A Paradox of Semiparametric Estimators with Infinite Dimensional Nuisance Parameters

K. Hitomi[a] **and Y. Nishiyama**[b]

[a]Kyoto Institute of Technology, Kyoto, Japan, E-mail: hitomi@kit.ac.jp
[b]Kyoto Institute of Economic Research, Kyoto University, Kyoto, Japan

Keywords: *semiparametric estimation, nonparametric nuisance parameter, tangent space, efficiency*

EXTENDED ABSTRACT

[1]Recently, a paradoxical phenomenon of semiparametric estimators was found that some semiparametric estimators are more efficient when infinite dimensional nuisance parameters are unknown. This paper examined the structure of the paradox.

Pierce (1982) found a paradoxical phenomenon. Let $\theta' = (\beta', \gamma')$ be parameters which we would like to estimate. In many cases, we are only interested in some parameters and the rest is nuisance parameters. Let β be parameters we were interested in and γ be nuisance parameters. Usually, an estimator of β has smaller variance when the nuisance parameter γ is known. Pierce (1982) found that under some conditions the variance of estimator of β with *unknown* γ is smaller than the one with *known* γ.

Robins *et al.* (1992), Robins *et al.* (1994) and Lawless *et al.* (1999) reported the same phenomenon in their semiparametric models. Henmi (2004) and Henmi and Eguchi (2004) investigated the paradox in semiparametric situations. Henmi (2004) and Henmi and Eguchi (2004) investigated the following problem. Let $\mathcal{M} = \{p(z; \theta, g)\}$ be a family of probability distribution, θ is a finite dimensional parameter and g is an infinite dimensional nuisance parameter. The parameter θ is composed of two parts, a parameter of interest β and a nuisance parameter γ. There is a moment condition

$$E(m(z, \theta)) = 0,$$

and θ could be estimated by solving

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\theta}) = 0.$$

Note that the estimation of the infinite dimensional parameter g is not required to be estimated in this setting. Under the above settings, Henmi (2004) and Henmi and Eguchi (2004) examined the phenomenon that the variance of estimator of β with *unknown* γ is smaller than the one with *known* γ. To summarize, they investigated only the situation that estimations of infinite dimensional nuisance parameters were not necessary.

The above setting does not suffice to analyze many semiparametric models. Many semiparametric estimators require estimation of infinite dimensional nuisance parameters. Estimation of infinite dimensional parameters (for example, density estimations or nonparametric regressions) is more difficult than estimation of finite dimensional parameters and the convergence rate is slower than finite dimensional cases.

This paper investigated the paradoxical phenomenon that the efficiency of estimator increased when the nuisance parameters are estimated. We found that the paradox might occur even if the nuisance parameters were infinite dimensional. We obtained the necessary and sufficient condition of the paradox, and showed a sufficient condition in an easy-to-understand way. Our sufficient condition depended on the relation of the projection of the estimating function on the tangent set.

In the last section, two semiparametric estimators, Kaplan-Meier integral and the average treatment effect, were examined. We found that both of the examples satisfied our sufficient condition of the paradox.

[1]This research was partially supported by Grant-in-Aid for Scientific Research (KAKENHI 14530031, 15330040) of Japan Society for the Promotion of Science.

1 INTRODUCTION

This paper investigated the following question. Suppose our estimation problem includes infinite dimensional nuisance parameters and we need to estimate the nuisance parameters. Is it possible that the variance of the parameters of interest with estimated nuisance parameters is smaller than the one with known nuisance parameters?

The following is the organization of the paper. The section 2 explains our model and the settings. In section 3, we state the conditions under which the paradox occurs. The section 4 includes examples. The section 5 is the concluding remarks.

2 MODEL

We investigate the following problem. Let $\{z_i | i = 1, \dots, n\}$ be an i.i.d. sample from a population and the following moment condition is satisfied,

$$E[m(z_i, \beta_0, g_0)] = 0, \tag{1}$$

where $m(\dots)$ is a $q \times 1$ vector valued function, β_0 is $q \times 1$ vector of parameters of interest and g_0 is some infinite dimensional nuisance parameters. For simplicity, we assume that g_0 does not depend on the finite dimensional parameter β. All examples in the later session satisfy this assumption.

