Semiparametric Efficient Estimation by Reproducing Kernel Hilbert Space

Masaaki Imaizumi*

Abstract

A *semiparametric model* is a class of statistical models, characterized by a finite-dimensional parameter and an infinite-dimensional parameter. In many cases, we are interested in inference of the finite-dimensional parameter. The efficiency of the inference is reduced by the nuisance infinite dimensional parameter, and we deem an estimator *semiparametric efficient* when the estimator of the finite dimensional parameter minimizes the efficiency loss. We suggest a general method to implement the semiparametric efficient estimation by representing the projection operator by the kernel function. Our proposed method does not restrict how the infinite-dimensional parameter is estimated, and thus our method can enable efficient estimation for a wide range of semiparametric models. We also prove the guarantee of our method, and conduct some numerical experiments.

Key Words: Semiparametric model, semiparametric efficient, reproducing kernel Hilbert space

1. Introduction

A semiparametric model is a class of statistical models, characterized by a finite-dimensional parameter $\theta \in \mathcal{R}^p$ and an infinite-dimensional parameter $\eta \in \mathcal{H}$. p is the number of dimensions of θ, and $\mathcal{H}$ is an Hilbert space. We consider the general scheme "M-estimator" where the estimator are obtained by minimizing criteria functions, and denote the criteria function of the semiparametric model as $m(Z, \theta, \eta)$ with n i.i.d. observations $\{Z_i\}_{i=1}^n$. We obtain the estimators $(\hat{\theta}, \hat{\eta})$ by optimizing the empirical mean of the criteria function $m(Z, \theta, \eta)$ as

$$(\hat{\theta}, \hat{\eta}) = \arg\max_{(\theta,\eta)} \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \theta, \eta).$$

The class of semiparametric models includes numerous statistical models, and such models are used in many fields. As examples, consider the partially linear regression model (Härdle and Liang (2007)) , Cox regression model (Cox (1972)), single index model (Ichimura (1993)), and so on. Our main interest is to conduct inference of θ, and we treat η as a nuisance parameter. In many cases, we separately estimate η by some nonparameteric estimation, and then estimate θ and evaluate the estimator by deriving an asymptotic distribution of the estimator of θ.

When statistical models have the nuisance parameter η, there is an efficiency loss of the estimator of θ. When the efficiency loss is minimized, we deem the estimator of θ to be *semiparametric efficient estimation*. Its properties are discussed in many literatures; for example, Begun *et al.* (1983), Bickel *et al.* (1993), van der Vaart (2000), Tsiatis (2007), and Kosorok (2007). According to the literatures, when the score function of θ is orthogonal to the score function of η, we can minimize the efficiency loss from the nuisance parameter η. To achieve the orthogonality, the *efficient score function* is defined as

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

where I is an identity operator and $\Pi_{\theta,\eta}$ is a projection operator onto a linear space spanned by the score function of η. The inference of θ with the efficient score function enable the semiparametric efficient estimation.

*University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo 113-8654

To evaluate the efficient score function, some researches provide methods for efficient estimation in general semiparametric models without the analytical form of the efficient score function. These researches does not depend on the functional form of the semiparametric models, thus they are acceptable for a wide range of the semiparametric models. Shen (1997) provides a method for semiparametric models when η is estimated by a sieve method and the estimation is to be a maximum likelihood estimation. Severini and Wong (1992) and Ai (1997) provide a method for models that can be decomposed into a specific form. The method by Ai and Chen (2003) can solve a wide range of semiparametric models by estimating η using the sieve method.

Our research provides a method to derive the efficient score function by estimating the projection operator $\Pi_{\theta,\eta}$ without its analytical form. Our method numerically approximates the projection operator in matrix form and derives the semiparametric efficient score function with n observations. When estimating the projection operator, we represent the operator by reproducing kernel function. The usage of the kernel function depends on a theory of *Reproducing kernel Hilbert space* (henceforth RKHS; for details, see Berlinet and Thomas-Agnan (2004)), and we handle the operator for the semiparametric score function by a linear combination of terms with the kernel function.

An advantage of our method is that our method can enable semiparametric efficient estimation without restriction on the estimation method for η, such as using only the sieve method. As long as η is in RKHS, we can accept any nonparametric estimation method for η, such as empirical distribution or Nadaraya-Watson estimation. Thus, our method can realize a better convergence rate for η than the sieve method. Moreover, we can avoid misspecification from the selection of the number of basis for the sieve estimation.

The rest of this paper is organized as follows. Section 2 provides a preliminary theoretical basis for semiparametric estimation, by drawing theories from existing works. Section 3 is the main part of this paper, wherein we provide an estimation method for semiparametric efficient estimation and its theoretical aspects. Section 4 contains numerical experiments, and section 5 concludes. The proofs of all theorems, lemma, and propositions are given in the supplementary materials.

2. Preliminary

As a short, we provide a general scheme of the semiparametric model. This section discusses the formation of semiparametric models, the asymptotic normality of a semiparametric estimator, and the semiparametric efficiency of the same.

