

Semiparametric Efficiency in Irregularly Identified Models*

Shakeeb Khan[†] and Denis Nekipelov[‡]

July 2008

ABSTRACT. This paper considers efficient estimation of structural parameters in a class of semiparametric models where these parameters are *irregularly* identified. For such models conventional semiparametric efficiency bound calculations of Stein(1956) are of little use as they do not result in finite bounds. Thus the notion of efficiency has to be reconsidered for this class of models and we attempt to address this gap in the literature. Along the lines of the minimax bounds attained in Ibragimov and Has'minskii(1981) we attain not only minimax from below orders of $n^{-1/2}$, but qualitative bounds as well. We focus on such minimax bounds for three examples- the regression coefficients in binary choice models under median and mean restrictions (Horowitz(1992,1993) and Lewbel(1998,2000)), the randomly censored regression model considered in Koul et al.(1981), and the ATE under unconfoundedness (Hahn(1998))), noting that under stated conditions, regular rates of convergence are unattainable.

*We thank H. Hong for helpful comments.

[†]Department of Economics, Duke University, 213 Social Sciences Building, Durham, NC 27708. Phone: (919) 660-1873.

[‡]Department of Economics, University of California - Berkeley, 508-1 Evans Hall, #3880, Berkeley, CA 94720. Support from the National Science Foundation is gratefully acknowledged.

1 Introduction

Efficiency bounds for structural parameters in econometric and statistical models have been of great interest for several decades. Bounds such as the Cramer-Rao bound for parameters in parametric models and semiparametric efficiency bounds, as introduced in Stein(1956) for finite dimensional parameters in models with nonparametric components, have proven to be very useful for a number of reasons. One is that they have help in quantifying the efficiency loss of adopting a given estimation procedure by comparing its variance to that of the bound. Second, they also provide information on the relative asymptotic efficiency of any particular consistent estimator. For example, in the class of parametric models, Cramer-Rao bounds have been used to show the efficiency (under appropriate regularity conditions) of the maximum likelihood estimation procedure.

Consequently, several important papers have proposed methods for computing these bounds. Important examples include Koshevnik and Levit(1976), Pfanzagl and Wefmeyer(1982), Begun et al.(1983), Newey(1990) and more recently Severini and Tripathi(2003). Severini and Tripathi(2003) proposed a simplified procedure for deriving bounds in a class of moment condition models, which was a welcome addition to the literature since the aforementioned previous work made calculation of bounds relatively difficult.

However, for a class of models the aforementioned bounds are not very useful. This is because, for these models, the aforementioned bounds are not finite, making standard optimality theorems- e.g. Hajek's(1970) convolution theorem inapplicable to ascertain optimal estimation procedures. Consequently, this paper aims to define and attain efficiency bounds for such models where standard notions cannot be applied.

Like in existing work, we will work with the concept of minimax risk on bounds to define our notion of optimality. For the class of irregular models we consider we will attain minimax risk bound from below which is of order slower than the regular rate of the square root of the sample size. This is analogous to existing results in the literature on attain optimal rates of convergence for the parameters of interest. However this paper aims to go one step further in the sense that we also aim to attain qualitative bounds.

The concept of optimality for irregular models is not new to the statistics and econometrics literature. The notion of efficiency bounds for irregular models, those where the optimal rate of convergence for the parameters of interest are different from the regular rate of the square root of the sample size, is covered in detail in van der Vaart(1991), who provides a series of examples for parametric models. Certain nonparametric models can fit into this setting as well, see for example Hasminskii and Ibragimov(1981) who consider efficient es-

timation of a density function. Our results in this paper are different in the sense that the parameter we consider conducting optimal inference is finite dimensional in a model with infinite dimensional nuisance parameters- i.e. a semiparametric model.

The class of irregular models that have been studied in Econometrics literature is substantially different from the models covered in this paper. Existing literature focuses on the cases of discontinuity of the likelihood function or derivatives which normally leads to the convergence rates which are faster than the parametric rate. For instance, Chernozhukov and Hong(2004) study a class of models with jumps in density functions and analyze the properties of Bayesian and classical estimators for these models. Hirano and Porter(2003) analyze the class of models with parameter-dependent support and use the notion of the asymptotic minimax risk to build tools for comparison of superefficient estimators. In our paper we deal with a substantially different case where the semiparametric model has zero information. In this case the definition of asymptotic efficiency will require not only the optimal choice of a normalizing sequence for the estimator, but also the choice of the “best” parametric sub-model of the semiparametric model.

The rest of the paper is organized as follows. The next section defines the notions we will be working with, such as minimax risk and minimax bounds, so that we can introduce our notion of optimality. Section 3 then provides the general theorem of this paper, which is analogous to deriving the formal efficiency bound as defined in the previous section. Section 4 then illustrates our results in the context of deriving the bounds for three specific parameters: 1)the regression coefficients in binary choice models under median and mean restrictions, 2)the regression coefficients in randomly censored regression models and 3) the average treatment effect (ATE) under an unconfoundedness condition. Section 5 then concludes by summarizing and discussing areas for future research.

2 Definitions

The notion of the efficiency bound naturally arises in parametric models. In that case we consider sampling of data from the parametrized family of distributions $\{F_\theta(\cdot), \theta \in \Theta \subset \mathbb{R}\}$ and the members of this family are free from complications such as parameter-dependent support, mass-points and density jumps. The lower bound for the variance of the estimator of a smooth function of θ , $\varphi(\theta)$ is provided by *Cramér-Rao bound* which is defined by the inverse Fisher information for θ and the Jacobi matrix for $\varphi(\cdot)$. The Fisher information requires \mathbf{L}_2 -differentiability of the square root of the density, and the regularity of a particular model

is associated with the existence of a finite Fisher information on the parameter set. In this paper we refer to a particular estimator T_n of a parameter θ as a regular estimator⁴ if there exists a Gaussian element $G(\cdot)$ such that

$$\sqrt{n}(T_n - \theta) \Rightarrow G(\cdot),$$

where \Rightarrow denotes weak convergence of a corresponding empirical process.

A more general approach to bounds in parametric setting was provided in Le Cam(1986) where the problem of parameter estimation is considered from the point of view of the concept of a statistical experiment. Evaluation of bounds in this setting will be associated with general loss functions and risk functions rather than the asymptotic variance as in the simplest case. A statistical experiment is characterized by a triple $(\mathcal{X}, \mathcal{B}, F_\theta)$ with $\theta \in \Theta$ where \mathcal{X} is the support of random variable, \mathcal{B} is a corresponding σ -algebra. Denote $T_{1/\sqrt{n}}$ a sequence of asymptotically normal estimators and $\theta_{1/\sqrt{n}}$ a sequence of estimators for θ . We consider a set of statistical experiments $\mathcal{E}_\epsilon = (\mathcal{X}^{(\epsilon)}, \mathcal{B}^{(\epsilon)}, F_\theta^{(\epsilon)})$ with $\theta \in \Theta$. Following Ibragimov and Hasminskii(1981) we can define that a sequence of estimators for θ : $\hat{\theta}_\epsilon$ is ℓ_ϵ -asymptotically efficient at point $\theta \in \Theta$ with respect to the family of loss functions $\ell_{\epsilon(\cdot)}$ if

$$\lim_{\delta \rightarrow 0} \lim_{\epsilon \rightarrow 0} \left[\inf_{T_\epsilon} \sup_{|u-\theta| < \delta} E_u^{(\epsilon)} \ell_\epsilon(T_\epsilon - u) - \sup_{|u-\theta| < \delta} E_u^{(\epsilon)} \ell_\epsilon(\hat{\theta}_\epsilon - u) \right] = 0. \quad (2.1)$$

This definition can be extended to the entire set Θ . In this paper we only consider subconvex loss functions. We call a particular loss function $\ell : \mathbb{R}^k \mapsto [0, \infty)$ subconvex if for each z the projection of its lower contour set $\{x : \ell(x) \leq z\}$ is convex and closed.