If we know true g_0, β could be estimated by solving

$$\frac{1}{n}\sum_{i=1}^{n} m(z_i, \tilde{\beta}, g_0) = 0. \tag{2}$$

In general, g_0 is unknown, however, an estimator based on $m(z_i, \beta, g_0)$ is not feasible. A feasible version might be based on a preliminary estimate $\hat{g}$ of g_0. We investigated a class of semiparametric estimators which are satisfied

$$\frac{1}{n}\sum_{i=1}^{n} m(z_i, \hat{\beta}, \hat{g}) = 0. \tag{3}$$

Many semiparametric estimators are included in this class. An important example is Robinson's (1988) estimator for the semiparametric regression model, where $g_0 = (E[y|v], E[x|v]) \equiv (g_{10}(v), g_{20}(v))$, $m(z, \beta, g) = (x - g_2(v))(y - g_1(v) - (x - g_2(v))'\beta)$.

3 MAIN RESULTS

In order to study the limiting distribution with estimated nonparametric components, impose the following high-level assumptions.

Assumption 1 (i) $\hat{\beta} \xrightarrow{p} \beta_0$; (ii) $m(z, \beta, g)$ is continuously differentiable with respect to β; (iii) for any $\bar{\beta} \xrightarrow{p} \beta_0$, $\partial(1/n)\sum_{i=1}^{n} m(z_i, \bar{\beta}, \hat{g})/\partial\beta \xrightarrow{p} M \equiv \partial E[m(z, \beta, g_0)]/\partial\beta|_{\beta=\beta_0}$ and M is nonsingular. (iv) g_0 does not depend on the finite dimensional parameter β.

Conditions (i) and (iii) assure of the consistency of $\hat{\beta}$ and $\partial(1/n)\sum_{i=1}^{n} m(z_i, \bar{\beta}, \hat{g})/\partial\beta$ that might require much work to check in a particular model. Condition (ii) is imposed for simplicity and it might be possible to weaken (ii).

Suppose $(1/\sqrt{n})\sum_{i=1}^{n} m(z_i, \beta_0, \hat{g})$ is bounded in probability, then usual mean value expansion gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = -M^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i, \beta_0, \hat{g}) + o_p(1).$$

Here calculating the asymptotic distribution reduces to finding a formula for the distribution of $(1/\sqrt{n})\sum_{i=1}^{n} m(z_i, \beta_0, \hat{g})$.

Assumption 2

(i)

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, \hat{g}) - E[m(z_1, \beta_0, \hat{g})] - \frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, g_0)\right\} = o_p(1);$$

(ii)

$$\sqrt{n}E[m(z_1, \beta_0, \hat{g})] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tau_i + o_p(1),$$

where $E[\tau_i \tau_i'] < \infty$ and $E[\tau_i] = 0$.

Note that the expectations in (i) and (ii) are taken for given $\hat{g}$.

Condition (i) is a type of stochastic equicontinuity conditions. Conditions (i) and (ii) imply that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i, \beta_0, \hat{g}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(m(z_i, \beta_0, g_0) + \tau_i) + o_p(1),$$

thus τ_i is a correction term for the effect of the estimation of $\hat{g}$.

With Assumption 1 and 2, we get the following lemma.

