||W||_2^2$. It is written as

$$\min_{W \in R^n} \frac{1}{n} ||Y - K[F \circ W]||_2^2 + \lambda_n ||W||_2^2,$$

where $\circ$ means element multiplication. Preliminary, we define an extended gram matrix $\overline{K} \in \mathcal{R}^n \times \mathcal{R}^n$ with its element as

$$\overline{K}_{ij} = k(Z_i, Z_j) f_j.$$

An optimal conditions provides us

$$\hat{W} = (\overline{K}^T \overline{K} + \lambda_n I_n)^{-1} \overline{K}^T Y,$$

where I_n is an identity matrix. Thus, we can estimate $\{w_i\}$ which is invariant to the f as the argument.

Thirdly, we provide the estimator of the projection operator $\Pi_{\theta,\eta}$. We define a weighted gram matrix $G \in \mathcal{R}^n \times \mathcal{R}^n$ with its element

$$G_{ij} = k(Z_i, Z_j) \hat{w}_j.$$

Then, we can evaluate the operator with n observations. Consider the vector of $f(Z_i)$. By multiplying G, we obtain a mapped function on n as

$$\begin{pmatrix} \hat{B}[f](Z_1) \\ \vdots \\ \hat{B}[f](Z_n) \end{pmatrix} = G \begin{pmatrix} f(Z_1) \\ \vdots \\ f(Z_n) \end{pmatrix}. \tag{14}$$

We define the estimator of the projection operator as

$$\hat{\Pi}_{\theta,\eta} = \hat{B}_{\theta,\eta}(\hat{B}^*_{\theta,\eta}\hat{B}_{\theta,\eta})^{-1}\hat{B}^*_{\theta,\eta}.$$

Then, an element representation of $\hat{\Pi}$ is written as

$$\hat{\Pi}_{\theta,\eta}[f](Z_i) = \sum_{j=1}^{n}[G(G^TG)^{-1}G^T]_{ij}f(Z_j). \tag{15}$$

By this form, we can evaluate the value of $\Pi_{\theta,\eta}f$ for all f with n observations.

Using the method, we can obtain the efficient score as

$$\tilde{m}(Z,\theta,\eta)[a^*] = (I - \hat{\Pi}_{\theta,\eta})m_1(Z,\theta,\eta).$$

It enables us to rewrite the optimal condition (9) by the matrix G. It is written as

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}(I - \hat{\Pi}_{\theta,\eta})m_1(Z,\theta,\eta) \\ &= \frac{1}{n}[I_n - G(G^TG)^{-1}G^T]\begin{pmatrix} m_1(Z_1,\theta,\eta) \\ \vdots \\ m_1(Z_n,\theta,\eta)\end{pmatrix}. \end{aligned} \tag{16}$$

3.3 Theory of method

We provide a theoretical aspect of our method. Purpose of this section is to show that the estimator $\hat{B}_{\theta,\eta}$ and $\hat{\Pi}_{\theta,\eta}$ are consistent, and the suggested estimation problem (16) can provide the asymptotic distribution with the efficient variance. Proofs of all lemmas and theorems are in Appendix.

Assumption are as follows.

Assumption 3. *Following four conditions are satisfied.*

1.Condition on Kernel *Kernel function k are continuous, bounded and positive semi-definite.*

2.Moment condition *There exists a positive constant C_f and σ, which satisfies $C_f > ||w_j^*k(\cdot,Z_j)f(Z_j)||$ and $\sigma^2 \geq E[\{w_j^*k(\cdot,Z_j)f(Z_j)\}^2]$ for all j and f.*

3.Boundedness *$||Z||_\infty < \infty$.*

4.Smoothness of score operator *For all $(\theta,\eta) \in \mathcal{M}_n$, $||B_{\theta,\eta} - B_{\theta_0,\eta_0}||_{op} \leq C_b\{||\theta - \theta_0|| + ||\eta - \eta_0||^{c^*}\}$ with some positive constant C_b*

The last assumption requires smoothness of the operator $B_{\theta,\eta}$. This is same condition to assumption 3. It is necessary to show root-n convergence of the estimator.