Expression (2.1) could be difficult to verify in practice. To make the study of ℓ_ϵ -asymptotic efficiency more tenable, we can adapt the techniques developed for parametric settings and verify that using these techniques it is possible to develop estimators consistent with (2.1). General results regarding the properties of the risk function in semi-parametric settings are provided in van der Vaart and Wellner(1996), and here we provide a general overview of the existing results. In the semiparametric setting we consider the problem of estimating a parameter

$$\psi(F) = \varphi_0 \left(\int \varphi(x) F(dx) \right),$$

where $F \in \mathcal{F}$ can be a substantially larger set of functions than $\{F_\theta, \theta \in \Theta\}$. In this model the generalized information will be associated with a functional derivative of F , and

⁴See Newey(1990) for a slightly different definition of a regular estimator.

correspondingly reflect the properties of the efficient score of the model. To find the value of the functional derivative, we consider a smooth parametric family $\eta = \{F_h(x)\} \in \mathcal{F}$ and denote $\psi(F) = t$. Then we impose that for $h = t$ $F_t = F$. This re-arrangement has “picked” a subset of F which is much easier to analyze than F itself. This subset is a parametric family of distributions and we can compute the information for this family in the parametric sense. For a precise definition of the information in the semiparametric case we impose restrictions that $\psi(F_h) = h + o(h - t)$ as $h \rightarrow t$. Moreover, the likelihood ratio can be defined as

$$\lambda_u = \left[\frac{dF_{t+u}^c}{dF_t}(X) \right]^{1/2},$$

and we assume that it has a mean-square derivative $\dot{\lambda}_0$ at $u = 0$ such that

$$\lim_{u \rightarrow 0} \frac{1}{u^2} E_F \left[\left[\frac{dF_{t+u}^c}{dF_t}(X) \right]^{1/2} - 1 - u \dot{\lambda}_0 \right]^2 = 0.$$

Then the information for the considered family of distributions η is

$$I(F, \eta) = 4E_F \lambda_0^2.$$

The information for the semiparametric model can be defined over all parametric families as:

$$\inf_{\varphi \in \mathcal{F}} I(F_0, \varphi) = I(F).$$

The existence of a finite generalized information in the semiparametric model implies the possibility of the parametric rate of convergence of an efficient semiparametric estimator. In particular, a generalized version of the asymptotic efficiency in this case will imply that for regular estimators $\hat{\psi}_n$ for $\psi(F)$

$$\begin{aligned} \lim_{n \rightarrow \infty} \inf_{\hat{\psi}_n} \sup_{F \in \mathcal{F}} E_F \ell \left(\sqrt{n} \left(\hat{\psi}_n - \psi(F) \right) \right) \\ \geq \sup_{F_0 \in \mathcal{F}} \frac{1}{\sqrt{2\pi}} \int \ell \left(I^{-1/2}(F_0) x \right) \exp \left(-\frac{x^2}{2} \right) dx. \end{aligned}$$

This concept is valid when the information is finite. In this paper we consider the case where the information is infinite, but there exists a normalizing sequence r_n such that a sequence of estimators $\hat{\psi}_n$ is regular with respect to this normalizing sequence

$$r_n \left(\hat{\psi}_n - \psi(F) \right) \xrightarrow{F} \mathcal{L},$$

for all $F \in \mathcal{F}$ and a fixed tight probability measure \mathcal{L} in a Banach space. Then the risk bound can take into account the fact that the asymptotic risk can depend on the choice of a normalizing sequence:

$$\sup_{S \subset \mathcal{F}} \liminf_{n \rightarrow \infty} \sup_{F \in S} E_F \ell \left(r_n \left[\hat{\psi}_n - \psi(F) \right] \right) \geq E \ell(G),$$

where S are proper subsets of \mathcal{F} and G is a Gaussian element.

3 Bounds for Irregular Models

Consider the problem of semiparametric M-estimation where the finite-dimensional parameter θ is defined by a semi-parametric moment condition. This moment condition is defined by the unknown density f_0 of random covariate z as:

$$E[m(z, \theta_0, f_0)] = 0.$$

This moment equation is assumed to suffer from the problem of *irregular identification* for θ as described by Khan and Tamer (2007). The irregularity of an estimator for the model with irregular identification generally leads to the singularity of the information matrix corresponding to the model. However, it is still possible to estimate the parameter of interest using estimation procedures converging at non-parametric rates. In the subsequent discussion we will call a pair (θ, f) - the parameters of the model, and in this pair θ is a Euclidean parameter of interest while f is a nuisance parameter denoting the distribution of covariate z .

We impose the following set of assumptions on functions and parameters of the model.

Assumption 1 A.1 $\theta \in \Theta \subset \mathbb{R}^k$, where Θ is compact with respect to the Euclidean norm in \mathbb{R}^k . f belongs to a subset \mathbf{F} of Sobolev space \mathbb{S}^d which is compact with respect to the Sobolev norm in \mathbb{S}^d . $z \in \mathbb{R}^p$ is a random variable with absolutely continuous density f_0 .

A.2 $m(z, \theta, f)$ is locally Lipschitz in θ in the neighborhood of (θ_0, f_0)

A.3 There exists a function $\Gamma(z)$ such that $E[\|\Gamma(z)\|^2] < \infty$ and a number $\|J_0\| < \infty$ such that

$$\left. \frac{\partial}{\partial \theta} E[\Gamma(z)m(z, \theta, f_0)] \right|_{\theta=\theta_0} = J_0.$$

A.4 There exists $\delta(z)$, $E[\|\delta(z)\|^2] < \infty$ such that for each δf such that $f_0 + \delta f$ is in the tangent set

$$\|m(z, \theta_0, f_0 + \delta f) - m(z, \theta_0, f_0) - \delta(z) \delta f\| \leq L(z) \|\delta f\|,$$

for $E[\|L(z)\|^2] < \infty$.

A.5 Maximum convergence rate for estimation of irregularly identified θ is γ_n

A.6 Equation $E[m(z, \theta, f_0)] = 0$ has a unique solution in Θ at point θ_0 .