Lemma 1. (Newey (1990) lemma A.3) *If Assumption 1 and 2 are satisfied, then $\hat{\beta}$ is asymptotically linear with influence function $-M^{-1}(m(z_i, \beta_0, g_0) + \tau_i)$.*

Proof. By Assumption 1(ii) and the mean value theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,\hat{g}) + M\sqrt{n}(\hat{\beta}-\beta_0) = o_p(1). \quad (4)$$

Also, by Assumption 2,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,\hat{g}) \\
= \;& \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n} m(z_i,\beta_0,\hat{g}) - E[m(z_i,\beta_0,\hat{g})] \right. \\
&\left. -\frac{1}{n}\sum_{i=1}^{n} m(z_i,\beta_0,g_0)\right\} \\
&+\sqrt{n}E[m(z_i,\beta_0,\hat{g})] + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,g_0) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(m(z_i,\beta_0,g_0)+\tau_i) + o_p(1). \qquad (5)
\end{aligned}$$

so by the central limit theorem, $(1/\sqrt{n})\sum_{i=1}^{n} m(z_i,\beta_0,\hat{g})$ is bounded in probability. Therefore, by equation (5), solving equation (4) for $\sqrt{n}(\hat{\beta}-\beta_0)$ gives

$$\begin{aligned}
\sqrt{n}(\hat{\beta}-\beta_0) &= -M^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,\hat{g}) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-M^{-1}(m(z_i,\beta_0,g_0)+\tau_i)\right\} \\
&\quad +o_p(1).
\end{aligned}$$

□

Note that Assumption 2 is required for linearizing $m(z_i,\beta_0,\hat{g})$ with respect to g. Thus, if another linearization method is available, Assumption 2 is not required.

If we know the true g_0, β could be estimated by solving (2). Let $\tilde{\beta}$ denote this estimator that utilizes the information of true g_0. The influence function of $\tilde{\beta}$ is $-M^{-1}m(z_i,\beta_0,g_0)$ and

$$\sqrt{n}(\tilde{\beta}-\beta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M^{-1}m(z_i,\beta_0,g_0).$$

For simplifying notations, let m_i denote $m(z_i,\beta_0,g_0)$. Suppose the following decomposition of m_i,

$$m_i = A\tau_i + u_i,$$

where A is $E(m_i\tau_i')\{E(\tau_i\tau_i')\}^{-1}$ and $u_i = m_i - A\tau_i$. With these notations, the influence functions of $\tilde{\beta}$ and $\hat{\beta}$ are

$$\begin{aligned}
-Mm_i &= -M(u_i + A\tau_i) \\
-M(m_i+\tau_i) &= -M(u_i + (A+I)\tau_i)
\end{aligned}$$

and the asymptotic variances could be expressed as

$$\begin{aligned}
V(\sqrt{n}\tilde{\beta}) &= M^{-1}\left[E(u_iu_i') + AE(\tau_i\tau_i')A'\right]M'^{-1} \\
V(\sqrt{n}\hat{\beta}) &= M^{-1}\left[E(u_iu_i')\right. \\
&\quad \left.+(A+I)E(\tau_i\tau_i')(A+I)'\right]M'^{-1},
\end{aligned}$$

where I is $q\times q$ identity matrix. Thus necessary and sufficient condition of the paradox is

$$\begin{aligned}
& V(\sqrt{n}\hat{\beta}) \le V(\sqrt{n}\tilde{\beta}) \\
\iff\; & (A+I)E(\tau_i\tau_i')(A+I)' \le AE(\tau_i\tau_i')A'.
\end{aligned}$$

We get the following theorem.

Theorem 2. *Suppose Assumption 1 and 2 are satisfied. Let $A\tau_i$ be the linear projection of $m(z_i,\beta_0,g_0)$ on τ_i. The necessary and sufficient condition of that the asymptotic variance of $\hat{\beta}$ is smaller than the asymptotic variance of $\tilde{\beta}$ is*

$$(A+I)E(\tau_i\tau_i')(A+I)' \le AE(\tau_i\tau_i')A'. \quad (6)$$

If the dimension of interested parameter β is one, the condition becomes a bit simpler.