The consistency of $\hat{B}_{\theta,\eta}$ is shown in a following lemma. This theory bases on a theory of kernel ridge regression and concentration inequality on empirical process. The convergence rate of the penalty parameter λ_n plays a critical role.

Lemma 1. *Consider $B \in \mathcal{L}_0(\mathcal{H}_k, \mathcal{H}_k)$, and $\hat{B}$ is an estimator of it defined in (14). If the assumption 3-i,3-ii and 3-iii are satisfied and $\lambda_n = O(\frac{1}{n})$, then with large probability,*

$$||B - \hat{B}||_{op} = O(\frac{1}{\sqrt{n}}).$$

Based on the lemma, the consistency of $\Pi_{\theta,\eta}$ is shown in following theorem. Compactness of operators becomes a key factors.

Lemma 2. *Consider $\Pi \in \mathcal{L}_0(\mathcal{H}_k, \overline{lin}(\mathcal{N})$ is a projection operator defined at (7), and $\hat{\Pi}$ is an estimator of it defined in (15). If the assumption 3-i,3-ii and 3-iii are satisfied, then with large probability,*

$$||\Pi - \hat{\Pi}||_{op} = O(\frac{1}{\sqrt{n}}).$$

Finally, we will show that the limit distribution of the suggested optimization problem.

Theorem 2. *Consider the estimation problem (16). If assumption 1, 2 and 3 are satisfied, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, H_*^{-1}\Sigma^* H_*^{-1}).$$

The asymptotic variance term is denoted in (10) and (11).

4 Numerical Experiment

In this section, we provide a result of numerical experiments to show the effectiveness of our method.

Firstly, we use the partially linear model with non-identical noise term. It is known that ordinal method cannot realize the efficient estimation. In contrast, our method can implement the efficient estimation, our method should have smaller variance.

We replicate the estimation for 200 times, and mean and variance of the result is in table 4. We can see that our method can improve the efficiency.

We provide same experiment for copula model with Clayton-type copula function (4). As discussed before, no method can carry out the efficient estimation. Then we provide our proposed method. Result is in table 4. We can see that our method can obtain the efficiency.

	$\theta = 1.0$		$\theta = 3.0$		$\theta = 5.0$	
$n = 100$	mean	variance	mean	variance	mean	variance
Ordinal	0.9699	0.1024	3.0084	0.1224	4.9574	0.0803
Kernel	0.9818	0.0849	2.9964	0.1050	4.9583	0.0795
$n = 200$	mean	variance	mean	variance	mean	variance
Ordinal	0.9970	0.0472	3.0068	0.0509	5.0127	0.0451
Kernel	0.9948	0.0410	3.0043	0.0415	5.0125	0.0438
$n = 300$	mean	variance	mean	variance	mean	variance
Ordinal	0.9949	0.0366	3.0084	0.0358	5.0329	0.0350
Kernel	0.9996	0.0313	3.0044	0.0332	5.0322	0.0346
$n = 400$	mean	variance	mean	variance	mean	variance
Ordinal	1.0075	0.0261	3.0166	0.0212	4.9695	0.0246
Kernel	1.0042	0.0241	3.0126	0.0195	4.9698	0.0242
$n = 500$	mean	variance	mean	variance	mean	variance
Ordinal	0.9987	0.0273	2.9970	0.0236	5.0001	0.0189
Kernel	1.0000	0.0225	2.9958	0.0202	4.9998	0.0184

Table 1: :Estimation of partially linear model for 200 times. Values are mean and variance of the replication.

	$\theta = 3.0$		$\theta = 4.0$		$\theta = 5.0$	
$n = 100$	mean	variance	mean	variance	mean	variance
Ordinal	3.6759	0.2307	4.6415	0.4024	5.3926	0.6680
Kernel	3.6152	0.1675	4.3464	0.3060	5.1199	0.6196
$n = 200$	mean	variance	mean	variance	mean	variance
Ordinal	3.7323	0.1222	4.5659	0.2180	5.5586	0.2415
Kernel	3.6624	0.0899	4.0244	0.1408	4.7989	0.2247
$n = 300$	mean	variance	mean	variance	mean	variance
Ordinal	3.7051	0.0785	4.7026	0.1322	5.5870	0.1678
Kernel	3.6100	0.0431	4.0848	0.1058	4.6194	0.1335
$n = 400$	mean	variance	mean	variance	mean	variance
Ordinal	3.7370	0.0626	4.6944	0.1018	5.5930	0.1448
Kernel	3.6543	0.0487	4.1008	0.0676	4.5497	0.0840
$n = 500$	mean	variance	mean	variance	mean	variance
Ordinal	3.7679	0.0559	4.7344	0.0795	5.6589	0.1430
Kernel	3.6080	0.0384	4.1522	0.0483	4.6544	0.0849