2.1 Semiparametric model

Consider a statistical model $\mathcal{P}_{\theta,\eta}(Z)$, where θ and η are parameters. θ is a finite-dimensional parameter $\theta \in \Theta \subset \mathcal{R}^p$, where Θ is a compact parameter space, and p denotes the number of dimensions. η is an infinite-dimensional parameter $\eta \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$. We have n i.i.d. observations $\{Z_i\}_{i=1}^n$ from $\mathcal{P}_{\theta,\eta}(Z)$. For $i = 1, \ldots, n$, $Z_i \in \mathcal{Z}$, where $\mathcal{Z}$ is a sample space. We consider that $m(Z, \theta, \eta) : \mathcal{Z} \times \Theta \times \mathcal{H} \to \mathcal{R}$ is a known deterministic criteria function. The value of the unique true parameters (θ_0, η_0) and the estimator for parameters $(\hat{\theta}, \hat{\eta})$ satisfy the following,

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

In most empirical researches, we treat η as a nuisance parameter, and thus we do not care about the inference of η but are interested in the value of θ. From now on, we discuss the estimation properties of θ.

2.2 Asymptotic properties of the estimator

In this part, we show the asymptotic properties of the estimator of θ, which is the parameter of interest. Mainly, we provide the asymptotic normality of $\hat{\theta}$. There are many approaches to show the asymptotic normality, for instance Bickel *et al.* (1993), van der Vaart (2000) and Tsiatis (2007). In this research, we show formulation based on Kosorok (2007).

We first define the following derivatives,

$$m_1(Z,\theta,\eta) := \frac{\partial}{\partial\theta} m(Z,\theta,\eta),$$
$$m_2(Z,\theta,\eta)[f] := \frac{\partial}{\partial\eta} m(Z,\theta,\eta)[f].$$

The first term $m_1(Z,\theta,\eta)$ is an ordinal partial derivative with respect to θ. The second term $m_2(Z,\theta,\eta)$ is a Gateaux derivative with respect to the infinite-dimensional parameter η. To implement the derivative, a derivative path function $f : \mathcal{Z} \to \mathcal{R}$ is needed. Thus, we obtain

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

In van der Vaart (2000), this is expressed as a *one-dimensional submodel*. The choice of f is arbitrary, and we denote a set of possible f as $\mathcal{A}$. How to select f from $\mathcal{A}$ is the critical point of the semiparameteric efficient estimation, and we will discuss later.

Higher-order derivatives are similarly defined as

$$m_{11}(Z,\theta,\eta) := \frac{\partial}{\partial\theta} m_1(Z,\theta,\eta),$$
$$m_{21}(Z,\theta,\eta)[f] := \frac{\partial}{\partial\theta} m_2(Z,\theta,\eta)[f],$$
$$m_{12}(Z,\theta,\eta)[f] := \frac{\partial}{\partial\eta} m_1(Z,\theta,\eta)[f],$$
$$m_{22}(Z,\theta,\eta)[f_1][f_2] := \frac{\partial}{\partial\eta} m_2(Z,\theta,\eta)[f_1][f_2],$$

where f_1 and f_2 are some derivative path functions in $\mathcal{A}$. Also, denote $\mathbf{f} = (f_1,\ldots,f_p)$ be a set of p path functions, and denote $m_2(Z,\theta,\eta)[\mathbf{f}] = (m_2(Z,\theta,\eta)[f_1],\ldots,m_2(Z,\theta,\eta)[f_p])$ be a p-dimensional vector. $m_{12}(Z,\theta,\eta)[\mathbf{f}], m_{21}(Z,\theta,\eta)[\mathbf{f}]$, and $m_{22}(Z,\theta,\eta)[\mathbf{f}_1][\mathbf{f}_2]$ are defined in the same say as $p\times p$ matrices.

We provide formulation for the estimation of θ. For general semiparametric M-estimators, the estimator is obtained by optimizing the empirical criteria function (2). We define a score function as

$$\tilde{m}(Z,\theta,\eta)[\mathbf{f}] := m_1(Z,\theta,\eta) - m_2(Z,\theta,\eta)[\mathbf{f}], \tag{3}$$

with some set of derivative path functions $\mathbf{f}$. Thus, we obtain the estimator $\hat{\theta}$ by solving the equation

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{m}(Z_i,\theta,\hat{\eta})[\mathbf{f}] = 0,$$

with some $\mathbf{f}$ and substituted estimator $\hat{\eta}$.

Using second-order derivatives, we define a Hesse matrix $H_{\theta,\eta}[f_1,f_2]$ as

$$\begin{aligned} H_{\theta,\eta}[\mathbf{f}_1,\mathbf{f}_2] :=& E\left[m_{11}(Z,\theta,\eta) + m_{12}(Z,\theta,\eta)[\mathbf{f}_2]\right. \\ &\left. + m_{21}(Z,\theta,\eta)[\mathbf{f}_1] + m_{22}(Z,\theta,\eta)[\mathbf{f}_1][\mathbf{f}_2]\right]. \end{aligned} \tag{4}$$

Using previous notation, we now show the conditions for asymptotic normality.

Assumption 1. *$\mathcal{Z}$ is compact and the criteria function $m(Z, \theta, \eta)$ is second-order Gateaux differentiable with respect to θ and η.*

Assumption 2. *Assume*

$$\|\hat{\theta} - \theta_0\| = o_P(1),$$
$$\|\hat{\eta} - \eta_0\|_{\mathcal{H}} = O_P(n^{-c_1}),$$

where c_1 is some positive constant.

Assumption 2 requires the consistency of estimators. In this research, we focus on the asymptotic distribution of the estimators, thus we omit the discussion about the consistency. The sufficient condition to obtain the consistency is discussed in detail in van de Geer (2000) and Kosorok (2007). The convergence rate of $\hat{\eta}$ is discussed in many literatures. For example, in Tsybakov (2009), when η is contained in the Sobolev class with order β, the sieve methods provides $\|\hat{\eta} - \eta_0\|_{\mathcal{H}} = O_P(n^{-\beta/(2\beta+1)})$.