Corollary 3. *Suppose Assumption 1 and 2 are satisfied and $q=1$. Let $a\tau_i$ be the linear projection of $m(z_i,\beta_0,g_0)$ on τ_i. If $a\le -1/2$, then the asymptotic variance of $\hat{\beta}$ is smaller than the asymptotic variance of $\tilde{\beta}$.*

3.1 A sufficient condition

One of sufficient conditions for (6) is

$$A = -I,$$

it implies that

$$\text{Proj}(m_i|\tau_i) = -\tau_i.$$

If this sufficient condition is satisfied, the influence function of $\hat{\beta}$ becomes $-Mu_i = -M(m_i - \text{Proj}(m_i|\tau_i))$.

For illustrating the sufficient condition, suppose a finite dimensional nuisance parameter case. The finite dimensional case provides insight concerning the infinite dimensional case. Suppose γ is a finite dimensional nuisance parameter vector. And $\hat{\gamma}$ is an estimator of γ solving $\sum_{i=1}^{n} h(z_i,\hat{\gamma}) = 0$. Let $\hat{\beta}$ be an estimator of β that satisfies the following moment condition,

$$\frac{1}{n}\sum_{i=1}^{n} m(z_i,\hat{\beta},\hat{\gamma}) = 0.$$

Let $\tilde{\beta}$ denote an estimator of β when γ is known to γ_0, thus it satisfies

$$\frac{1}{n}\sum_{i=1}^{n} m(z_i,\tilde{\beta},\gamma_0) = 0.$$

It is well known that under general conditions the distribution of $\tilde{\beta}$ is

$$\begin{aligned}\sqrt{n}(\tilde{\beta}-\beta_0) &= -\left(E\left(\frac{\partial m}{\partial \beta}\right)\right)^{-1} \\ &\quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,\gamma_0) \\ &\quad +o_p(1) \qquad (7)\end{aligned}$$

Under general regularity conditions, the distribution of $\hat{\beta}$ is

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= -\left(E\left(\frac{\partial m}{\partial \beta}\right)\right)^{-1} \\ &\quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,\hat{\gamma}) + o_p(1),\end{aligned}$$

$(1/\sqrt{n})\sum m(z_i,\beta_0,\hat{\gamma})$ could be decomposed as the following,

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,\hat{\gamma}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i,\beta_0,\gamma_0) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(z_i,\beta_0,\gamma_0)}{\partial\gamma} \\ &\quad \sqrt{n}(\hat{\gamma}-\gamma_0)+o_p(1)\end{aligned}$$

and

$$\sqrt{n}(\hat{\gamma}-\gamma_0) = \left(E\left(\frac{\partial h}{\partial \gamma}\right)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(z_i,\gamma_0)+o_p(1).$$

A more illuminating expression can be obtained from the generalized information matrix equality

$$\begin{aligned}E\left(\frac{\partial m}{\partial \beta}\right) = -E\left(mS'_\beta\right), E\left(\frac{\partial m}{\partial \gamma}\right) = -E\left(mS'_\gamma\right) \\ E\left(\frac{\partial h}{\partial \gamma}\right) = -E\left(hS'_\gamma\right) \quad \text{and so on,}\end{aligned}$$

where S_β and S_γ are scores for β and γ, respectively. With the above expressions, we get

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= \left(E\left(mS'_\beta\right)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(z_i,\beta_0,\gamma_0) \\ &\quad -E\left(mS'_\gamma\right)\left(E\left(hS'_\gamma\right)\right)^{-1}h(z_i,\gamma_0)\} \\ &\quad +o_p(1). \qquad (8)\end{aligned}$$

Usually, the additional term $E\left(mS'_\gamma\right)\left(E\left(hS'_\gamma\right)\right)^{-1}h(z_i,\gamma_0)$ makes the variance of $\hat{\beta}$ lager than the variance of $\tilde{\beta}$. Suppose that $\hat{\gamma}$ is efficiently estimated, thus $h(z_i,\gamma) = S_\gamma(z_i,\gamma)$, then the second term in parentheses becomes

$$\begin{aligned}& E\left(mS'_\gamma\right)\left(E\left(hS'_\gamma\right)\right)^{-1}h(z_i,\gamma_0) \\ =\ & E\left(mS'_\gamma\right)\left(E\left(S_\gamma S'_\gamma\right)\right)^{-1}S_\gamma(z_i,\gamma_0).\end{aligned}$$

