Econometric Theory, **24**, 2008, 1717–1728. Printed in the United States of America.
doi:10.1017/S0266466608080699

NOTES AND PROBLEMS

A PUZZLING PHENOMENON IN SEMIPARAMETRIC ESTIMATION PROBLEMS WITH INFINITE-DIMENSIONAL NUISANCE PARAMETERS

KOHTARO HITOMI
Kyoto Institute of Technology

YOSHIHIKO NISHIYAMA
Kyoto University

RYO OKUI
Hong Kong University of Science and Technology

This note considers a puzzling phenomenon that is observed in some semiparametric estimation problems. In some cases, using estimated values of the nuisance parameters provides a more efficient estimator for the parameters of interest than does using the true values. This phenomenon takes place even in cases of semi-nonparametric models in which the nuisance parameters are infinite-dimensional and cannot be estimated at the parametric rate. We examine the structure and present the necessary and sufficient condition for the occurrence of this puzzle. We also provide a simple sufficient condition. It shows that the puzzle occurs when the term accounting for the effect of estimation of nuisance parameters is included in the tangent space. This condition is often satisfied when the estimating equation does not bring any restriction on the form of the nuisance parameters. Our simple sufficient condition can be applied to many important estimators.

Valuable comments were obtained from two anonymous referees, co-editor Paulo Paruolo, as well as Takeshi Amemiya, Songnian Chen, Xiaohong Chen, Wolfgang Härdle, Hidehiko Ichimura, Ingrid van Keilegom, Keiji Nagai, Byeong Uk Park, Werner Ploberger, Artem Prokhorov, and seminar participants at various institutions. This research was partially supported by a Grant-in-Aid for Scientific Research (KAKENHI 14530031, 15330040, 18203014, 19530182) of the Japan Society for the Promotion of Science. Okui acknowledges financial support from HKUST under Project DAG05/06.BM16 and from Research Grants Council under Project HKUST643907.
Address correspondence to Ryo Okui, Department of Economics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; e-mail: okui@ust.hk.

© 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

1. MOTIVATION

This paper considers a puzzling phenomenon that is observed in some semiparametric estimation problems and that is particularly relevant in the missing data and treatment effect literature. It has been observed in many important statistical models that using estimated values of the nuisance parameters provides a more efficient estimator for the parameters of interest than does using the true values. This is a puzzle, as knowing the true value of the nuisance parameters should not be detrimental. Moreover, such a phenomenon takes place even in models in which the nuisance parameters are infinite-dimensional and are estimated nonparametrically.

Pierce (1982) is a classic paper that provides the puzzle. This problem is particularly relevant in the missing data literature. The set of papers reporting this phenomenon in the case of finite-dimensional nuisance parameters includes Rosenbaum (1987), Robins, Mark, and Newey (1992), Imbens (1992), Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995), Lawless, Kalbfleisch, and Wild (1999), and Wooldridge (1999, 2001), among others. We observe the puzzling phenomenon also in semi-nonparametric estimation problems in which the nuisance parameters are infinite-dimensional. Examples include: Hirano, Imbens, and Ridder (2003), Chen, Hong, and Tarozzi (2008), Magnac and Maurin (2007) on the estimator of Lewbel (2000), Lewbel and Schennach (2007), Newey and Ruud (2005), and Suzukawa (2004) on the Kaplan–Meier integral. Some of these examples are examined in Section 3.

This paper aims at clarifying the process of this puzzle. We present the necessary and sufficient condition for the puzzle to take place in semiparametric estimation problems that require the estimation of infinite-dimensional nuisance parameters. We also provide a simple sufficient condition, which yields an interesting insight into the problem, based on the theory of Newey (1990). The sufficient condition is that the term accounting for the effect of estimating the nuisance parameters ("the correction term") belongs to the tangent space with respect to the nuisance parameters. When this condition is satisfied, the influence function of the estimator using estimated nuisance parameters becomes the residual of the projection of the influence function of the estimator with known nuisance parameters on the tangent space. The variance of the residual is smaller than the variance of the original term. As a result, the puzzling phenomenon occurs. This condition is typically satisfied when the estimating equation does not bring any restriction on the form of the nuisance parameters because the tangent space in those cases becomes large and often contains any possible form of the correction term.