Consider a family of local perturbations of parameters (θ_t, f_t) similarly to Ibragimov and Has'minskii(1981):

$$\theta_t = \theta_0 + \Gamma(z)t\Delta_\theta, \quad \text{and} \quad f_t(z) = f_0(z)[1 + \lambda(z)t],$$

where $\lambda(z)$ is such that $f + f\lambda \in \mathbb{S}^d$, $E[f(z)\lambda(z)] = 0$ and $E[|\lambda(z)|^2] < \infty$. We chose the parameters such that $t \in [0, 1]$ and :

$$E_t[m(z, \theta_t, f_t)] = 0.$$

Analogously to the terminology developed for regular semi-parametric estimation, we will call the closure of the linear envelope of the set of scores of the model corresponding to the perturbed nuisance parameter *the local tangent set* of the model. In the examples that we consider in this paper we explicitly perform such local perturbations, and thus explicitly show that the local tangent set is not empty.

The paths along the local perturbations pass through (θ_0, f_0) . At the next step we compute the derivative of the moment condition along the path indicated by local perturbations. In this case we relate $\lambda(z)$ to the score of the model $S_t(z)$ along the parametrization path t . Then we can express Δ_θ as:

$$J_0 \Delta_\theta = -E[(m(z, \theta_0, f_0) + \delta(z)) S_t(z)].$$

This expression defines the quantity Δ_θ which can be characterized as a normalized pathwise derivative of the Euclidean parameter. For the expectation on the right-hand side standard projection results in Newey(1994) will hold. In fact, our modification of the usual score calculus is augmented only in the part when we compute the Jacobi matrix of the moment equation while the pathwise derivative for the nuisance parameter is computed in the standard way. As a result, we can find the quasi-influence function of the model:

$$\psi(z) = -\{m(z, \beta_0, f_0) + (\delta(z) - E[\delta(z)])\}. \quad (3.2)$$

Consider a class of the generalized linear estimators $\hat{\theta}$ denoting $\bar{\theta} = E[\hat{\theta}]$ such that:

$$\gamma_n(\hat{\theta} - \bar{\theta}) = \frac{1}{n} \sum_{i=1}^n \zeta_n(z_i) \tilde{\psi}(z_i) + o_p(1),$$

where $E[\zeta_n(z_i) \tilde{\psi}(z)] = 0$, $\lim_{n \rightarrow \infty} E[n^{-1} |\zeta_n(z_i)|] < \infty$ and the conditions of the Lindeberg-Feller CLT are satisfied. We will show next that the estimator for irregularly identified

parameter θ belongs to the class of the generalized linear estimators. Moreover, the quasi-influence function $\psi(\cdot)$ plays the role of $\tilde{\psi}(\cdot)$.

Suppose that moment vector $m(\cdot)$ forms a system of moments which exactly identifies the parameter θ . Consider an estimator $\hat{f}^{(n)}$ for the density f_0 . Denote

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

We will first introduce a high-level assumption regarding the asymptotic behavior of the empirical moment function and later give primitive conditions justifying this assumption.

Assumption 2 A.1 *Then assume that for all $\tilde{\theta} \xrightarrow{p} \theta_0$ there exists $\epsilon > 0$ such that for all $\|\delta\theta\| < \epsilon$*

$$\hat{m}^{(n)}\left(\tilde{\theta} + \gamma_n^{-1} \delta\theta\right) - m^{(n)}\left(\tilde{\theta}\right) - J_0 \delta\theta = o_p(\|\delta\theta\|).$$

A.2 *Empirical moment function evaluated at the true Euclidean parameter value*

$$\sqrt{n}m^{(n)}(\theta_0) \xrightarrow{d} N(0, \sigma_m^2).$$

The estimator for β solves the system of equations:

$$\hat{m}^{(n)}(\theta) = 0. \tag{3.3}$$

The following theorem summarizes the properties of the estimator of interest:

Theorem 3.1 *Suppose that Assumptions 1.1-1.5 and 2.1-2.2 are satisfied. Then $\gamma_n(\hat{\theta} - \bar{\theta})$ is generalized asymptotically linear with*

$$\psi(z_i) = m(z_i, \theta_0, f_0) + \delta(z_i) - E[\delta(z_i)],$$

and $\zeta_n(z_i) = \sqrt{n}J_0$. Moreover, the asymptotic variance of $\gamma_n(\hat{\theta} - \bar{\theta})$ does not depend on the type of the estimator for the unknown density.

Proof:

Note that γ_n/\sqrt{n} approaches zero. Therefore, defining $\delta\theta = (\hat{\theta} - \bar{\theta}) \frac{\gamma_n}{\sqrt{n}}$ and applying Assumptions 2.1-2.2 and the definition of the estimator, we obtain that the estimator is asymptotically normal. Then using the proof of Theorem 2.1 in Newey(1994) we verify that the influence function associated with the estimator will be defined by (3.2). \square

Now let us analyze lower level assumptions required to obtain the result of Theorem 3.1 without using Assumptions 2.1-2.2. First of all, note that the Lipschitz constant in Assumption 1.4 is a function of z only. From this assumption we can formulate a requirement on the convergence rate of the estimator for the density of covariate z .

Assumption 3

$$\sqrt{n} \left\| \hat{f}^{(n)} - f_0 \right\|^2 \xrightarrow{p} 0.$$

Assumption 3 implies that the convergence rate for the non-parametric density estimate should not be slower than $n^{1/2}$ to not impact the asymptotic variance of $\hat{\theta}$. As γ_n reflects non-parametric convergence rate, for problems that we are interested in this convergence rate can be significantly lower than $n^{-1/4}$. Such estimators are readily available for problems with sufficiently smooth distributions of covariates.

The next condition assuring the appropriate behavior of the non-parametric estimate is the condition for the stochastic equicontinuity of the moment equation. By Assumption 1.4 the moment equation belongs to the type II class of functions in the definition of Andrews(1994). Therefore, as shown by Andrews(1994), the Pollard’s entropy condition should hold for these functions with the envelope $C\delta(\cdot)$, where C is a fixed constant. Next, we need an additional requirement for the moments of the envelope function $\delta(\cdot)$.

Assumption 4 For some $\epsilon > 0$ $E[\delta(z)^{2+\epsilon}] < \infty$.

Assumptions 1.4 and 4 assure stochastic equicontinuity of the moment equation. The next step is to introduce quantity $\delta(z)f_0(z) - E[\delta(z)f_0(z)]$. The following assumption sets the restriction on $\delta(\cdot)f_0(\cdot)$ so that it mitigates differences between the integration over the non-parametrically estimated distribution $\hat{F}^{(n)}$ and the empirical distribution.

Assumption 5

$$\sqrt{n} \left\{ \int \delta(z) [\hat{f}^{(n)} - f_0] f_0 dz - \frac{1}{n} \sum_{i=1}^n [\delta(z_i) f_0(z_i) - E[\delta(z) f_0(z)]] \right\} \xrightarrow{p} 0.$$

To see the intuition behind this assumption, note that

$$\int \delta(z) [\hat{f}^{(n)} - f_0] f_0 dz = \int [\delta(z) f_0(z) - E(\delta(z) f_0(z))] \hat{f}^{(n)} dz.$$

This means that the error in the moment condition due to plug-in estimate of density is small enough that does not exceed the sampling error, represented by the empirical distribution.