It is the projection of $m(z_i,\beta_0,\gamma_0)$ on S_γ. Thus, if γ is efficiently estimated,

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= \left(E\left(mS'_\beta\right)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(z_i,\beta_0,\gamma_0) \\ &\quad -\mathrm{Proj}(m(z_i,\beta_0,\gamma_0)|S_\gamma)\}+o_p(1).\end{aligned}$$

To extend this result to semiparametric models, we focus on a parametric submodel and a tangent set. A parametric submodel is a parameterization of g, say $g(\gamma)$, such that $g(\gamma_0) = g_0$ for some γ_0. We can construct the likelihood function or scores of the parametric submodels. An example of the parameterization is such that $g(\gamma) = (1-\gamma)g_0 + \gamma g_1$. A tangent set in nonparametric direction is the mean-square closure of linear combinations of scores S_γ for the nonparametric component of parametric submodels;

$$\begin{aligned}\mathcal{T} &= \{t\in R^q : E[\|t\|^2]<\infty, \\ &\quad \exists B_j S_{\gamma j}\ \text{with}\ \lim_{j\to\infty}E[\|t-B_jS_{\gamma j}\|^2]=0\}.\end{aligned}$$

In the finite dimensional case, the projection of m on the score of the nuisance parameter γ plays the most important role in the paradox. In the semiparametric case, the projections on the tangent set $\mathcal{T}$ have the similar role.

Let $u \equiv m - \mathrm{Proj}(m|\mathcal{T})$ be the residual from the projection of m on $\mathcal{T}$ and $S \equiv S_\beta - \mathrm{Proj}(S_\beta|\mathcal{T})$ be the residual from the projection of S_β on $\mathcal{T}$. The following lemma is a special case of Theorem 4.3 in Newey (1993).

Lemma 4. *Suppose Assumptions 1 and 2 are satisfied and $\hat{\beta}$ is regular. If $\forall i \quad \tau_i \in \mathcal{T}$, $\hat{\beta}$ has an influence function $-M^{-1}u$.*

Proof. Since g_0 does not depend on β, τ_i also does not depend on β. It implies $E(\partial\tau_i/\partial\beta) = E(\tau_i S_\beta) = 0$. Thus $E(uS) = E(mS_\beta)$. The rest of proof is the same as the proof of Theorem 4.3 of Newey (1990). □

A direct implication of the lemma is the following.

Theorem 5. *Suppose Assumptions 1 and 2 are satisfied and $\hat{\beta}$ is regular. If $\forall i \quad \tau_i \in \mathcal{T}$, then*

$$V(\sqrt{n}\hat{\beta}) \le V(\sqrt{n}\tilde{\beta})$$

and a strict inequality holds when $\mathrm{Proj}(m|\mathcal{T}) \neq 0$.

As noted before, Assumption 2 is required only for establishing (5). Thus, another method for (5) is available, Theorem 5 does not require Assumption 2.

4 EXAMPLES

4.1 Kaplan-Meier integral

Kaplan-Meier integral is an estimator of

$$\int \psi(x)dF(x),$$

where $\psi(x)$ is a known function and $F(x)$ is an unknown distribution function of non-negative random variable x. When x is randomly right censored, a natural estimator of $\int \psi dF$ is obtained by plugging the Kaplan-Meier estimator $\hat{F}(x)$ into $F(x)$, which is called a Kaplan-Meier integral.

Let us define some notations for explaining this example. Let $\{X_i | i = 1, \ldots, n\}$ be i.i.d. positive random variables with a distribution function F. Let $\{Y_i | i = 1, \ldots, n\}$ be i.i.d. positive random variables with a distribution function G and independent of X's. The Y's represent censoring time and G is a censoring distribution. In the randomly right censored data, the pairs (X_i, Y_i), $i = 1, \ldots, n$ are not observed. One observes the pairs (Z_i, δ_i), $i = 1, \ldots, n$, where

$$Z_i = \min(X_i, Y_i) \text{ and } \delta_i = 1(X_i \le Y_i),$$

with $1(.)$ denoting the indicator function.