There have been several papers that investigate this puzzling phenomenon, such as Henmi (2004), Henmi and Eguchi (2004), Wooldridge (2007), and Prokhorov and Schmidt (2006). The main contribution of this paper is that our framework allows the nuisance parameters to be infinite-dimensional and they do not need to be estimated at the parametric rate. To the best of our knowl-

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

edge, this is an original contribution to the literature. The puzzle appears in several semi-nonparametric estimation problems, as mentioned previously. Moreover, the puzzle is more puzzling in estimation problems that require estimation of infinite-dimensional nuisance parameters than in cases in which we need to estimate only finite-dimensional parameters, because estimators of infinite-dimensional parameters may converge at a rate that is slower than the parametric rate.

2. RESULTS

Let $\{z_i\}_{i=1}^n$ be an independently and identically distributed sample from a population and the following moment condition be satisfied:

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

where $m(\cdot)$ is a known $q \times 1$ vector valued function, β_0 is a $q \times 1$ vector of parameters of interest, and g_0 is an infinite-dimensional nuisance parameter. For simplicity, we assume that g_0 is a scalar function and does not depend on the finite-dimensional parameter β and that $m(\cdot)$ is continuously differentiable. This setting includes all the examples considered in Section 3.

We consider the following two estimators of β. Let $\hat{g}$ be an estimate of g_0. The first estimator, $\hat{\beta}$, is given as the solution to

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

The second estimator, $\tilde{\beta}$, is given as the solution to the following equation:

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

This estimator is feasible only when g_0 is known. We study and compare the asymptotic variances of the estimators $\tilde{\beta}$ and $\hat{\beta}$.

Assumption 1.

(a) $z_i, i = 1, \ldots, n$ are independently and identically distributed;
(b) $\mathrm{E}\{m(z_i, \beta_0, g_0)m(z_i, \beta_0, g_0)'\} < \infty$;
(c) $m(z_i, \beta, g)$ is continuously differentiable with respect to β;
(d) $M \equiv \partial \mathrm{E}\{m(z, \beta_0, g_0)\}/\partial\beta$ is nonsingular;
(e) g does not depend on the parameter β;
(f) the estimators exhibit the following expansions:

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M^{-1}\{m(z_i, \beta_0, g_0) + \tau(z_i)\} + o_p(1),$$

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

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

where $\mathrm{E}\{\tau(z_i)\} = 0$ and $\mathrm{E}\{\tau(z_i)\tau(z_i)'\} < \infty$.

The term $\tau(z_i)$ is considered to be the "correction" term, which represents the effect of the first-stage estimation of g. Finding the term $\tau(z_i)$ is one of major tasks in semiparametric estimation problems. The formulas in Lee (1996, Ch. 10) would be useful to obtain the form of $\tau(z_i)$.

The asymptotic linearity of the estimators (Assumption 1(f)) is a key in our theory. Our main results presented subsequently essentially do not depend on what assumptions we use to guarantee the asymptotic linearity of the estimators. For example, Assumptions 4.1–2 of Newey (1990) may be employed to guarantee that the estimators are asymptotically linear.

2.1. The Necessary and Sufficient Condition

We derive the necessary and sufficient condition for the estimator $\hat{\beta}$ to be more efficient than the estimator $\tilde{\beta}$. Let $m_0(z_i)$ denote $m(z_i, \beta_0, g_0)$. Let $A \equiv \mathrm{E}\{m_0(z_i)\tau(z_i)'\}[\mathrm{E}\{\tau(z_i)\tau(z_i)'\}]^{-1}$ and $u(z_i) \equiv m_0(z_i) - A\tau(z_i)$. The term $A\tau(z_i)$ is the projection of $m_0(z_i)$ on $\tau(z_i)$. The term $u(z_i)$ is the residual of that projection.