Assumptions 1.4, 3, 4, 5 validate Lemma 5.1 in (Newey, 1994). As a result, we conclude that for

$$\Omega = \text{Var} [m(z, \theta_0, f_0) + \delta(z) - E[\delta(z)]],$$

the moment equation evaluated at the true value of the Euclidean parameter converges to normal distribution at the parametric rate. Thus,

$$\sqrt{n}\hat{m}^{(n)}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [m(z_i, \theta_0, f_0) + \delta(z_i) - E[\delta(z_i)]] + o_p(1) \xrightarrow{d} N(0, \Omega).$$

Consistency of the estimator $\hat{\theta}$ is provided by Assumptions 1.4 and 1.6. Asymptotic normality of the estimate of θ_0 normalized by the Jacobi matrix will follow from the same arguments as in Andrews(1991) as structure of the Jacobi matrix and the vector of moment equations has similar properties to that problem.

Theorem 3.1 suggests that asymptotic properties of the semiparametric estimate of the finite-dimensional parameters, normalized by the Jacobi matrix do not depend on a particular selection of the sieve functions used to compute non-parametric components of the model. In our case, however, the estimation method will influence the variance of the unnormalized parameters. We consider local representations for coefficients with

$$\hat{\theta}_n - \bar{\theta}_n = (\hat{\theta}_n - \theta_0) - (\bar{\theta}_n - \theta_0).$$

We consider a normalization sequence γ_n such that

$$\gamma_n^{-1} J_n = J_0^\eta + o_p(1),$$

$$\gamma_n (\bar{\theta}_n - \theta_0) = h^\eta + o_p(1).$$

We label both the bias and the normalized Jacobi matrix by η to indicate that, generally speaking, these limits depend on the choice of structure used to estimate infinite-dimensional nuisance parameter of the model. Then the asymptotic distribution of parameter of interest can be expressed as:

$$\gamma_n (\hat{\theta}_n - \theta_0) \xrightarrow{d} N \left(h^\eta, (J_0^\eta)^{-1} \Omega (J_0^\eta)^{-1} \right)$$

The following theorem states the main result in this section

Theorem 3.2 *For a chosen normalizing sequence γ_n corresponding to the fastest convergence rate for estimators of $\theta_0 \in \Theta \subset \mathbb{R}^k$, estimators $\hat{\theta}_n^\eta$ converging at rate γ_n and depending on the estimation method $\eta \in \Xi$ for nuisance parameter, and a subconvex loss function $\ell(\cdot)$:*

$$\inf_{\eta \in \Xi} \lim_{n \rightarrow \infty} \inf_{\hat{\psi}_n^\eta} \sup_{F \in \mathcal{F}} E_F \ell \left(\gamma_n (\hat{\theta}_n^\eta - \theta_0(F)) \right) > 0.$$

This theorem establishes the minimax bound for the estimator of θ_0 across possible non-parametric distributions. The intuition behind this result is that it suggests that minimization over estimation procedures and over all estimators for the "coarsest" non-parametric distribution does not make the minimum risk with normalization by γ_n degenerate.

4 Bounds for Irregularly Identified Semiparametric Models

In this section we illustrate the relevance of Theorems 3.1 and 3.2 by applying them to widely studied semiparametric models whose irregularity properties have been established in the literature. Specifically, we will consider regression coefficients in semiparametric binary choice models under conditional mean and median conditions, as well the average treatment effect (ATE) in the treatment effect model under unconfoundedness.

4.1 (Smoothed) Maximum Score Model

Consider the model with the binary outcome

$$y_i = \mathbf{1}[x_i'\beta_0 - \epsilon_i > 0]$$

where x_i is a k - dimensional vector of observed covariates, and the unobserved error term ϵ_i satisfies a conditional median restriction, thus allowing for the conditional heteroskedasticity. The object of interest is the k - dimensional vector of regression coefficients β_0 .

Manski(1975,1985) established the identification of β_0 under mild smoothness conditions on the distributions of $x_i'\beta_0$ and ϵ_i . Horowitz(1992) strengthened these smoothness conditions in order to attain faster rates of convergence for estimating β_0 and proposed the smoothed maximum score estimator which has a limiting gaussian distribution. Horowitz(1993) established the upper bounds on achievable rates of convergence for estimating β_0 , but did not provide qualitative bounds for estimators that do indeed achieve the optimal rate. Horowitz(1992) also establishes a minimized MSE for his proposed estimator but this bound is left as a function of the kernel function chosen for implementation.

Here we apply Theorems 3.1 and 3.2 to provide bounds for the class of estimators achieving the optimal rate established in Horowitz(1993). We can do so because the identification

of β_0 explicitly stems from the conditional moment condition:

$$\beta_0 = -E \left[\frac{\delta(\mathcal{P}(x_i) - \frac{1}{2})}{f_X(x_i)} x_i \right]$$

where to simplify exposition we assumed the model has a single continuous regressor x_i whose coefficient has been normalized to 1 and whose density function is denoted by $f_X(\cdot)$, $\mathcal{P}(x_i) = E[y_i|x_i]$ is the propensity score, and $\delta(\cdot)$ denotes the Dirac delta function.

The idea of the smoothed maximum score estimator is that the delta-function can be approximated by a family of smooth functions $\delta_\epsilon(\cdot)$ such that $\delta_\epsilon(\cdot) \xrightarrow{\epsilon \rightarrow 0} \delta(\cdot)$, where we consider weak convergence in the class of generalized functions. The idea of the estimator will be to consider a set of approximating moment equations with the moment function

$$l_\epsilon(y_i, x_i) = -\frac{\delta_\epsilon(\mathcal{P}(x_i) - \frac{1}{2})}{f_X(x_i)} x_i.$$

The optimal estimation problem in this case will boil down to the optimal choice of the smoothing sequence ϵ_n as $n \rightarrow \infty$. We assume that the optimal convergence rate for the estimator $\hat{\beta}_n$ is γ_n and it is known. In this case for each ϵ

$$\gamma_n \left(\hat{\beta}_n - \beta_0 \right) \xrightarrow{d} N(b_{\epsilon^*}, V_{\epsilon^*}).$$

This shows that for a specific sequence ϵ_n there will be an asymptotic bias. We can compute the asymptotic variance as:

$$V_{\epsilon^*} = E \left[\frac{1}{f_X(x_i)^2} (\delta_{\epsilon^*}(u_i)x_i - E[\delta_{\epsilon^*}(u_i)x_i])^2 \right],$$

and squared bias is

$$b_{\epsilon^*}^2 = E \left[\frac{\delta_{\epsilon^*}(u_i)x_i}{f_X(x_i)} \right]^2.$$

The minimum risk for the estimator of β_0 is evaluated in this case as:

$$r_\epsilon = \sup_{\epsilon^* > 0} \frac{1}{\sqrt{2\pi}} \int \ell \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \exp(-x^2/2) dx.$$

For a smooth parametrization the least upper bound will be non-zero and can be solved for by solving the non-linear first order condition:

$$\int \ell' \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \left[-\frac{1}{2} V_{\epsilon^*}^{-3/2} V_{\epsilon^*}' (-b_{\epsilon^*} + x) - V_{\epsilon^*}^{-1/2} b_{\epsilon^*}' \right] \exp(-x^2/2) dx = 0.$$