Suzukawa (2004) found that the Kaplan-Meier integral could be represented as follow,

$$\int \psi d\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \psi(Z_i)}{1 - \hat{G}(Z_i -)} \equiv \hat{\beta}, \qquad (9)$$

where $\hat{G}$ is the Kaplan-Meier estimator of G. So, if G is *known*, we could utilize this information and could construct an estimator

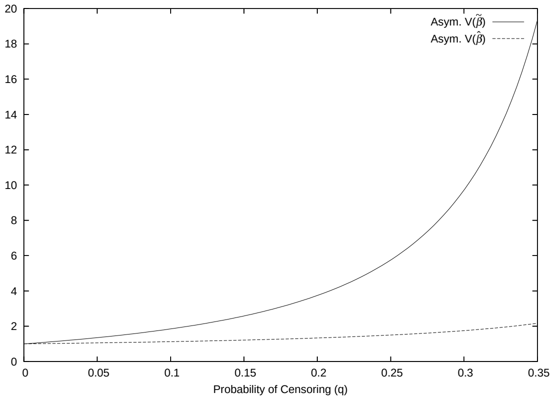
$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \psi(Z_i)}{1 - G(Z_i -)}. \qquad (10)$$

Suzukawa (2004) investigated the estimator $\tilde{\beta}$ and found that $\tilde{\beta}$ has the smaller bias, however, it has the larger variance than $\hat{\beta}$. He constructed the following numerical example. Let $F(x) = 1 - \exp(-x)$, $G(y) = 1 - \lambda \exp(-\lambda y)$ and $\psi(x) = x$. Denote q be the censoring probability $q = P(\delta = 0)$, then $q = \lambda/(1 + \lambda)$. Figure 1 shows the asymptotic variances of $\hat{\beta}$ and $\tilde{\beta}$ in this settings. It might be surprising that the asymptotic variance of the estimator with *know* G is about 10 times lager than the asymptotic variance of the estimator with *unknown* G.

Based on the Suzukawa's representation, define the following two step estimator which satisfies

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \delta_i, \hat{\beta}, \hat{G}) = 0,$$

Figure 1. Asymptotic Variance of $\hat{\beta}$ and $\tilde{\beta}$



where $m(Z_i, \delta_i, \beta, \hat{G}) = (\delta_i \psi(Z_i))/(1 - \hat{G}) - \beta$ and

$$1 - \hat{G}(z) \equiv \prod_{i=1}^{n} \left(1 - \frac{1 - \delta_{(i)}}{n - i + 1}\right)^{1(Z_{(i)} \le z)}$$

is the Kaplan-Meier estimator of G. Hence, the Kaplan-Meier integral is included in our two step estimator framework. We examine that the Suzukawa's Kaplan-Meier integral estimator satisfies the conditions for Theorem 5. The following conditions are assumed.

Assumption 3: (i) the supports of X and Y are both $[0, \infty)$ and F and G have density functions f and g, respectively. (ii) $E[(\delta\psi(z)/(1 - G(z-)))^2] < \infty$, (iii) $\int |\psi(x)| C^{1/2}(x) dF(x) < \infty$, where

$$C(x) = \int_0^{x-} \frac{1}{(1 - H(y))(1 - G(y))} dG(y),$$

and $H(z)$ is the distribution function of the observed Z's.

Condition (i) is just for simplicity and it is easy to weaken. Condition (ii) is corresponding to the existence of variance of $\psi(z)$. Condition (iii) controls the bias, and Stute (1994) gives a detailed account of this issue.