Let $V(\hat{\beta})$ and $V(\tilde{\beta})$ denote the asymptotic variance matrix of $\hat{\beta}$ and $\tilde{\beta}$, respectively. These asymptotic variances are expressed as

$$V(\hat{\beta}) = M^{-1}[\mathrm{E}\{u(z_i)u(z_i)'\} + (A + I)\mathrm{E}\{\tau(z_i)\tau(z_i)'\}(A + I)'](M^{-1})',$$

$$V(\tilde{\beta}) = M^{-1}[\mathrm{E}\{u(z_i)u(z_i)'\} + A\mathrm{E}\{\tau(z_i)\tau(z_i)'\}A'](M^{-1})',$$

where I is the $q \times q$ identity matrix. By comparing these two asymptotic variances, we obtain the following theorem.

THEOREM 1. *Suppose that Assumption 1 is satisfied. Then,* $V(\hat{\beta}) \leq V(\tilde{\beta})$ *if and only if*

$$(A + I)\mathrm{E}\{\tau(z_i)\tau(z_i)'\}(A + I)' \leq A\mathrm{E}\{\tau(z_i)\tau(z_i)'\}A'.$$

Moreover, when $q = 1$, *then* $V(\hat{\beta}) \leq V(\tilde{\beta})$ *if and only if* $A \leq -\frac{1}{2}$.

Note that $V(\hat{\beta}) \leq V(\tilde{\beta})$ means that $V(\hat{\beta}) - V(\tilde{\beta})$ is nonpositive definite.

2.2. A Sufficient Condition

In Section 2.1, we obtained the necessary and sufficient condition for the asymptotic variance of $\hat{\beta}$ to be smaller than that of $\tilde{\beta}$. In this section, we present a simple sufficient condition.

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

We first introduce the concept of "tangent space," which is a key to our sufficient condition. See, e.g., Newey (1990), Bickel, Klassen, Ritov, and Wellner (1993), and Lee (1996) for more detailed discussions on tangent space. The tangent space is the semi-nonparametric counterpart of score function. To define the tangent space, first we consider a parametric submodel. We parameterize g, say, $g(\gamma)$, such that $g(\gamma_0) = g_0$ for some γ_0, where the parameter γ is finite-dimensional. In this way, we can construct a parametric submodel. Let $S_\gamma(z_i)$ denote the score vector with respect to γ of such a parametric submodel (at $\gamma = 0$), which is a function of z_i and is a random vector. Let r be the dimension of z_i. A tangent space in nonparametric direction is the mean-square closure of all linear combinations of scores $S_\gamma(z_i)$ for smooth parametric submodels:

$$\mathcal{T} = \Big\{ t(\cdot) : R^r \to R^q \quad \text{with } \mathrm{E}\{\|t(z_i)\|^2\} < \infty, \quad \exists B_j S_{\gamma j}(\cdot),$$

$$\text{such that } \lim_{j\to\infty} \mathrm{E}\{\|t(z_i) - B_j S_{\gamma j}(z_i)\|^2\} = 0 \Big\},$$

where $\{B_j\}_{j=1}^{\infty}$ is a sequence of nonrandom matrices. For example, consider the case in which the nuisance parameter is f where f is the density of a scalar random variable x and assume that there is no restriction on the form of f. A parametric submodel can be $f(x, \gamma) = f_0(x)\{1 + \gamma a(x)\}$, where f_0 is the true density and $a(x)$ is a bounded scalar function that satisfies $\mathrm{E}\{a(x)\} = 0$ so that $f(x, \gamma)$ is a density function for any γ. The score function with respect to γ at $\gamma = 0$ is $a(x)$. This implies that the tangent space includes all bounded mean-zero functions of x, and it becomes the entire Hilbert space, i.e., $\{t(\cdot) : R \to R^q$ with $\mathrm{E}\{t(x)\} = 0, \mathrm{E}\{\|t(x)\|^2\} < \infty\}$.