Assumption 3. *For all $\mathbf{f}_1, \mathbf{f}_2 \in \mathcal{A}^p$, the Hessian $H_{\theta_0,\eta_0}[\mathbf{f}_1, \mathbf{f}_2]$ and $E[\tilde{m}(Z, \theta_0, \eta_0)[\mathbf{f}_1]\tilde{m}(Z, \theta_0, \eta_0)[\mathbf{f}_1]^T]$ are invertible and their determinants are finite.*

This assumption is about the regularity condition of the Hesse matrix.

Assumption 4. *Let $\mathcal{M}_n = \{(\theta, \eta) : \|\theta - \theta_0\| = o_P(1), \|\eta - \eta_0\|_{\mathcal{H}} = O_P(n^{-c_1})\}$ 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 of $\mathcal{F}$ with norm $\|\cdot\|$.

Assumption 4 yields that the criteria function $m(Z, \theta, \eta)$ satisfies stochastic equicontinuity in $\mathcal{M}_n$. The class of functions that satisfy the entropy condition is called Donsker, and the criteria function, belonging to the Donsker class, has asymptotic equicontinuity.

Assumption 5. *Let $c_2 > \max\{1, \frac{1}{2c_1}\}$ and δ_n be a sequence of positive constant converges to 0. For all $(\theta, \eta) \in \mathcal{M}_n$ and $f \in \mathcal{A}$,*

$$\begin{aligned}
&|E[\{m_1(Z, \theta, \eta) - m_1(Z, \theta_0, \eta_0)\} - m_{11}(Z, \theta, \eta)(\delta_\theta) - m_{12}(Z, \theta, \eta)[\delta_\eta / \|\delta_\eta\|_{\mathcal{H}}] \\
&= o(\|\delta_\theta\|) + O(\|\delta_\eta\|_{\mathcal{H}}^{c_2}), \\
&|E[\{m_2(Z, \theta, \eta) - m_2(Z, \theta_0, \eta_0)\} - m_{21}(Z, \theta, \eta)[\mathbf{f}](\delta_\theta) - m_{22}(Z, \theta, \eta)[\mathbf{f}][\delta_\eta / \|\delta_\eta\|_{\mathcal{H}}]\|\delta_\eta\|_{\mathcal{H}}]| \\
&= o(\|\delta_\theta\|) + O(\|\delta_\eta\|_{\mathcal{H}}^{c_2}),
\end{aligned}$$

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

Assumption 5 is for local smoothness. This assumption means that the criteria function has higher smoothness in a small ball around the true parameter (θ_0, η_0). This kind of condition is common in semi-parametric literature.

Given the above conditions, the following theorem for asymptotic normality is obtained.

Theorem 1. *Consider the estimator in (2). If assumptions 1,2,3,4, and 5 hold, then $\forall \mathbf{f}_1, \mathbf{f}_2 \in \mathcal{A}^p$,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, H_{0,\mathbf{f}_1,\mathbf{f}_2}^{-1} \Sigma_{\mathbf{f}_1} H_{0,\mathbf{f}_1,\mathbf{f}_2}^{-1}),$$

where

$$H_{0,\mathbf{f}_1,\mathbf{f}_2} = H_{\theta_0,\eta_0}[\mathbf{f}_1, \mathbf{f}_2], \tag{5}$$
$$\Sigma_{\mathbf{f}_1} = E\left[\tilde{m}(Z, \theta_0, \eta_0)[\mathbf{f}_1]\tilde{m}(Z, \theta_0, \eta_0)[\mathbf{f}_1]^T\right]. \tag{6}$$

This result is based on a theorem by Kosorok (2007) and its proof is in appendix.

2.3 Semiparametric efficient estimation

In this part, we review and formalize a theory of semiparametric efficient estimation, that is discussed by van der Vaart (2000), Tsiatis (2007), and other researches in great detail. To simplify, we set $p = 1$, and conduct the semiparametric efficient estimation. In $p > 1$ case, we can trivially expand the following discussion to the multi-dimensional case. Preliminarily, we consider a set of the score function of η as $\mathcal{N} = \{m_2(Z_i, \theta, \eta)[f] | f \in \mathcal{A}\}$, and space $\overline{\text{lin}}(\mathcal{N})$, named a nuisance tangent space.

The score function (3) is affected by the perturbations of θ and η. When conducting the inference of η, the perturbation of η reduces the efficiency of the inference of θ. Thus, we let the score function (3) be orthogonal to the perturbation of η, by the proper choice of f, which is arbitrary in previous part. To enable the efficient estimation, the score function $\tilde{m}(Z, \theta, \eta)[f]$ should be orthogonal to the nuisance tangent space,

$$\tilde{m}(Z, \theta, \eta)[f] \perp\!\!\!\perp m_2(Z, \theta, \eta)[f].$$

We denote f^* as the derivative path that satisfies the above orthogonal condition, and we call $\tilde{m}(Z, \theta, \eta)[f^*]$ as an *efficient score function*. To derive the efficient score function with proper derivative path f^*, it is necessary to obtain a projection operator to the nuisance tangent space.