Table 2: :Estimation of copula model with Clayton copula function for 200 times. Values are mean and variance of the replication.

5 Conclusion

Our research provide a method to implement the semiparametric efficient estimation (9). When the ordinal estimator of θ with root-n convergence and asymptotic normality, our method can carry out the estimation with semiparametric efficient variance. The conditions to implement our method is very weak. Other methods which require strong restriction such as sieve estimation or loss function formation. In contrast, our method is applicable to wide range of semiparametric models.

A Proof of theorem 1

This proof is mainly provided by [9]. Most notations and conditions are same to theorem 1 in [9]. A main difference is about the entropy condition in the assumption 1. Based on the [11], the entropy condition leads stochastic equicontinuity.

Denote an empirical process as $G_n f = \sqrt{n}\frac{1}{n}\{\sum_{i=1}^{n} f(X_i) - E[f(X)]\}$. If the entropy condition 1-iii is satisfied, then for all $\delta \to 0$ and a constant C,

$$\sup_{(\theta,\eta)\in\mathcal{N}_n} |G_n\tilde{m}(X,\theta,\eta) - \tilde{m}(X,\theta_0,\eta_0)| = op(1).$$

Thus, the condition A3 of the theorem 1 in [9] is satisfied, and the asymptotic normality is obtained.

B Proof of proposition 1

Firstly, we show that the representation (12) is valid. Following discussion traces the proof of [4].

Consider a linear space $\mathcal{L} = lin\{f(\cdot)k(z,\cdot) : f \in \mathcal{H}_k, z \in \mathcal{Z}\}$, with inner product $\langle B, B'\rangle = \sum_i \sum_j w_i f(z_i) f(z_j) w'_j k(z_i, z_j)$. By the representation of (12), for all $B \in \mathcal{L}$ and f, a function $B[f]$ is written as linear sum of kernels with some $\{w_i\}$. By this form we get $B[f](\cdot) \in \mathcal{H}_k$, and completeness of $\mathcal{L}$ is provided since $\mathcal{H}_k$ is complete.

Next we try to show $\mathcal{L}_0(\mathcal{H}_k, \mathcal{H}_k)$ and $\mathcal{L}$ are equivalent. Let $\{B_j\}$ be a Cauchy sequence in $\mathcal{L}$, and also denote $[\{B_j\}]$ be an equivalent class including $\{B_j\}$. When $[\{B_j\}] = [\{B'_j\}]$, we have $||B_j[f] - B'_j[f]|| \leq ||B_i - B_j||||f|| \to 0$, for all $f \in \mathcal{H}_k$. Then we obtain $\lim_{j\to\infty} B_j[f] = \lim_{j\to\infty} B'_j[f]$, and the limit value of the sequence is independent from the element of equivalent classes. Thus, we can define a mapping $\Phi : \mathcal{L} \to \mathcal{L}_0$. When $\forall f, \lim_{j\to\infty} B_j[f] = 0$, we obtain $[\{B_j\}] = 0$. Thus, Φ is injection and linear, then we can see that $\mathcal{L}$ and $\mathcal{L}_0$ are equivalent.