Next, we define the projection on the tangent set. The projection $\mathrm{Proj}\{m_0(z_i)|\mathcal{T}\}$ is defined as a vector belonging to $\mathcal{T}$ that satisfies $\mathrm{E}[\{m_0(z_i) - \mathrm{Proj}(m_0(z_i)|\mathcal{T})\}a] = 0$ for all $a \in \mathcal{T}$. Let $u(z_i) \equiv m_0(z_i) - \mathrm{Proj}\{m_0(z_i)|\mathcal{T}\}$ be the residual from the projection of $m_0(z_i)$ on $\mathcal{T}$. We have the following result, which is a special case of Theorem 4.3 in Newey (1990).

LEMMA 2. *Suppose that Assumption 1 is satisfied and $\hat{\beta}$ is regular.*[1] *If $\tau(z_i) \in \mathcal{T}$, then $\hat{\beta}$ has an influence function, $-M^{-1}u(z_i)$.*

The proof is omitted because it is essentially identical to that of Theorem 4.3 in Newey (1990). The idea of the proof is the following. First we note that the influence function is orthogonal to any score for a regular semiparametric submodel, S_γ, because of the generalized information equality. Therefore, we have $\mathrm{E}[\{m_0(z_i) + \tau(z_i)\}S_\gamma] = \mathrm{E}[\{m_0(z_i) - (-\tau(z_i))\}S_\gamma] = 0$. This equality implies that $-\tau(z_i)$ is the projection of $m_0(\cdot)$ on $\mathcal{T}$. In this case, the matrix A in Section 2.1 becomes $A = -I$. Note that the proof does not essentially depend on the assumptions used for showing the asymptotic linearity of the estimator.

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

The variance of $u(z_i)$ is smaller than the variance of $m_0(z_i)$. Thus, we have the following theorem, which is our main result, as a direct implication of the lemma.

THEOREM 3. *Suppose that Assumption 1 is satisfied and* $\hat{\beta}$ *is regular. If* $\tau(z_i) \in \mathcal{T}$, *then* $V(\hat{\beta}) \leq V(\tilde{\beta})$. *A strict inequality holds when* $\text{Proj}\{m_0(z_i)|\mathcal{T}\} \neq 0$.

This theorem demonstrates that the asymptotic variance of $\hat{\beta}$ never exceeds the asymptotic variance of $\tilde{\beta}$ if the correction term, $\tau(z_i)$, is included in the tangent space.

Remark 1. Some of the nuisance parameters may not explicitly appear in the moment condition. We do not need to estimate them in such a case. Then, in verifying the condition $\tau(z_i) \in \mathcal{T}$ of Theorem 3, it suffices to show $\tau(z_i) \in \mathcal{T}_g$ where $\mathcal{T}_g$ is the tangent space with respect to g only. This is because $\mathcal{T}_g \subset \mathcal{T}$, noting that $\mathcal{T}$ is the tangent space with respect to all the nuisance parameters in the model. In some cases, $\mathcal{T}$ may not be easy to obtain, but deriving $\mathcal{T}_g$ is much easier; then checking $\tau(z_i) \in \mathcal{T}_g$ is much more convenient. In the next section, where we apply the theorem to several estimators, we only consider $\mathcal{T}_g$.

Remark 2. Our result includes the case of finite-dimensional nuisance parameters. As observed in Pierce (1982), Henmi and Eguchi (2004), Wooldridge (2007), and others, the puzzling phenomenon occurs in finite-dimensional cases when the nuisance parameters are estimated by maximum likelihood estimators. An example of the puzzle in finite-dimensional cases is given in Remark 6 of Section 3.1. Suppose that γ is the p-dimensional vector of nuisance parameters and S_γ is the score with respect to γ. The tangent space in finite-dimensional cases is the linear subspace spanned by S_γ, i.e., $\{BS_\gamma : B \in R^{q\times p}\}$. When the nuisance parameters are estimated by maximum likelihood estimators, the correction term becomes a linear transformation of the score and belongs to the tangent space. Thus, our theory can be used to explain the puzzle in cases in which finite-dimensional nuisance parameters are estimated by maximum likelihood estimators, although the main focus of this paper is the puzzle in semi-nonparametric estimation problems.