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

$$B_{\theta,\eta}[f] := m_2(Z, \theta, \eta)[f], \tag{7}$$

where $B : \mathcal{A} \to \overline{\text{lin}}(\mathcal{N})$ is an operator. It is a mapping that takes a derivative path function f as argument, and returns the score function of η. Denote $B^*_{\theta,\eta}$ is an adjoint operator of $B_{\theta,\eta}$. For efficient estimation, the following conditions must hold.

Assumption 6. *The following conditions hold.*

1. $B_{\theta,\eta}$ *is a bounded linear operator.*
2. $B^*_{\theta,\eta} B_{\theta,\eta}$ *is continuously invertible.*
3. $\exists f^*, \forall f \in \mathcal{H},\ E[m_{12}(Z, \theta, \eta)[f] - m_{22}(Z, \theta, \eta)[f][f^*]] = 0.$

From the above assumption, we can denote the projection operator $\Pi_{\theta,\eta} : \mathcal{A} \to \overline{\text{lin}}(\mathcal{N})$ as

$$\Pi_{\theta,\eta} = B_{\theta,\eta}(B^*_{\theta,\eta} B_{\theta,\eta})^{-1} B^*_{\theta,\eta}. \tag{8}$$

Given the projection operator, we can obtain the efficient score function with f^* by the following calculation with some f;

$$\begin{aligned}
\tilde{m}(Z, \theta, \eta)[f^*] &= (I - \Pi_{\theta,\eta})\tilde{m}(Z, \theta, \eta)[f] \\
&= (I - \Pi_{\theta,\eta})(m_1(Z, \theta, \eta) - m_2(Z, \theta, \eta)[f]) \\
&= m_1(Z, \theta, \eta) - m_2(Z, \theta, \eta)[f] - \Pi_{\theta,\eta} m_1(Z, \theta, \eta) + m_2(Z, \theta, \eta)[f] \\
&= (I - \Pi_{\theta,\eta}) m_1(Z, \theta, \eta),
\end{aligned} \tag{9}$$

where I is an identity operator. Using this operation, we can evaluate the score function in the orthogonal complement space of the nuisance tangent space. Thus, $(I - \Pi_{\theta,\eta}) m_1(Z, \theta, \eta)$ is the efficient score function, and we can evaluate it without deriving f^* directly.

Finally, the semiparametric efficient estimation is enabled given the following estimation equation:

$$\frac{1}{n} \sum_{i=1}^{n} (I - \Pi_{\theta,\eta}) m_1(Z_i, \theta, \eta) = 0. \tag{10}$$

The above estimation equation (10) is orthogonal to the score function of η, thus the estimation avoids the efficiency loss from the nuisance parameter η.

2.3.1 Relation with efficient variance bound

For maximum likelihood estimation, $m(Z, \theta, \eta) = \log(p(Z, \theta, \eta))$, and the asymptotic variance from the efficient score estimation corresponds to a semiparametric efficient variance bound. This is discussed in detail in van der Vaart (2000).

3. Proposed method

In this section, we suggest a method to implement semiparametric efficient estimation for general semiparametric models. For the estimation, our method evaluate the efficient score (9) by estimating the projection operator (8), and solve the estimation equation (10) with the efficient score. We evaluate the score operator (7) by representation based on the theory of the reproducing kernel Hilbert space (RKHS). The representation of the score operator enables us to evaluate the projection operator and construct the efficient estimation equation.

3.1 Kernel representation

A representation based on the RKHS is a way of representing a component of Hilbert space by a kernel function. Consider a kernel function satisfying $k(z, z') : \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$, which satisfies symmetry; wherein $\forall z, z', k(z, z') = k(z', z)$ and positive semi-definite; and wherein $\forall z_i, \forall \{c_i\} \in \mathcal{R}, \sum_{i,j} c_i c_j k(z_i, z_j) \geq 0$.

Denote $\mathcal{H}_k$ is an RKHS with kernel k, with an inner product $\langle \cdot, \cdot \rangle$. Aronszajn (1950) shows that kernel k uniquely determines RKHS $\mathcal{H}_k$. The space has properties: (i) $k(\cdot, z) \in \mathcal{H}_k$, (ii) linear hull of $\{k(\cdot, z)\}$ is dense in $\mathcal{H}_k$, and (iii) *reproducing property*, that is $\forall f(\cdot) \in \mathcal{H}_k, f(z) = \langle f, k(\cdot, z) \rangle$. These theories are described in Berlinet and Thomas-Agnan (2004).

Our purpose is to evaluate the projection operator $\Pi_{\theta,\eta}$. First, we estimate the score operator $B_{\theta,\eta}$ by a representation based on the reproducing kernel. We restrict $B_{\theta,\eta}$. Consider some function $f \in \mathcal{H}_k$ and $B_{\theta,\eta}[f]$ with some fixed θ and η. Then, we consider a class of linear operators,

$$\mathcal{T}_k := \left\{ B : \mathcal{H}_k \to \mathcal{H}_k \,\middle|\, B[f](\cdot) = \int_{\mathcal{Z}} \lambda(z) k(\cdot, z) f(z) d\mu(z), \forall f \in \mathcal{H}_k \right\}, \tag{11}$$

with some function $\lambda(z) \in \mathcal{H}_k$ and $\mu(\cdot)$ is a measure of z. This is a part of the Fredholm integral operator. The Fredholm integral operator is an operator with form

$$B[f] = \int_{\mathcal{Z}} \overline{K}(\cdot, z) f(z) d\mu(z),$$

with integral kernel function $\overline{K} : \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$ and where $\overline{K}$ is not necessarily symmetric. When $\overline{K}$ can be decomposed as $\overline{K}(z, z') = \lambda(z) k(z', z)$, the operator defined by $\overline{K}$ belongs to $\mathcal{T}_k$.