Secondly, we apply the representation theorem [13] for our optimization problem. Let $\mathcal{L}_0^n = lin\{f(z_i)k(z,z_i) : i = 1,\ldots,n, f \in \mathcal{H}_k, z \in \mathcal{Z}\}$ be a linear space spanned with n observations. We implement orthogonal decomposition $\mathcal{L}_0 = \mathcal{L}_0^n \bigoplus \mathcal{L}_0^{\perp}$, where $\mathcal{L}_0^{\perp}$ is an orthogonal complement space of $\mathcal{L}_0^n$. We also decompose $B = B^n + B^{\perp}$ where $B^n \in \mathcal{L}_0^n$ and $B^{\perp} \in \mathcal{L}_0^{\perp}$. Because of the orthogonality, $B[f](Z_i) = B^n[f](Z_i) + B^{\perp}[f](Z_i) = B^n[f](Z_i)$. Furthermore, we obtain $||B||^2 = ||B^n||^2 + ||B^{\perp}||^2$, hence $\Omega(||B^n||) \leq \Omega(||B||)$. According to the discussion, using only n bases does not affect the value of $B[f](Z_i)$, and also it minimizes the value of penalty. Hence, we obtain the result of the proposition.

C Proof of lemma 1

For all θ and η, the operator norm is decomposed as

$$
\begin{aligned}
&||\hat{B}_{\theta,\eta} - B_{\theta,\eta}||_{op} \\
&= \sup_{||f||=1} ||\hat{B}_{\theta,\eta}[f] - B_{\theta,\eta}[f]|| \\
&\leq \sup_{||f||=1} ||\sum_{j=1}^{n} \hat{w}_j k(\cdot, Z_j) f(Z_j) - \sum_{j=1}^{n} w_j^* k(\cdot, Z_j) f(Z_j)|| \\
&+ \sup_{||f||=1} ||\sum_{j=1}^{n} w_j^* k(\cdot, Z_j) f(Z_j) - B_{\theta,\eta}[f]||,
\end{aligned}
$$

where $\{w_j^*\}$ is a sequence of weight which satisfies

$$
B_{\theta,\eta}[f] = \sum_{j=1}^{\infty} w_j^* k(z, Z_j) f(Z_j).
$$

About the first term,

$$
\begin{aligned}
&\sup_{||f||=1} ||\sum_{j=1}^{n} \hat{w}_j k(\cdot, Z_j) f(Z_j) - \sum_{j=1}^{n} w_j^* k(\cdot, Z_j) f(Z_j)|| \\
&= \sup_{||f||=1} ||\sum_{j=1}^{n} (\hat{w}_j - w_j^*) k(\cdot, Z_j) f(Z_j)|| \\
&\leq \sup_{||f||=1} ||\sum_{j=1}^{n} (\hat{w}_j - w_j^*) k(\cdot, Z_j) f(Z_j)|| \\
&\leq \sup_{||f||=1} \sum_{j=1}^{n} |\hat{w}_j - w_j^*| ||k(\cdot, Z_j) f(Z_j)||.
\end{aligned}
$$

The $\hat{w}_j - w_j^*$ is difference of ridge estimator and true value and its convergence rate is $O(\frac{\lambda_n}{\sqrt{n}})$. Since kernel k and argument f are bounded, we obtain that

$$
\sup_{||f||=1} \sum_{j=1}^{n} |\hat{w}_j - w_j^*| ||k(\cdot, Z_j) f(Z_j)|| = O(\lambda_n \sqrt{n}).
$$

About the second term, we apply Talagrand's inequality by [16] and [7]. By the assumption 3, each boundedness is obtained and it is represented by C_b and σ. Then,

with probability $1-\delta$,

$$\begin{aligned}
&\sup_{||f||=1} ||\sum_{j=1}^{n} w_j^* k(\cdot, Z_j) f(Z_j) - B_{\theta,\eta}[f]|| \\
&\leq \frac{1}{\sqrt{n}}\sqrt{2\sigma^2 \log\frac{1}{\delta}} + \frac{1}{n}\left(\sqrt{4C_f B_{\theta,\eta}[f] \log\frac{1}{\delta} + \frac{1}{3}C_f \log\frac{1}{\delta}}\right) \\
&= O\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Then, we get that

$$||\hat{B}_{\theta,\eta} - B_{\theta,\eta}||_{op} = O(\lambda_n\sqrt{n}) + O(\frac{1}{\sqrt{n}}).$$

The convergence rate of λ_n provides the result.