Denoting derivative of the approximating function with respect to the smoothness parameter $\delta'_\epsilon(\cdot)$, the optimal ϵ^* solves

$$\int \ell'(\xi) \left[-\xi V_{\epsilon^*}^{-1} E \left[\frac{\delta'_{\epsilon^*} x_i - E[\delta'_{\epsilon^*} x_i]}{f_X^2(x_i)} (\delta_{\epsilon^*} x_i - E[\delta_{\epsilon^*} x_i]) \right] - V_{\epsilon^*}^{-1/2} E \left[\frac{\delta'_{\epsilon^*} x_i}{f_X(x_i)} \right] \right] \exp(-x^2/2) dx = 0. \quad (4.4)$$

Therefore we can apply the previous theorems to attain the following qualitative bound for the maximum score model:

Theorem 4.1 *Under Assumptions 1-10 in Theorem 3.5 in Horowitz(1998) the bound for estimating β_0 is:*

$$r_{\epsilon^*} = \frac{1}{\sqrt{2\pi}} \int \ell \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \exp(-x^2/2) dx,$$

where ϵ^* solves (4.4)

In practical applications the loss function is restricted by the the type of the applied problem for which a particular irregularly identified model is used. For instance, if the problem requires the best in-sample mean-squared prediction, then the associated loss function will be quadratic. In this case the derivation of the minimax bound for the problem reduces to the evaluation of the asymptotic variance of the estimator, and the asymptotic bias arising due to the dominating role of the non-parametric component of the model. Below we derive the minimax bound for such a concrete example.

To simplify the notation we assume that the parameter of interest is an intercept term, i.e. the binary model has only one varying regressor whose coefficient has been normalized to 1. Denote the kernel function by $K(\cdot)$. Then by picking the bandwidth such that the estimator converges at its fastest rate, the asymptotic bias, b_{ϵ^*} , and variance, V_{ϵ^*} are:

$$b_{\epsilon^*} = c_2 \mathcal{Q}^{-1} \sum_{j=1}^s \frac{1}{j!(s-j)!} F^{(j)}(0) f_X^{(s-j)}(0)$$

$$V_{\epsilon^*} = c_1 \mathcal{Q}^{-1} f_X(0) \mathcal{Q}^{-1}$$

where

$$\mathcal{Q} = 2[F^{(1)}(0|0)f_X(0)].$$

In these expressions $F^{(1)}(\cdot|\cdot)$ denotes the first derivative of the error distribution function in the binary model conditional on regressor x , c_i are constants associated with the kernel

function, s denotes the assumed order of smoothness of the probability and index density functions and the superscript j denotes the j^{th} order derivative. Coefficients c_1 and c_2 will be determined by the optimal kernel structure and the order of smoothness of estimated distributions. Below we give an example of computing these constants if the degree of smoothness of the propensity score is up to four.

One concrete example when we can provide the constant for the minimum risk in case of quadratic loss function is the case where the function $\delta(\cdot)$ is approximated by a sequence of functions (kernels) with finite support. Consider the following procedure for estimating parameters in the model.

- **Step 1**

Estimate functions $\mathcal{P}(\cdot)$ and $f_X(\cdot)$ non-parametrically denoting estimates $\widehat{\mathcal{P}}(\cdot)$ and $\widehat{f}_X(\cdot)$

- **Step 2**

estimate parameters of interest using kernel $K(\cdot)$

$$\hat{\beta} = -\frac{1}{nh} \sum_{i=1}^n \frac{K\left(\frac{\widehat{\mathcal{P}}(x_i) - \frac{1}{2}}{h}\right)}{\widehat{f}_X(x_i)} \widehat{\mathcal{P}}'(x_i) x_i,$$

for a particular choice of bandwidth h .

This procedure is constructed to approximate the infeasible estimator

$$\beta_0 = -E \left[\frac{\delta_{\mathcal{P}^{-1}(\frac{1}{2})}(x)}{f_X(x)} x \right],$$

where $\delta_a(x)$ is the Dirac delta-function. We assume that a compact set \mathcal{X} contains all covariates. We also assume that estimators $\widehat{\mathcal{P}}(\cdot)$, $\widehat{\mathcal{P}}'(\cdot)$ and $\widehat{f}_X(\cdot)$ are such that for each $x \in \mathcal{X}$ they converge with rate of at least $n^{-1/4}$. The existence of such estimators will be assured by smoothness of these functions, compactness of their support and their continuity on the boundary of support. Given that $h = O(n^{-1/3})$, this means that error due to approximation error in $\widehat{\mathcal{P}}(\cdot)$ and $\widehat{f}_X(\cdot)$ is negligible. Therefore, we consider the asymptotics for

$$n^\alpha (\hat{\beta} - \beta_0) = n^\alpha \left(-\frac{1}{nh} \sum_{i=1}^n \frac{K\left(\frac{\mathcal{P}(x_i) - \frac{1}{2}}{h}\right)}{f_X(x_i)} \mathcal{P}'(x_i) x_i - \beta_0 \right) + o_p(1),$$

where α is the convergence rate determined by the order of smoothness of $\mathcal{P}(\cdot)$. We assume that $\mathcal{P}(\cdot)$ is locally monotone and $2k$ times differentiable. The kernel is symmetric and integrates to one. Consider

$$\begin{aligned}\beta_h &= -E \left[\frac{1}{h} \frac{K\left(\frac{P(x)-\frac{1}{2}}{h}\right)}{f_X(x)} \mathcal{P}'(x)x \right] = -\frac{1}{h} \int K\left(\frac{P(x)-\frac{1}{2}}{h}\right) \mathcal{P}'(x)x \, dx \\ &= -\int K(u) \mathcal{P}^{-1}\left(\frac{1}{2} + uh\right) \, du.\end{aligned}$$

To derive the minimax bound over kernel-based procedures we consider cases where the degree of smoothness of the propensity score functions is from 1 to 4. In the Appendix we demonstrate the procedure to derive the bound. In the first step we determine the convergence rate of the estimator which can be found by balancing the bias and the variance. In the second step, we optimize the kernel structure using the fact that we impose restrictions on the degree of smoothness of the propensity score function. Thus, we can find the corresponding structure of the kernel by representing it in terms of a basis expansion in an appropriate functional space. We find that for the cases with 2 and 3 - times differentiable propensity score the optimal Kernel has Epanechnikov structure with

$$K(u) = \frac{1}{\sqrt{2}}L_0(u) - \frac{1}{\sqrt{10}}L_2(u) = \frac{3}{4}(1 - u^2),$$

where $L_i(\cdot)$ are normalized Legendre polynomials of degree i . The minimum mean-squared error of estimating the kernel can be computed as

$$n^{4/5}\text{MSE}(\hat{\beta}) = \frac{5}{4} \left(\frac{9\beta_0^4}{80}\right)^{2/5} \left(\frac{\mathcal{P}''(-\beta_0)}{\mathcal{P}'(-\beta_0)f_X^2(-\beta_0)}\right)^{2/5}.$$

Similarly, for 4 and 5-times differentiable propensity score cases the optimal kernel can be written as

$$K(u) = \frac{1}{\sqrt{2}}L_0(u) - \frac{\sqrt{10}}{4}L_2(u) + \frac{36\sqrt{2} - 3\sqrt{38}}{136}L_4(u).$$

A detailed derivation can be found in Appendix.