Assumption 7. *For all $\theta, \eta \in \mathcal{M}_n$, $B_{\theta,\eta} \in \mathcal{T}_k$.*

This assumption is critical. A linear bounded operator that satisfies this assumption with restriction has a finite-dimensional kernel and cokernel, and accepts the decomposition $\overline{K}(z, z') = \lambda(z) k(z', z)$. A multiplication kernel satisfies the above assumption by reproductivity, and some semiparametric models such as partially linear models satisfy it directly.

With fixed f and reproducing k, we construct an estimator of $B_{\theta,\eta}$ of the following form:

$$\hat{B}_{\theta,\eta}[f] = \sum_{i=1}^{\infty} w'_{\theta,\eta,i} k(\cdot, z_i) f(z_i), \tag{12}$$

where $\{w'_{\theta,\eta,i}\}$ is a set of weights. We now mention the representer theorem described in Schölkopf *et al.* (2001). This theorem guarantees that the sum of n kernels is sufficient to optimize an empirical minimization problem with n observations and penalty. In other words, only n bases can optimize an optimization problem with n observation. The representer theorem justifies the representation of the function by n kernels, and we apply it for the operator.

Drawing from the above, we theoretically show that n kernels are sufficient to estimate an operator is an empirical minimization problem. Define a norm of $B_{\theta,\eta}$ as $\|B_{\theta,\eta}\|_B = \sqrt{\sum_i w^2_{\theta,\eta,i}}$, and increasing penalty function $\Omega(\cdot)$.

Proposition 1. *Consider an optimization problem with n observations and a loss function $l(\cdot) : \mathcal{Z}^n \times \mathcal{T}_k \to \mathcal{R}$,*

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

Then, the minimizer of the problem is represented by $\{f(Z_i)k(Z_i, z)\}_{i=1}^n$.

The proof is in the appendix. Finally, by calculating only n components, $\{w_{\theta,\eta,i}\}_{i=1}^n$, we evaluate operator $B_{\theta,\eta}$, for n observations and an arbitrary f.

3.2 Operator estimation

In this subsection, we show how to estimate $B_{\theta,\eta}$ using the kernel, and represent $\Pi_{\theta,\eta}$. There are three steps: (i) obtain numerical functional derivative values $\{m_2(Z_i, \theta, \eta)[f]\}_{i=1}^n$ for n samples with some function $f \in \mathcal{H}_k$, (ii) estimate the weight $\{w_{\theta,\eta,i}\}_{i=1}^n$ with the derivative coefficients as training data, and (iii) construct $\hat{\Pi}_{\theta,\eta}$ by using $\hat{B}_{\theta,\eta}$ from the estimated $\{w_{\theta,\eta,i}\}_{i=1}^n$. Thus, we solve the efficient estimation problem.

(i) Obtain derivatives: Calculate the functional derivative values $m_2(Z, \theta, \eta)[f]$ numerically with observations $\{Z_i\}_{i=1}^n$ with given (θ, η). Here, the derivative path function f is an arbitrary measurable function. Then, we derive empirical numerical values of the derivative coefficients as

$$\hat{m}_2(Z_i, \theta, \eta)[f] = \lim_{t \to 0} \frac{m(Z_i, \theta, \eta + tf) - m(Z_i, \theta, \eta)}{t},$$

and obtain numerical value of $\{\hat{m}_2(Z_i, \theta, \eta)[f]\}_{i=1}^n$.

(ii) Estimate weights of the estimated operators: We estimate $B_{\hat{\theta},\hat{\eta}}$. In line with the previous proposition, we represent the estimator as

$$\hat{B}[f](z) = \frac{1}{n} \sum_{i=1}^n w_i f(Z_i) k(Z_i, z), \tag{13}$$

where $\{w_i\}_{i=1}^n$ denotes some weights. To obtain the weights, we consider the following problem

$$\min_{\hat{B} \in \mathcal{T}_k} \frac{1}{n} \sum_{i=1}^n \left(\hat{m}_2(Z_i, \theta, \eta)[f] - \hat{B}[f](Z_i) \right)^2 + \kappa_n \Omega(\|B\|_B), \tag{14}$$

where κ_n is a positive finite penalty coefficient, and converges to zero as $n \to \infty$. The penalty term is to avoid overfitting in the estimation problem. Then, this optimization problem can be rewritten as

$$\min_{\{w\}_{j=1}^n} \frac{1}{n} \sum_{i=1}^n \left(\hat{m}_2(Z_i, \theta, \eta)[f] - \frac{1}{n} \sum_{j=1}^n w_j f(Z_j) K(Z_j, Z_i) \right)^2 + \kappa_n \Omega(\|B\|_B). \tag{15}$$

We rewrite this problem in matrix form:

$$Y = \begin{pmatrix} \hat{m}_2(Z_1, \theta, \eta)[f] \\ \vdots \\ \hat{m}_2(Z_n, \theta, \eta)[f] \end{pmatrix}, K = \begin{pmatrix} k(Z_1, Z_1) & \dots & k(Z_1, Z_n) \\ \vdots & \ddots & \vdots \\ k(Z_n, Z_1) & \dots & k(Z_n, Z_n) \end{pmatrix},$$
$$F = \begin{pmatrix} f(Z_1) \\ \vdots \\ f(Z_n) \end{pmatrix}, W = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}.$$