Remark 3. Henmi and Eguchi (2004) provide an alternative explanation of the puzzle. Our theory and that of Henmi and Eguchi (2004) are different in two ways. First, Henmi and Eguchi (2004) consider the puzzling phenomenon for finite-dimensional nuisance parameter cases only.[2] The second difference is that we essentially examine the projection of the estimating equation on the score, but Henmi and Eguchi (2004) examine the projection of the score on the estimating equation. We can handle semi-nonparametric estimation problems by examining the projection on the tangent space. On the other hand, it appears to be difficult to extend the approach of Henmi and Eguchi (2004) to infinite-

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

dimensional cases as it requires an explicit examination of the score and the equation used for the estimation of the nuisance parameters.

Remark 4. An important implication of this theorem is as follows. When there is no restriction on the form of the nuisance parameter, the tangent space with respect to the nuisance parameter is typically large and can include almost any possible form of the correction term. Thus, the puzzling phenomenon appears to occur when the moment condition does not impose any restriction on the form of the nuisance parameter. In fact, the moment conditions do not impose any restriction on the form of the nuisance parameters in all the examples discussed in the next section. This observation is somewhat similar to the result of Prokhorov and Schmidt (2006, Statements 8 and 9 in Thm. 2.2). They argue that the puzzle occurs when the moment condition used for estimating the parameters of interest is not useful in estimating the nuisance parameters.

3. DISCUSSION

This section discusses three semi-nonparametric estimation problems in which the puzzling phenomenon takes place.[3] The sufficient condition presented in the previous section turns out to be applicable to all three examples.

3.1. Hirano et al. (2003)

Hirano et al. (2003) propose a semiparametrically efficient estimator for the average treatment effect. Suppose that we have a random sample of size n: $\{z_i\}_{i=1}^n \equiv \{T_i, Y_i, X_i\}_{i=1}^n$. The binary variable T_i indicates the treatment status. 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 observe only $Y_i \equiv T_i Y_i(1) + (1 - T_i) Y_i(0)$. The vector X_i contains covariates.

Our objective is to estimate the average treatment effect, $\mathrm{E}\{Y_i(1) - Y_i(0)\} \equiv \beta_0$. Let $p(x)$ denote the propensity score: $p(x) = \Pr(T_i = 1 | X_i = x) = \mathrm{E}(T_i | X_i = x)$. Under the assumption $T \perp \{Y(0), Y(1)\} | X$, the parameter β_0 is identified because $\mathrm{E}\{Y_i T_i / p(X_i)\} = \mathrm{E}\{Y(1)\}$ and $\mathrm{E}\{Y_i(1 - T_i)/p(X_i)\} = \mathrm{E}\{Y(0)\}$. With this observation, Hirano et al. (2003) suggest the following semi-nonparametric estimator for the average treatment effect β_0:

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

where $\hat{p}(x)$ is the series logit estimator of the propensity score (for more details, see Hirano et al., 2003). Let $\tilde{\beta}$ denote the estimator obtained by replacing $\hat{p}(x)$ by the true propensity score $p(x)$ in equation (1). Hirano et al. (2003) demonstrate that $V(\hat{\beta}) \leq V(\tilde{\beta})$.

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

The estimator $\hat{\beta}$ is included in our framework with the moment condition:

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

The influence function of $\hat{\beta}$ is provided in Appendix B of Hirano et al. (2003), and it is $m_0(z_i) + \tau(z_i)$, where

$$\tau(z_i) = -\left[\frac{\mathrm{E}\{Y_i(1)|T = 1, X = X_i\}}{p(X_i)} + \frac{\mathrm{E}\{Y_i(0)|T = 0, X = X_i\}}{1 - p(X_i)} \right] \{T_i - p(X_i)\}.$$

Hahn (1998) examines the semiparametric efficiency bound of this model and characterizes the tangent space with respect to $p(x)$. It is

$$\mathcal{T}_p = \{t : \exists a(x) \quad \text{such that } t = a(x)\{T - p(x)\}\},$$

where $a(x)$ is any square integrable measurable function. Thus, $\tau(z_i)$ is included in the tangent space, and the estimator $\hat{\beta}$ is more efficient than $\tilde{\beta}$.