4.2 Binary Choice Model with Exclusion Restriction

In this section we consider a model with the binary outcome

$$y_i = I[x_i'\beta_0 + v_i - \epsilon_i > 0],$$

where v_i is a continuously distributed regressor whose coefficient has been normalized to 1 and whose distribution is assumed to be independent from ϵ_i conditional on x_i . Lewbel established the identification of β_0 under support conditions on ϵ_i and proposed a consistent estimator for β_0 . Khan and Tamer(2007) established an impossibility theorem for this model that is analogous to the theorem in Chamberlain(1986). They further showed how optimal rates, slower than the parametric rate of the square root of the sample size, depend on relative tail behavior conditions on v_i and ϵ_i , but did not provide qualitative bounds. The Lewbel(2001) identification result is based on the moment condition:

$$\beta_0 = E[(y_i - I\{v_i > 0\})f(v_i)^{-1}] \equiv E[(y_i - I\{v_i > 0\})w(v_i)]$$

with the weight function $w(v_i) = f(v_i)^{-1}$ where $f(\cdot)$ here denotes the density of v_i , where here again we assumed a single regressor, v_i , to ease the exposition.

The nature of the irregularity here is of the same kind as in case of the smooth maximum score with the only difference that the irregularity occurs at the tails of the regressor. In this case it would be convenient to introduce a family of trimmed moment functions:

$$l_\epsilon(y_i, x_i, v_i) = \frac{(y_i - \mathbf{1}\{v_i > 0\})}{f(v_i)} \mathbf{1}\{v_i < 1/\epsilon\}$$

Note that this moment function converges to the original moment function for $\epsilon \rightarrow +0$. Estimation will proceed given an optimal choice of the trimming sequence $\epsilon_n \rightarrow 0$. We assume that the optimal convergence rate for the estimator of β_0 is known and corresponds to the optimal choice of the trimming sequence. In this case for the optimal sequence ϵ^*

$$\gamma_n \left(\hat{\beta}_n - \beta_0 \right) \xrightarrow{d} N(b_{\epsilon^*}, V_{\epsilon^*}).$$

This shows that for a specific sequence ϵ_n there will be an asymptotic bias. We can compute the asymptotic variance as:

$$V_{\epsilon^*} = E \left[\left(\frac{(y_i - \mathbf{1}\{v_i > 0\})}{f(v_i)} \mathbf{1}\{v_i < 1/\epsilon^*\} - E \left[\frac{(y_i - \mathbf{1}\{v_i > 0\})}{f(v_i)} \mathbf{1}\{v_i < 1/\epsilon^*\} \right] \right)^2 \right],$$

and squared bias is

$$b_{\epsilon^*}^2 = E \left[\frac{(y_i - \mathbf{1}\{v_i > 0\})}{f(v_i)} \mathbf{1}\{v_i < 1/\epsilon^*\} \right]^2.$$

The minimum risk for the estimate of β_0 is evaluated in this case as:

$$r_\epsilon = \sup_{\epsilon^* > 0} \frac{1}{\sqrt{2\pi}} \int \ell \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \exp(-x^2/2) dx.$$

For smooth parametrization the least upper bound will be non-zero and can be obtained by solving a non-linear first order condition:

$$\int \ell' \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \left[-\frac{1}{2} V_{\epsilon^*}^{-3/2} V_{\epsilon^*}' (-b_{\epsilon^*} + x) - V_{\epsilon^*}^{-1/2} b_{\epsilon^*}' \right] \exp(-x^2/2) dx = 0.$$

Denote $z_i = \frac{(y_i - \mathbf{1}\{v_i > 0\})}{f(v_i)}$ then

$$V_{\epsilon^*}' = \left(E \left[z_i^2 \middle| v_i = \frac{1}{\epsilon^*} \right] - 2E \left[z_i \mathbf{1} \left\{ v_i < \frac{1}{\epsilon^*} \right\} \right] E \left[z_i \middle| v_i = \frac{1}{\epsilon^*} \right] \right) f_v \left(\frac{1}{\epsilon^*} \right).$$

Similarly, we find the derivative of the bias as

$$b_{\epsilon^*}' = E \left[z_i \middle| v_i = \frac{1}{\epsilon^*} \right] f_v \left(\frac{1}{\epsilon^*} \right).$$

This leads to the equation for the optimal ϵ^* :

$$\begin{aligned} \int \ell'(\xi) f_v \left(\frac{1}{\epsilon^*} \right) V^{-1} \left(\xi E \left[\frac{z_i^2}{2} - z_i E \left[z_i \mathbf{1} \left\{ v_i < \frac{1}{\epsilon^*} \right\} \right] \middle| v_i = \frac{1}{\epsilon^*} \right] \right. \\ \left. + V^{1/2} E \left[z_i \middle| v_i = \frac{1}{\epsilon^*} \right] \right) = 0. \end{aligned} \tag{4.5}$$

From the identification condition for the parameter vector β we can derive qualitative bounds for the parameter of interest for the class of estimators converging at the optimal rate:

Theorem 4.2 *Under Assumptions 1-5 in Lewbel(2001) the bound for estimating β_0 is:*

$$r_{\epsilon^*} = \frac{1}{\sqrt{2\pi}} \int \ell \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \exp(-x^2/2) dx,$$

where ϵ^* solves (4.5).

As the result of the theorem suggests, the minimax bound is associated with the optimal choice of the trimming sequence which keeps the integral in the expectation from explosive behavior. Our result, however, does not provide a recipe for constructing such a sequence. Next we provide the results for the minimax bound when we consider a quadratic loss function, where we can associate the minimax bound for the estimator with a particular way of trimming.

To simplify notation we will assume that the parameter of interest is an intercept term, i.e. the binary model has only one varying regressor whose coefficient has been normalized to

1. We will examine the asymptotic bias and variance of the infeasible version of the estimator where the density of the special regressor is assumed to be known. The asymptotic bias, b_{ϵ^*} and variance V_{ϵ^*} are

$$b_{\epsilon^*} = \lim_{\gamma_n \rightarrow \infty} \gamma_n (1 - \mathcal{P}(\gamma_n))$$

$$V_{\epsilon^*} = \lim_{\gamma_n \rightarrow \infty} \gamma_n^{-1/2} \int_{-\infty}^{\gamma_n} \frac{\mathcal{P}(v) (1 - \mathcal{P}(v))}{f(v)} dv.$$

where $f(\cdot)$ denotes the density of the special regressor and $\mathcal{P}(\cdot)$ denotes the propensity score function, which in this model is only a function of the values that the special regressor takes. In this case the bias and the variance are determined solely by the functional form of density and the propensity score.