Then, we rewrite problem (14) as

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

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

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

From the optimal conditions of the problem, the minimizer can be written as

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

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

(iii) Construct operators: Provide the estimator of $\Pi_{\theta,\eta}$. We define a weighted gram matrix $G_{\theta,\eta} \in \mathcal{R}^{n \times n}$ with elements

$$G_{\theta,\eta,ij} = k(Z_i, Z_j) \hat{w}_{\theta,\eta,j}.$$

Then, we can evaluate the operator with n observations. Consider the vector of $f(Z_i)$. By multiplying $G_{\theta,\eta}$, we obtain an image of $f(Z_i)$:

$$\begin{pmatrix} \hat{B}_{\theta,\eta}[f](Z_1) \\ \vdots \\ \hat{B}_{\theta,\eta}[f](Z_n) \end{pmatrix} = G_{\theta,\eta} \begin{pmatrix} f(Z_1) \\ \vdots \\ f(Z_n) \end{pmatrix}. \tag{16}$$

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, the element representation of $\hat{\Pi}$ can be given as

$$\hat{\Pi}_{\theta,\eta}[f](Z_i) = \sum_{j=1}^{n} [G_{\theta,\eta} (G^T_{\theta,\eta} G_{\theta,\eta})^{-1} G^T_{\theta,\eta}]_{ij} f(Z_j). \tag{17}$$

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

Using this method, we can obtain an estimator of the efficient score:

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

It enables us to rewrite the optimal condition (10) using matrix $G_{\theta,\hat{\eta}}$,

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}(I - \hat{\Pi}_{\theta,\hat{\eta}})m_1(Z_i,\theta,\hat{\eta}) \\ &= \frac{1}{n}[I_n - G_{\theta,\hat{\eta}}(G_{\theta,\hat{\eta}}^T G_{\theta,\hat{\eta}})^{-1}G_{\theta,\hat{\eta}}^T] \begin{pmatrix} m_1(Z_1,\theta,\hat{\eta}) \\ \vdots \\ m_1(Z_n,\theta,\hat{\eta}) \end{pmatrix}. \end{aligned} \tag{18}$$

Then, let the solution of the equation (18) be $\hat{\theta}^*$, and this is our proposed estimator.

Because the estimator from (18) is Z-estimator, we can improve its efficiency by one-step operation if necessary. For detail, see van der Vaart (2000).

3.3 Theory underlying the method

We provide a theoretical aspect for our method. In this section, we show that $\hat{B}_{\hat{\theta},\hat{\eta}}$ and $\hat{\Pi}_{\hat{\theta},\hat{\eta}}$ are consistent estimators, and that the suggested estimation problem (18) can provide an asymptotic distribution with efficient variance. The proofs of all lemmas and theorems are in the appendix.

The following assumption is made. Denote $\lambda_{\theta,\eta}(z)$ be a function to represent $B_{\theta,\eta} \in \mathcal{T}_k$, in other words, $B_{\theta,\eta}[f] = \int \lambda_{\theta,\eta}k(\cdot,z)f(z)dz$.

Assumption 8. *The following conditions hold.*

1. *Kernel function k is continuous, bounded, and positive semi-definite.*
2. *Let $\Lambda_n = \{\lambda_{\theta,\eta}(z)k(z,z')f(z) : z,z' \in \mathcal{Z}, (\theta,\eta) \in \mathcal{M}_n, f \in \mathcal{H}_k\}$. Then,*

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

The first condition in the above assumption is a general condition for RKHS theory. The second condition requires that $\Lambda_{\theta,\eta}(Z)k(Z,Z')f(Z)$ is Donsker class, discussed in Assumption 4.

Finally, we show the limit distribution of the suggested optimization problem. Preliminary, we denote the Hesse matrix and asymptotic variance using the efficient score. Let

$$H^* := E\left[m_{11}(Z,\theta_0,\eta_0) + m_{22}(X,\theta_0,\eta_0)[f^*][f^*]\right], \tag{19}$$

$$\Sigma^* := E[\tilde{m}(Z,\theta_0,\eta_0)[f^*]\tilde{m}(Z,\theta_0,\eta_0)[f^*]^T], \tag{20}$$

Then, the variance of the asymptotic distribution of $\hat{\theta}$ becomes $H_*^{-1}\Sigma^* H_*^{-1}$ by Theorem 1.

Theorem 2. *Consider the estimation problem (18). If Assumptions 1-8 hold,*

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

where $\kappa_n = o(1/\sqrt{n})$.

The proof is in the appendix.

3.4 Relation with existing methods

A novelty of our proposed method is that our method can enable semiparametric efficient estimation without restriction on the estimation method for the infinite dimensional parameters η. The existing methods, such as Ai and Chen (2003), use the sieve method to estimate η, and implement the semiparametric efficient estimation based on the properties of the sieve method. In contrast, our method focuses on the estimation of the operator, thus we can accept any nonparametric estimation method for η, such as empirical distribution or Nadaraya-Watson estimation. As a result, our method can avoid the difficulty of the sieve method, such as a selection problem of the number of basis functions, and also can enjoy the benefit of other nonparametric estimation methods.

4. Numerical Experiment

In this section, we provide results of numerical experiments to show the effectiveness of our method. Throughout this section, we call our proposed method with the efficient score function as 'our kernel method', and the estimation without the efficient score function as 'the ordinal method'. Purpose of this section is to illustrate that our kernel method reduces the variance of the estimator than the ordinal method. We conduct the experiments using some specific semiparametric models, such as the partially linear model and the copula model.