D Proof of theorem 2

To show the consistency of $\hat{\Pi} = \hat{B}_{\theta,\eta}(\hat{B}^*_{\theta,\eta}\hat{B}_{\theta,\eta})^{-1}\hat{B}^*_{\theta,\eta}$, we show the consistency of $\hat{B}^*_{\theta,\eta}\hat{B}_{\theta,\eta}$ and $(\hat{B}^*_{\theta,\eta}\hat{B}_{\theta,\eta})^{-1}$.

The first point is shown as

$$\begin{aligned}
||B^*B - \hat{B}\hat{B}^T|| &\leq ||B^*B - B^*\hat{B}|| + ||B^*\hat{B} - \hat{B}^T\hat{B}|| \\
&\leq ||B^*||||B - \hat{B}|| + ||B^* - \hat{B}^T||||\hat{B}|| \\
&= O(\frac{1}{\sqrt{n}}).
\end{aligned}$$

The second point is

$$\begin{aligned}
||(B^*B)^{-1} - (\hat{B}^T\hat{B})^{-1}|| &= ||(B^*B)^{-1}\{B^*B - \hat{B}\hat{B}^T\}(\hat{B}^T\hat{B})^{-1}|| \\
&= O(\frac{1}{\sqrt{n}}).
\end{aligned}$$

Then, the consistency of $\hat{\Pi}_{\theta,\eta}$ is written as

$$
\begin{aligned}
&||\hat{\Pi} - \Pi|| \\
&= ||\hat{B}(\hat{B}^T\hat{B})^{-1}\hat{B}^T - B(B^*B)^{-1}B^*|| \\
&\leq ||(\hat{B} - B)(\hat{B}^T\hat{B})^{-1}\hat{B}^T|| + ||\hat{B}(\hat{B}^T\hat{B})^{-1}(\hat{B}^T - B^*)|| \\
&+ ||B\{(B^*B)^{-1} - (\hat{B}^T\hat{B})^{-1}\}B^*|| \\
&\leq ||\hat{B} - B||||\hat{B}^T|| + ||\hat{B}||||\hat{B}^T - B^*|| \\
&+ ||B(B^*B)^{-1/2}\{(B^*B)^{1/2}(B^*B)^{-1}(B^*B)^{1/2} - I\}(B^*B)^{-1/2}B^*|| \\
&\leq ||\hat{B} - B||||\hat{B}^T|| + ||\hat{B}||||\hat{B}^T - B^*|| \\
&+ ||B||||B^*||||(B^*B)^{1/2}(B^*B)^{-1}(B^*B)^{1/2} - I|| \\
&\leq ||\hat{B} - B||||\hat{B}^T|| + ||\hat{B}||||\hat{B}^T - B^*|| \\
&+ ||B||||B^*||||(\hat{B}^T\hat{B})^{-1/2}(B^*B)(\hat{B}^T\hat{B})^{-1/2} - I|| \\
&\leq ||\hat{B} - B||||\hat{B}^T|| + ||\hat{B}||||\hat{B}^T - B^*|| \\
&+ ||B||||B^*||||(\hat{B}^T\hat{B})^{-1/2}(B^*B - \hat{B}^T\hat{B})(\hat{B}^T\hat{B})^{-1/2}|| \\
&\leq ||\hat{B} - B||||\hat{B}^T|| + ||\hat{B}||||\hat{B}^T - B^*|| + ||B||||B^*||||B^*B - \hat{B}^T\hat{B}||.
\end{aligned}
$$

By using the boundedness of $||B||$ and $||\hat{B}||$, the consistency is obtained as

$$||\hat{\Pi}_{\theta,\eta} - \Pi_{\theta,\eta}|| = O(\frac{1}{\sqrt{n}}).$$

E Proof of lemma 2

Firstly, we use the assumption of consistency and the entropy condition. By boundedness of $\hat{\Pi}_{\theta,\eta}$,

$$(I - \hat{\Pi}_{\hat{\theta},\hat{\eta}})G_n m_1(Z, \hat{\theta}, \hat{\eta}) = (I - \hat{\Pi}_{\hat{\theta},\hat{\eta}})G_n m_1(Z, \theta_0, \eta_0) + op(1).$$