We can demonstrate a concrete example of the choice of trimming sequence for the example with one special regressor and one intercept term. We also assume that it is known that the propensity score is twice differentiable. We consider a family of weighting functions generated by ratios of kernels

$$\omega_h(v) = 1 + h^{-2} \frac{K_1\left(\frac{v}{h}\right)}{K_2\left(\frac{v}{h}\right)}, \quad (4.6)$$

where each of the kernels $K_i(\cdot)$ is symmetric. We assume that

$$\lim_{h \rightarrow 0} h^{-2} \frac{K_1\left(\frac{v}{h}\right)}{K_2\left(\frac{v}{h}\right)} = 0, \quad \text{for each } v \in \mathbb{R},$$

and

$$\int_{-\infty}^{+\infty} \int_0^x \frac{K_1(z)}{K_2(z)} dz dx = 0.$$

We use the weights to generate the sample moment condition

$$\widehat{\beta}_h = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \mathbf{1}\{v_i > 0\})}{f(v_i)} \omega_h(v_i).$$

Note that

$$\begin{aligned} \beta_h &= E \left[\widehat{\beta}_h \right] = \int_{-\infty}^{+\infty} \int_{\epsilon}^{\infty} [\mathbf{1}\{v + \beta_0 - \epsilon > 0\} - \mathbf{1}\{v > 0\}] \omega_h(v) dv dF_{\epsilon} \\ &= - \int_{\epsilon}^{\epsilon - \beta_0} \left(\int_0^{\epsilon - \beta_0} \omega_h(v) dv \right) dF_{\epsilon} = \beta_0 + \int f_{\epsilon}(\beta_0 + hu) \left(\int_0^u \frac{K_1(z)}{K_2(z)} dz \right) du. \end{aligned}$$

Using the properties of the weighting function, we find that

$$\beta_h - \beta_0 = h\mathcal{P}''(0) \int u \left(\int_0^u \frac{K_1(z)}{K_2(z)} dz \right) du + o(h).$$

To compute the variance note that

$$\begin{aligned} \text{Var} \left(\widehat{\beta}_h \right) &= \frac{1}{nh^4} E_\epsilon \left[\text{sign}\{\epsilon - \beta_0\} \int_0^{\epsilon - \beta_0} f(v)^{-1} \frac{K_1^2\left(\frac{v}{h}\right)}{K_2^2\left(\frac{v}{h}\right)} dv \right] \\ &= \frac{2}{nh^2} \frac{\mathcal{P}'(0)}{f_v(0)} \int_0^{+\infty} \int_0^u \frac{K_1^2(z)}{K_2^2(z)} dz du + o(n^{-1}h^{-2}). \end{aligned}$$

This provides an expression for the minimum mean-squared error

$$n^{1/2} \text{MSE} \left(\widehat{\beta} \right) = 2\sqrt{2} \left(\frac{\mathcal{P}'(0)(\mathcal{P}''(0))^2}{f_v(0)} \right)^{1/2} \left[\int_{-\infty}^{+\infty} \int_0^u u \frac{K_1(z)}{K_2(z)} dz du \right] \left[\int_0^{+\infty} \int_0^u \frac{K_1^2(z)}{K_2^2(z)} dz du \right]^{1/2}.$$

The optimal bandwidth sequence is $h_n = O(n^{-1/4})$. The optimal structure of the kernels can be found if one represents them as expansions in the basis of Legendre polynomials. The coefficients of these expansions can be found numerically.

4.3 Treatment Effects Model under Exogenous Selection

A central problem in evaluation studies is that potential outcomes that program participants would have received in the absence of the program is not observed. Letting d_i denote a binary variable taking the value 1 if treatment was given to agent i , and 0 otherwise, and letting y_{0i}, y_{1i} denote potential outcome variables, we refer to $y_{1i} - y_{0i}$ as the *treatment effect* for the i 'th individual. A parameter of interest for identification and estimation is the *average treatment effect*, defined as:

$$\beta = E[y_{1i} - y_{0i}] \tag{4.7}$$

One identification strategy for β was proposed in Rosenbaum and Rubin(1983), who imposed the following:

- (i) There exists an observed variable x_i s.t. d_i and (y_{0i}, y_{1i}) are independent of each other given x_i .
- (ii) $0 < P(d_i = 1|x_i) < 1 \quad \forall x_i$

Hirano et al.(1999) showed the following inverse weighting identification result:

$$\beta = E \left[\frac{y_i(d_i - \mathcal{P}(x_i))}{\mathcal{P}(x_i)(1 - \mathcal{P}(x_i))} \right] \quad (4.8)$$

whereas Khan and Tamer(2007) show identification is irregular for a wide class of models satisfying the above assumptions, precluding estimation at the parametric rates. They derive upper bounds for optimal rates for estimating β that they show depends on the density of the propensity score near the limit points of 0 and 1.

Consider a simple case of one-dimensional x , then we can provide a trimmed version of the moment equation for β by

$$l_\epsilon(y_i, x_i) = \frac{y_i(d_i - \mathcal{P}(x_i))}{\mathcal{P}(x_i)(1 - \mathcal{P}(x_i))} \mathbf{1}\{\varphi(x_i) < \epsilon\},$$

where $\varphi(\cdot)$ is the function indicating the domain of the propensity score. This function is constructed such that $\epsilon \rightarrow 0$, then the moment equation converges to the original moment equation. We choose the optimal trimming sequence corresponding to the optimal convergence rate of the estimator for treatment effect. In this case for the optimal sequence ϵ^* convergence of the estimator is described as

$$\gamma_n \left(\hat{\beta}_n - \beta_0 \right) \xrightarrow{d} N(b_{\epsilon^*}, V_{\epsilon^*}).$$

As in the previous examples, estimation at an optimal rate will be associated with an asymptotic bias. We can compute the asymptotic variance as:

$$V_{\epsilon^*} = E \left[\left(\frac{y_i(d_i - \mathcal{P}(x_i))}{\mathcal{P}(x_i)(1 - \mathcal{P}(x_i))} \mathbf{1}\{\varphi(x_i) < \epsilon^*\} - E \left[\frac{y_i(d_i - \mathcal{P}(x_i))}{\mathcal{P}(x_i)(1 - \mathcal{P}(x_i))} \mathbf{1}\{\varphi(x_i) < \epsilon\} \right] \right)^2 \right],$$

and squared bias is

$$b_{\epsilon^*}^2 = E \left[\frac{y_i(d_i - \mathcal{P}(x_i))}{\mathcal{P}(x_i)(1 - \mathcal{P}(x_i))} \mathbf{1}\{\varphi(x_i) < \epsilon\} \right]^2.$$

The minimum risk for the estimate of β_0 is evaluated in this case as:

$$r_\epsilon = \sup_{\epsilon^* > 0} \frac{1}{\sqrt{2\pi}} \int \ell \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \exp(-x^2/2) dx.$$

For smooth parametrization the least upper bound will be non-zero and can be solved for by solving a non-linear first order condition:

$$\int \ell' \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \left[-\frac{1}{2} V_{\epsilon^*}^{-3/2} V_{\epsilon^*}' (-b_{\epsilon^*} + x) - V_{\epsilon^*}^{-1/2} b_{\epsilon^*}' \right] \exp(-x^2/2) dx = 0.$$