4.1 Partially linear model

The partially linear model is a typical semiparametric model. Let $Z = (Y, X_1, X_2)$, where Y is a respondent variable, and X_1 and X_2 are covariates. We assume that the observation is generated 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. When $\eta(\cdot)$ is nonlinear, this model can treat a nonlinear relationship between X_2 and Y. Since $\eta(\cdot)$ is an infinite-dimensional parameter, we estimate $\eta(\cdot)$ using some nonparametric method.

Usually, the parameters are estimated using the least square method as

$$(\hat{\theta}, \hat{\eta}) = \arg\min \frac{1}{n} \sum_{i=1}^{n} \left[y_i - x_{1,i}^T \theta - \eta(x_{2,i}) \right]^2. \tag{21}$$

The estimation of $\hat{\eta}(\cdot)$ is implemented by the sieve method or the Nadaraya-Watson method.

We estimate θ using the ordinal method and our proposed kernel method. In the ordinal method, we estimate η by the linear sum of polynomial sieve functions and estimate θ based on $\hat{\eta}$. We construct the total score function $\tilde{m}$ of the partially linear model by differentiating the loss function in (21). Our kernel method also uses the same method to estimate $\eta(\cdot)$, and we construct the efficient estimation equation (18), which is based on $\tilde{m}$ obtained from the ordinary method and a Gaussian kernel. To implement the kernel method, we select the tuning parameters $h = 1.0$ and $\kappa_n = 5.0$.

We replicate the estimation 200 times, for different values of θ and for different sample sizes. The mean and variance of the estimator of θ are in Table 1. For every sample size and parameter value, both methods provide consistent estimators. Variance is plotted in Figure 1, and the proposed kernel method reduces the variance of the estimator.

n	Ordinal:Mean(Var)	Kernel:Mean(Var)
$\theta = 1.0$		
100	0.970(0.102)	0.982(0.085)
200	0.997(0.047)	0.995(0.041)
300	0.995(0.037)	1.000(0.031)
400	1.007(0.026)	1.004(0.024)
500	0.999(0.027)	1.000(0.023)
$\theta = 3.0$		
100	3.008(0.122)	2.996(0.105)
200	3.007(0.051)	3.004(0.041)
300	3.008(0.036)	3.004(0.033)
400	3.017(0.021)	3.013(0.020)
500	2.997(0.024)	2.996(0.020)
$\theta = 5.0$		
100	4.957(0.080)	4.958(0.079)
200	5.013(0.045)	5.013(0.044)
300	5.033(0.035)	5.032(0.035)
400	4.970(0.025)	4.970(0.024)
500	5.000(0.019)	5.000(0.018)

Table 1: Mean and variance of the estimator of θ in partially linear model, from replication for 200 times with each different true value of θ. The variance of the estimator is in parentheses.

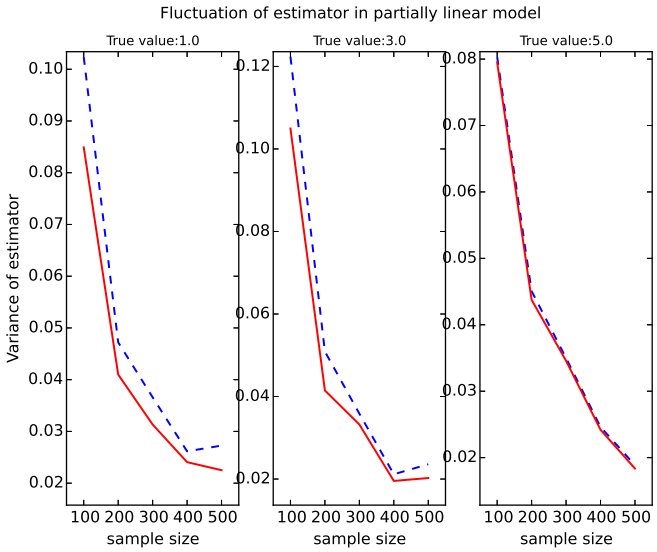

Figure 1: Variance of the estimator of θ in partially linear model, from replication for 200 times with each different true value of θ. Numerical value is from table 1. Blue dashed line is by the ordinal method, and red solid line is by the proposed kernel method.

4.2 Copula model

We conduct an experiment using a copula model with a Clayton-type copula function. The copula model is used for representing a relationship between variables. In this case, we consider the copula model for two

variables. Let $Z = (X_1, X_2)$. The correlation of X_1 and X_2 is represented by a joint distribution function. Let $C(x_1, x_2; \theta) \to \mathcal{X} \times \mathcal{X} \times \Theta \to [0, 1]$ be a copula function and let the joint distribution function be

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

where $F_j(\cdot)$ is a univariate distribution function of X_j. Functional forms have been suggested as the copula function, and an example would be the Clayton-Cook-Johnson function of form

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

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

The parameter of interest θ is estimated using the maximum likelihood function as

$$\hat{\theta} = \arg\max \frac{1}{n} \sum_{i=1}^{n} \log c(\hat{F}_1(x_{1,i}), \hat{F}_2(x_{2,i}); \theta). \tag{22}$$

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

We now provide our proposed method. We estimate θ using the ordinal method and our proposed kernel method. In the ordinal method, we estimate $F_1(x_1)$ and $F_2(x_2)$ using an empirical distribution estimator, and then substitute them into the copula function to estimate θ. We construct the total score function $\tilde{m}$ of the copula model by differentiating the loss function in (22). Our kernel method also uses the same method to estimate $F_1(x_1)$ and $F_2(x_2)$, and we construct an efficient estimation equation (18), which is based on $\tilde{m}$ obtained from the ordinary method and a Gaussian kernel. To implement the kernel method, we select the tuning parameters $h = 1.0$ and $\kappa_n = 5.0$.