Remark 5. Chen et al. (2008) consider a related estimator called the inverse probability weighted generalized method of moments (GMM) estimator in moment restriction models with nonclassical measurement error. The puzzling phenomenon is also present in their estimator. A similar argument to the preceding discussion shows that our sufficient condition is satisfied in their estimator.

Remark 6. A leading example of the puzzling phenomenon in finite-dimensional cases is the estimation problem that is similar to that of Hirano et al. (2003) but in which the propensity score is modeled parametrically. Suppose that the propensity score is modeled as $p(x) = p(x, \gamma_0)$, where γ_0 is finite-dimensional. Suppose that we estimate γ_0 by the maximum likelihood estimator so that $\hat{\gamma} = \arg\max \sum_{i=1}^{n} [T_i \log p(X_i, \gamma) + (1 - T_i) \log\{1 - p(X_i, \gamma)\}]$ and then estimate β by formula (1) replacing $\hat{p}(X_i)$ with $p(X_i, \hat{\gamma})$. This estimator is also more efficient than $\tilde{\beta}$ (see, e.g., Wooldridge, 2007). As stated in Remark 2, the tangent space with respect to γ is the linear subspace spanned by the score with respect to γ, and the correction term in this case is a linear transformation of the score because the correction term is determined by $\hat{\gamma} - \gamma_0$, whose first-order term is a linear transformation of the score. Thus, our theory can explain the puzzling phenomenon in both the semi-nonparametric and parametric versions of this estimation problem.

3.2. Lewbel (2000)

The second example is the estimator proposed by Lewbel (2000) for binary choice models with possibly endogenous regressors. Suppose that we have independently and identically distributed samples of $z_i \equiv (y_i, v_i, w_i, x_i)$ for

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

$i = 1, \ldots, n$. The binary variable y_i is generated by the data generating process of $y_i = \mathbf{1}\{w_i\beta + v_i + \epsilon_i \geq 0\}$, where ϵ_i is an unobservable disturbance term and β is the vector of parameters of interest. The scalar v_i is independent of ϵ_i, and the coefficient on v_i is normalized to 1. The vector variable w_i might be correlated with ϵ_i. Suppose that x_i is correlated with w_i and is independent of ϵ_i. Lewbel (2000) shows that under regularity conditions, the following moment condition holds:

$$\mathrm{E}\left[x\left\{\frac{y - \mathbf{1}\{v > 0\}}{f(v|w,x)} - w\beta\right\}\right] = 0,$$

where $f(v|w,x)$ is the conditional density function of v given w and x and $\mathbf{1}$ is the indicator function.

The Lewbel (2000) estimator $\hat{\beta}$ is obtained by solving the sample analog of the moment condition replacing $f(v|w,x)$ by the kernel density estimator $\hat{f}(v|w,x)$. That is, the estimator $\hat{\beta}$ solves

$$\frac{1}{n}\sum_{i=1}^{n}\left[x_i\left\{\frac{y_i - \mathbf{1}\{v_i > 0\}}{\hat{f}(v_i|w_i,x_i)} - w_i\hat{\beta}\right\}\right] = 0.$$

Let $\tilde{\beta}$ denote the estimator that uses the true $f(v|w,x)$. Magnac and Maurin (2007) note that the asymptotic variance of $\hat{\beta}$ is smaller than that of $\tilde{\beta}$.