Basically we need to check the following conditions,

1. the linearity of $\sqrt{n}E(m(z_i, \beta_0, \hat{G}))$: $\sqrt{n}E(m(z_i, \beta_0, \hat{G})) = (1/\sqrt{n})\sum_{i=1}^{n} \tau_i + o_p(1)$,
2. the efficiency condition: $\tau_i \in \mathcal{T}_{\mathcal{G}}$, where $\mathcal{T}_{\mathcal{G}}$ is the tangent set with respect to G.

The form of $\mathcal{T}_{\mathcal{G}}$ is well known (for example see Chapter 6.6 of Bickel, Klassen, Ritov and Wellner

(1993)) and it is

$$\begin{aligned}\mathcal{T}_{\mathcal{G}} &= \left\{(1-\delta)b(z) + \delta\frac{1}{1-G(z)}\int_z^\infty b(x)dG(x)\right| \\ &\qquad \forall b(x) \quad \int b(x)dG(x) = 0\Bigg\} \\ &\subset L_2(G).\end{aligned}$$

First, we examine the linearity condition. Let $H(z)$ denote the distribution function of the observed Z's. Thus, H satisfies that $1 - H = (1-F)(1-G)$. An important role in our analysis will be played by the subdistribution functions

$$\begin{aligned}H^0(z) &= P(Z \le z, \delta = 0) = \int_0^z (1-F(y))dG(y) \\ H^1(z) &= P(Z \le z, \delta = 1) = \int_0^z (1-G(y-))dF(y).\end{aligned}$$

Let $H_n^0(z)$ and $H_n^1(z)$ denote the empirical version of $H^0(z)$ and $H^1(z)$, respectively,

$$\begin{aligned}H_n^0(z) &= \frac{1}{n}\sum_{i=1}^n (1-\delta_i)1(Z_i \le z) \\ H_n^1(z) &= \frac{1}{n}\sum_{i=1}^n \delta_i 1(Z_i \le z).\end{aligned}$$

$E(m(z, \delta, \beta_0, \hat{G}))$ could be decomposed as the following,

$$\begin{aligned}E(m(z, \delta, \beta, \hat{G})) &= \frac{1}{n}\sum_{i=1}^n \{(1-\delta_i)(\gamma_1(Z_i) - \gamma_2(Z_i)) \\ &\quad -\delta_i\gamma_2(Z_i)\} + o_p(n^{-1/2}),\end{aligned}$$

where $\gamma_1(Z_i)$ and $\gamma_2(Z_i)$ are

$$\begin{aligned}\gamma_1(Z_i) &= \frac{1}{1-H(Z_i)}\int_z 1(Z_i < z)\psi(z)\gamma_0(z)dH^1(z) \\ \gamma_2(Z_i) &= \int_x\int_z \frac{1(x < Z_i)1(x < z)\psi(z)\gamma_0(z)}{(1-H(x))^2} \\ &\quad dH^0(x)dH^1(z),\end{aligned}$$

and $\gamma_0(z) = 1/(1-G(z-))$.

With simple integration, it could be shown that

$$\int \{\gamma_1(x) - \gamma_2(x)\}dG(x) = 0,$$

and

$$\frac{1}{1-G(Z_i-)}\int_{Z_i}^\infty (\gamma_1(w) - \gamma_2(w))dG(w) = -\gamma_2(Z_i).$$

It also possible to show that $E[\tau_i'\tau_i] < \infty$. It implies that $\tau_i \equiv (1-\delta_i)(\gamma_1(Z_i) - \gamma_2(Z_i)) - \delta_i\gamma_2(Z_i)$ is included in $\mathcal{T}_{\mathcal{G}}$. Thus, we could apply Theorem 5.

4.2 Average treatment effect

Hirano, Imbens, and Ridder (2003) proposed a semiparametrically efficient estimator for the average treatment effect. Their estimator used the propensity score, and they found that their estimator was more efficient when a nonparametrically estimated propensity score rather than the true propensity score was used.