Simple calculation yields

$$\sqrt{n}(I - \hat{\Pi}_{\hat{\theta},\hat{\eta}})E[m_1(Z, \hat{\theta}, \hat{\eta}) - m_1(Z, \theta_0, \eta_0)] = -(I - \hat{\Pi}_{\hat{\theta},\hat{\eta}})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_1(Z_i, \theta_0, \eta_0) + op(1).$$

Consider the estimation error of the projection operator. For all Z and $\theta, \eta \in \mathcal{M}$,

$$
\begin{aligned}
&|\hat{\Pi}_{\hat{\theta},\hat{\eta}}m_1(Z, \theta, \eta) - \Pi_{\theta,\eta}m_1(Z, \theta, \eta)| \\
&\leq |\hat{\Pi}_{\hat{\theta},\hat{\eta}}m_1(Z, \theta, \eta) - \Pi_{\hat{\theta},\hat{\eta}}m_1(Z, \theta, \eta)| + |\Pi_{\hat{\theta},\hat{\eta}}m_1(Z, \theta, \eta) - \Pi_{\theta,\eta}m_1(Z, \theta, \eta)|.
\end{aligned}
$$

By the theorem 1 and smoothness assumption 1, we obtain

$$
\begin{aligned}
&|\hat{\Pi}_{\hat{\theta},\hat{\eta}}m_1(Z, \theta, \eta) - \Pi_{\hat{\theta},\hat{\eta}}m_1(Z, \theta, \eta)| + |\Pi_{\hat{\theta},\hat{\eta}}m_1(Z, \theta, \eta) - \Pi_{\theta,\eta}m_1(Z, \theta, \eta)| \\
&O(\frac{1}{\sqrt{n}}) + o(||\hat{\theta} - \theta||) + O(||\hat{\eta} - \eta||^{c^*}).
\end{aligned}
$$

From the definition of projection operator, remember that

$$\Pi_{\theta,\eta} m_1(Z,\theta,\eta) = m_2(Z,\theta,\eta)[a^*].$$

Then we obtain the efficient derivative path function, and also we obtain

$$\begin{aligned}(I - \Pi_{\theta,\eta} m_1(Z,\theta,\eta)) &= m_1(Z,\theta,\eta) - m_2(Z,\theta,\eta)[a^*] \\ &= \tilde{m}(Z,\theta,\eta)[a^*].\end{aligned}$$

Then, we obtain a following equation

$$\begin{aligned}&\sqrt{n}E[\tilde{m}(Z,\hat{\theta},\hat{\eta})[a^*] - \tilde{m}(Z,\theta_0,\eta_0)[a^*]] \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{m}(Z_i,\theta_0,\eta_0)[a^*] + op(1) + o(||\hat{\theta} - \theta_0||) + O(||\hat{\eta} - \eta_0||^{c^*}).\end{aligned}$$

Rest of this proof is as same as theorem 1. The smoothness implies that the estimation problem 16 can obtain the asymptotic distribution with efficient variance.

References

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *The Journal of Machine Learning Research*, 10:803–826, 2009.

[2] C. Ai. A semiparametric maximum likelihood estimator. *Econometrica*, pages 933–963, 1997.

[3] C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404, 1950.

[5] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Springer, 2004.

[6] P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[7] O. Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[8] K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, pages 1871–1905, 2009.

[9] S. Ma and M. R. Kosorok. Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96(1):190–217, 2005.

[10] W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[11] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York., 1984.

[12] P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, pages 931–954, 1988.

[13] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer, 2001.

[14] T. A. Severini and W. H. Wong. Profile likelihood and conditionally parametric models. *The Annals of Statistics*, pages 1768–1802, 1992.

[15] X. Shen. On methods of sieves and penalization. *The Annals of Statistics*, pages 2555–2591, 1997.

[16] M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

[17] A. Tsiatis. *Semiparametric theory and missing data*. Springer, 2007.

[18] H. Tsukahara. Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375, 2005.

[19] S. A. van de Geer. *Empirical Processes in M-estimation*, volume 105. Cambridge university press Cambridge, 2000.

[20] A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.