We replicate the estimation 200 times, for different values of θ and for different sample sizes. The mean and variance of the estimators of θ are in Table 2. When θ is small, both methods provide consistent estimators. As θ increases, both estimators develop a bias of almost the same size. Variance is plotted in Figure 2, the proposed kernel method reduces the variance of the estimator.

n	Ordinal:Mean(Var)	Kernel:Mean(Var)
$\theta = 3.0$		
100	3.676(0.231)	3.615(0.167)
200	3.732(0.122)	3.662(0.090)
300	3.705(0.079)	3.610(0.043)
400	3.737(0.063)	3.654(0.049)
500	3.768(0.056)	3.608(0.038)
$\theta = 4.0$		
100	4.642(0.402)	4.346(0.306)
200	4.566(0.218)	4.024(0.141)
300	4.703(0.132)	4.085(0.106)
400	4.694(0.102)	4.101(0.068)
500	4.734(0.079)	4.152(0.048)
$\theta = 5.0$		
100	5.393(0.668)	5.120(0.620)
200	5.559(0.241)	4.799(0.225)
300	5.587(0.168)	4.619(0.134)
400	5.593(0.145)	4.550(0.084)
500	5.659(0.143)	4.654(0.085)

Table 2: Mean and variance of the estimator of θ in copula model, from replication for 200 times with each different true value of θ. The variance of the estimator is in parentheses.

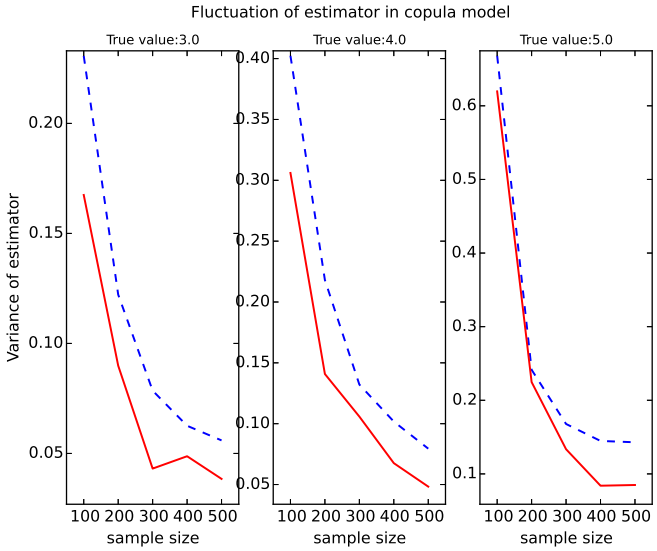

Figure 2: Variance of the estimator θ in copula model, from replication for 200 times with each different true value of θ. Numerical value is from table 2. Blue dashed line is by the ordinal method, and red solid line is by the proposed kernel method.

5. Conclusion

We provide a method that enalbes semiparametric efficient estimation for a wide range of semiparametric models. When the ordinal estimator of θ has root-n consistency and asymptotic normality, our method enables semiparametric efficient estimation with some assumptions. Unlike other methods, our method does not restrict how one estimates η. Thus, our method is applicable to a wide range of semiparametric models.

References

Ai, C. (1997) A semiparametric maximum likelihood estimator, *Econometrica*, pp. 933–963.

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

Aronszajn, N. (1950) Theory of reproducing kernels, *Transactions of the American mathematical society*, pp. 337–404.

Begun, J. M., Hall, W., Huang, W.-M., Wellner, J. A. *et al.* (1983) Information and asymptotic efficiency in parametric-nonparametric models, *The Annals of Statistics*, **11**, 432–452.

Berlinet, A. and Thomas-Agnan, C. (2004) *Reproducing kernel Hilbert spaces in probability and statistics*, vol. 3, Springer.

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

Cox, D. (1972) Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220.

Härdle, W. and Liang, H. (2007) *Partially linear models*, Springer.

Ichimura, H. (1993) Semiparametric least squares (sls) and weighted sls estimation of single-index models, *Journal of Econometrics*, **58**, 71–120.

Kosorok, M. R. (2007) *Introduction to empirical processes and semiparametric inference*, Springer Science & Business Media.

Schölkopf, B., Herbrich, R. and Smola, A. J. (2001) A generalized representer theorem, in *Computational learning theory*, Springer, pp. 416–426.

Severini, T. A. and Wong, W. H. (1992) Profile likelihood and conditionally parametric models, *The Annals of Statistics*, pp. 1768–1802.

Shen, X. (1997) On methods of sieves and penalization, *The Annals of Statistics*, pp. 2555–2591.

Tsiatis, A. (2007) *Semiparametric theory and missing data*, Springer.

Tsybakov, A. B. (2009) Introduction to nonparametric estimation, *Springer Series in Statistics*.

van de Geer, S. A. (2000) *Empirical Processes in M-estimation*, vol. 105, Cambridge university press Cambridge.

van der Vaart, A. W. (2000) *Asymptotic statistics*, vol. 3, Cambridge university press.