We show that our sufficient condition is satisfied in the Lewbel (2000) estimator. Lewbel (2000) and Magnac and Maurin (2007) provide the form of the correction term $\tau(z_i)$ for this estimator:

$$\tau(z_i) = -x_i\mathrm{E}\left[\frac{y_i - \mathbf{1}\{v_i > 0\}}{f(v_i|w_i,x_i)}\,|\,v_i,w_i,x_i\right] + x_i\mathrm{E}\left[\frac{y_i - \mathbf{1}\{v_i > 0\}}{f(v_i|w_i,x_i)}\,|\,w_i,x_i\right].$$

If there is no restriction on the form of $f(v|w,x)$, the tangent space with respect to $f(v|w,x)$ is easily obtained, and it is

$$\mathcal{T} = \{A(v,w,x)|\mathrm{E}(A|w,x) = 0, \mathrm{E}(A^2|w,x) < \infty\}.$$

In this model, the moment condition does not impose any restriction on the form of $f(v|w,x)$. Therefore, the tangent space with respect to $f(v|w,x)$ is just $\mathcal{T}$. It is easy to see that $\tau \in \mathcal{T}$.

3.3. Newey and Ruud (2005)

Last, we consider the estimator of Newey and Ruud (2005) for single index models. Suppose that we have independently and identically distributed samples of $z_i = (y_i, x_i)$ for $i = 1, \ldots, n$. The variable y_i is a scalar, and its conditional expectation given the random vector x_i has the single index form $\mathrm{E}(y|x) = \tau(x'\beta)$, where β is a vector of parameters of interest and τ is an unknown

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

function. Newey and Ruud (2005) propose the inverse density weighted least-squares estimator for β. Let $f(x)$ be a density with the linear conditional expectation property[4] and compact support.[5] Let $h(x)$ be the density of x. Let $X = (1, x')'$ and $\gamma = (\gamma_0, \beta)$, where γ_0 is the intercept parameter. The following moment condition holds:

$$\mathrm{E}\left\{\frac{f(x)}{h(x)}(y - \gamma' X)\right\} = 0.$$

See Newey and Ruud (2005) for more detailed discussion on this moment condition.

We consider the following two estimators. Let $\hat{\gamma}$ be the vector solving

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{\hat{h}(x_i)}(y_i - \hat{\gamma}' X_i) = 0, \tag{2}$$

where $\hat{h}(x)$ is the kernel density estimator. The first estimator $\hat{\beta}$ is the coefficient on x in $\hat{\gamma}$. Let $\tilde{\gamma}$ be the vector obtained by replacing $\hat{h}(x_i)$ by the true $h(x_i)$ in equation (2). The second estimator $\tilde{\beta}$ is the coefficient on x in $\tilde{\gamma}$.

Newey and Ruud (2005) note that $\hat{\beta}$ is more efficient than $\tilde{\beta}$. We show that our sufficient condition is satisfied in the current case. Newey and Ruud (2005) give the form of the correction term:

$$\tau(z_i) = w(x_i)x_i\{\mathrm{E}(y_i|x_i) - x_i'\gamma_0\}.$$

The tangent set with respect to $h(x)$ is

$$\mathcal{T} = \{A(x)|\mathrm{E}(A) = 0, \mathrm{E}(A^2) < \infty\}.$$

It is easy to derive the tangent set in this case because as in Section 3.2, the moment condition does not impose any restriction on the form of $f(x)$. It is easy to see that $\tau \in \mathcal{T}$.

Remark 7. In all the examples, it is the inverse of the nuisance parameter that enters the moment condition. One may try to find a possible relationship between the form of the moment condition and occurrence of the puzzle. However, there is only a limited connection between these two. The puzzle appears to occur when the moment condition does not impose any restriction on the form of the nuisance parameter. The functional form of the moment condition matters because it determines whether or not the moment condition places some restriction on the form of the nuisance parameter. We present an example in which we do not observe a puzzling phenomenon even though the nuisance parameter is in the denominator of the moment condition. Let f be the density of a random variable x. Consider the moment condition $\mathrm{E}[\{y/f(x)\} - x\beta_0] = 0$, where y is another random variable and β_0 is the parameter of interest. This