Suppose we have a random sample of size n. For each sample, we observe (T_i, Y_i, X_i). T_i indicate whether the treatment of interest was received ($T_i = 1$) or not ($T_i = 0$). Let $Y_i(0)$ denote the outcome for i under control and $Y_i(1)$ under treatment. We are not able to observe both $Y_i(0)$ and $Y_i(1)$, we can only observe $Y_i \equiv T_iY_i(1) + (1-T_i)Y_i(0)$. In addition, we observe a vector of covariates denoted by X_i. To solve the identification problem, the unconfoudedness assumption is assumed.

Assumption 4 (Unconfounded Treatment Assignment)

$$T \perp (Y(0), Y(1))|X.$$

For estimating the average effect of the treatment $E[Y(1) - Y(0)] \equiv \beta_0$, Hirano, Imbens and Ridder (2003) suggested the following semiparametric estimator

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_iT_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)}\right),$$

where $\hat{p}(x)$ was a nonparametric estimator of the propensity score $p(x)$

$$p(x) \equiv P(T = 1|X = x) = E[T|x].$$

Let $\tilde{\beta}$ denote a estimator that uses the true propensity score,

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_iT_i}{p(X_i)} - \frac{Y_i(1-T_i)}{1-p(X_i)}\right).$$

Their estimator is included in the class of our two step semiparametric estimators with a moment condition

$$m((T_i, Y_i, X_i), \beta, \hat{p}) = \frac{Y_iT_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)} - \beta.$$

The influence function of $\hat{\beta}$ was shown in Appendix B of Hirano, Imbens, and Ridder (2003) and it is

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n (m_i + \tau_i) + o_p(1),$$

where

$$\begin{aligned}\tau_i &= -\left(\frac{E[Y_i(1)|T=1,X=X_i]}{p(X_i)}\right.\\ &\quad\left.+\frac{E[Y_i(0)|T=0,X=X_i]}{1-p(X_i)}\right)\\ &\quad\times(T_i-p(X_i)).\end{aligned}$$

On the other hand, Hahn (1998) examined the semiparametric efficiency bound of this model and characterized the tangent space with respect to $p(x)$. It is

$$\mathcal{T}_p = \{a(x)(T-p(x))\},$$

where $a(x)$ is any square integrable measurable function. Thus τ_i is included in the tangent space and it implies

$$\tau_i = -\text{Proj}(m_i|\mathcal{T}_p).$$

Theorem 5 could be applied.

5 CONCLUDING REMARKS

This paper investigated the paradoxical phenomenon that the efficiency of estimator increased when the nuisance parameters are estimated. We found that the paradox might occur even if the nuisance parameters were infinite dimensional. We obtained the necessary and sufficient condition of the paradox, and showed a sufficient condition in an easy-to-understand way. Our sufficient condition depended on the relation of the projection of the estimating function on the tangent set.

The stochastic equicontinuity condition was used for linearizing the moment condition with respect to infinite dimensional parameter. It might be possible to utilize other linearization methods, for example Fréchet Differentiation.

Only two example, the Kaplan-Meier integral and the average treatment effect, are explained in the paper. It might be interesting to inspect how popular is this paradoxical phenomenon.

6 REFERENCES

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

Bickel, P. J., Klassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66:429–436.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Henmi, M. (2004). A paradoxical effect of nuisance parameters on efficiency of estimators. *Journal of the Japan Statistical Society*, 34(1):75–86.

Henmi, M. and Eguchi, S. (2004). A paradox consterning nuisance parameters and projected estimating functions. *Biometrika*, 91:929–941.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B*, 61:413–438.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, 10(2):475–478.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always oberved. *Journal of the American Statistical Association*, 89:846–866.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56:931–954.

Stute, W. (1994). The bias of kaplan-meier integrals. *Scandinavian Journal of Statistics*, 21:475–484.

Stute, W. (1995). The central limit theorem under random censorship. *The Annals of Statistics*, 23(2):422–439.

Stute, W. and Wang, J. L. (1993). The strong low under random censorship. *The Annals of Statistics*, 21:1591–1607.

Suzukawa, A. (2004). Unbiased estimation of functionals under random censorship. *Journal of the Japan Statistical Society*, 34(2):153–172.