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

moment condition puts a restriction on $f(x)$, as we have $\int xf(x)\,dx\beta_0 = \int \mathrm{E}(y|x)\,dx$. In fact, estimating f yields a less efficient estimator in this case.[6]

NOTES

1. Let $\theta = (\beta, \eta)$ where η is the possibly infinite-dimensional nuisance parameter (g is a part of η) and θ characterizes the distribution of z_i. Let θ_0 be the true value of θ. Let $\beta(\theta)$ be the value of β in θ. We consider a local data generating process in which the data are generated according to θ_n and $\sqrt{n}(\theta_n - \theta_0)$ is bounded. The estimator $\hat{\beta}$ is regular if the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta(\theta_n))$ under θ_n does not depend on the local data generating process. See Newey (1990, p. 102) for more details.

2. Although infinite-dimensional nuisance parameters appear in the Henmi and Eguchi (2004) discussion (κ in the notation of Hemi and Eguchi), they are not estimated. Therefore, their results do not explain the present phenomenon.

3. The working paper version of this paper, which is available upon request, also examines the Kaplan–Meier integral. The Kaplan–Meier integral is an integral with respect to the Kaplan–Meier estimator of the distribution of a censored random variable. Suzukawa (2004) determines that the Kaplan–Meier integral can be written in terms of the Kaplan–Meier estimator of the distribution of the censoring time and finds that estimating the censoring distribution yields a more efficient estimator than does using the true distribution.

4. The density of x is said to possess the linear conditional expectation property when the conditional expectation of x given any linear combination of x is linear in that combination (Newey and Ruud, 2005).

5. Note that Newey and Ruud (2005) consider that $f(x)$ depends on a vector of unknown parameters θ that is used to center the location and scale of the density. However, it is not related to the puzzle, and we shall ignore the estimation of θ.

6. Let $u = \{y/f(x)\} - x\beta_0$. The asymptotic variance of the estimator using a nonparametric estimator of f is $\mathrm{var}[\{u - \mathrm{E}(u|x)\} + \{x - \mathrm{E}(x)\}\beta_0]$, and that of the estimator using the true $f(x)$ is $\mathrm{var}\{u - \mathrm{E}(u|x)\}$. These asymptotic variances can be obtained by employing, e.g., formula (2.6) in Lee (1996, Ch. 10). The latter is smaller than the former, as $u - \mathrm{E}(u|x)$ and $x - \mathrm{E}(x)$ are uncorrelated.

REFERENCES

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

Chen, X., H. Hong, & A. Tarozzi (2008) Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics* 36, 808–843.

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

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

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

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

Imbens, G.W. (1992) An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* 60, 1187–1214.

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

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press

Lee, M.J. (1996) *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*. Springer-Verlag.

Lewbel, A. (2000) Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics* 97, 145–177.

Lewbel, A. & S.M. Schennach (2007) A simple ordered data estimator for inverse density weighted expectations. *Journal of Econometrics* 136, 189–211.

Magnac, T. & E. Maurin (2007) Identification and information in monotone binary models. *Journal of Econometrics* 139, 76–104.

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

Newey, W.K. & P.A. Ruud (2005) Density weighted linear least squares. In D.W.K. Andrews & J.H. Stock (eds.), *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, pp. 554–573. Cambridge University Press.

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

Prokhorov, A. & P. Schmidt (2006) GMM Redundancy Results for General Missing Data Problems. Mimeo, Concordia University, Canada.

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

Robins, J.M. & A. Rotnitzky (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90, 122–129.

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

Rosenbaum, P.R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association* 82, 387–394.

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

Wooldridge, J.M. (1999) Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica* 67, 1385–1406.

Wooldridge, J.M. (2001) Asymptotic properties of weighted *M*-estimators for standard stratified samples. *Econometric Theory* 17, 451–470.

Wooldridge, J.M. (2007) Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141, 1281–1301.

https://doi.org/10.1017/S0266466608080699 Published online by Cambridge University Press