Denote $z_i = \frac{y_i(d_i - \mathcal{P}(x_i))}{\mathcal{P}(x_i)(1 - \mathcal{P}(x_i))}$ then

$$V'_{\epsilon^*} = \left(E \left[z_i^2 \middle| x_i = \varphi^{-1}(\epsilon^*) \right] - 2E[z_i \mathbf{1}\{\varphi(x_i) < \epsilon^*\}] E \left[z_i \middle| x_i = \varphi^{-1}(\epsilon^*) \right] \right) f_X(\varphi^{-1}(\epsilon^*)).$$

Similarly, we find the derivative of the bias as

$$b'_{\epsilon^*} = E \left[z_i \middle| x_i = \varphi^{-1}(\epsilon^*) \right] f_X(\varphi^{-1}(\epsilon^*)).$$

This leads to the equation for the optimal ϵ^* :

$$\begin{aligned} \int \ell'(\xi) f_X(\varphi(\epsilon^*)) V^{-1} \left(\xi E \left[\frac{z_i^2}{2} - z_i E[z_i \mathbf{1}\{x_i < \varphi^{-1}(\epsilon^*)\}] \middle| x_i = \varphi^{-1}(\epsilon^*) \right] \right. \\ \left. + V^{1/2} E[z_i | x_i = \varphi^{-1}(\epsilon^*)] \right) = 0. \end{aligned} \quad (4.9)$$

Given the moment condition in (4.8) we can apply the theorems from the previous section to derive bounds:

Theorem 4.3 *Under Assumptions 1-3 in Hahn(1998) the bound for estimating β is:*

$$r_{\epsilon^*} = \frac{1}{\sqrt{2\pi}} \int \ell \left(V_{\epsilon^*}^{-1/2} (-b_{\epsilon^*} + x) \right) \exp(-x^2/2) dx,$$

where ϵ^* solves (4.9).

Similarly to the case with inverse density weighting, both the bias and the variance are only determined by the properties of the propensity score.

A concrete example of the minimax bound with a particular degree of smoothness of the propensity score can be taken directly from the binary choice model with an exclusion restriction. Then using the same weighting function (4.6) we can form the estimator

$$\widehat{\beta}_h = \frac{1}{n} \sum_{i=1}^n \frac{y_i(d_i - \mathcal{P}(x_i))}{\mathcal{P}(x_i)(1 - \mathcal{P}(x_i))} \omega_h(x_i).$$

The structure of the optimal bandwidth sequence as well as the structure of the optimal mean-squared error will be equivalent to that in the case of the binary choice model with an exclusion restriction.

5 Simulation Results

In this section we illustrate the irregularity of the identification attained for some of the models we derived efficiency bounds for. Results are illustrated in tables I-V, which report the basic statistics mean bias, median bias and MSE, for various distributional assumptions on the error term and regressors. Results are reported for sample sizes of 100, 400, 1600, and 6400, using 10000 replications for each design.

Tables I and II consider estimation of the intercept term (up to scale) in a binary choice model. Table I reports results for the case when the density of the special regressor is known. The three designs considered were 1) both regressor and error standard normal distributed independently of each other, 2) both regressor and error distributed standard logistic independent of each other, and 3) regressor distributed standard normal and error distributed (independently) logistic. As the results illustrate, the estimator, although performing very well, does not converge at the parametric (root- n) rate, as its MSE does not decline at the rate the sample size increases, for any of the 3 designs. Table II considers the same exercise, but now with the density of the regressor estimated using a normal kernel function and Silverman's rule of thumb for bandwidth selection. This estimator performs better than the previous one in terms of MSE than the previous one, but also does not converge at the parametric rate, as indicated by the rate of decline of its MSE.

Tables III and IV report results for estimation of the ATE. Here we simulated from a simple model where the treatment equation was an indicator taking the value one if the sum of a regressor and an error term exceeded 0, and 0 otherwise. The counterfactual outcome equations were the sum of the same regressor and an independently distributed standard normal error for the treated and untreated, but the treatment outcome also had an intercept term of 1. Table III reports results for the case where the propensity score was known, and Table IV reports results for when it was estimated, using a Nadaraya-Watson kernel estimator with a normal kernel function and the Silverman rule of thumb for bandwidth selection. As with the binary choice model, the MSE does not decline at the parametric rate, although this only becomes apparent in the larger sample sizes.

Table V reports results for the smooth maximum score estimator, using integrals of the normal and logistic p.d.f.s as kernel functions and a bandwidth sequence of the sample size raised to the minus one fifth. The design simulated here involved one regressor distributed standard normal, and a homoskedastic standard normal error. The slower than root- n rate of convergence is demonstrated here as well. Note that the MSE's tabulated here can give indication of the relative efficiency of the SMS compared to the bound derived in this paper.

Note the minimum MSE for the simple design here can be calculated from the expression on page 14, which provides us with a value of 1.623. For the normal kernel case, $n^{4/5}$ times the calculated MSE gives us values of 21.350, 4.163, 3.431, 3.105, and for the logistic kernel 332.149, 12.108, 11.743, 11.661. We note that the difference between the results attained in the simulation study and the bound are likely due to our choice of bandwidth selection, which is suboptimal compared to that proposed in Horowitz(1992).

6 Conclusions

This paper considered developing a notion for efficiency bounds for finite dimensional parameters in a class of models that are irregular in the sense that these parameters cannot be (Hajek-regularly) estimated at the parametric rate. As in existing work our evaluation of bounds will be associated with subconvex loss and minimax risk functions, but our results differ in the sense that there exist a sequence of estimators which are regular with respect to a normalizing sequence which diverges at a rate slower than the square-root of the sample size. Under our choice of loss and risk functions optimality criteria are directly related to minimizing asymptotic mean squared error.

We apply these results to three specific models which have already been shown in the literature to be irregular under the stated conditions- the regression coefficients in two binary choice models, and the average treatment effect in a treatment model under exogenous selection.

The work in this paper immediately suggests areas for future research. For one, it would be useful to evaluate the relative efficiency of exiting estimation procedures for these models that have been proposed in the literature. Second, it would be useful to use these bounds to propose estimators which are asymptotically efficient in the sense their limiting distributions attain or come arbitrarily close to attaining these bounds. Finally, are notion of efficiency and method of deriving bounds could be applied to other irregularly identified models - see, e.g. Hahn(1994), and Ridder and Wouterstan(2003). We leave all of these for future work.

References

1. Andrews, D.W.K. (1991), "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models, *Econometrica*, 59, 307-346.

2. Andrews, D.W.K. (1994), "Empirical Process Methods in Econometrics", in Engle, R.F. and D.McFadden (eds.), *Handbook of Econometrics, Vol. 4*, Amsterdam: North-Holland.
3. Begun, J., W. Hall, W. Huang, and J. Wellner (1983), "Information and Asymptotic Efficiency in parametric-nonparametric models", *Annals of Statistics*, 11, 432-452.
4. Chamberlain, G. (1986), "Asymptotic Efficiency in Semiparametric Models with Censoring", *Journal of Econometrics*, 32, 189-218.
5. Chernozhukov, V. and Hong, H. (2004), "Likelihood Estimation and Inference in a Class of Nonregular Econometric Models", *Econometrica*, 72, 1445-1480.
6. Hajek, J. (1970), "A Characterization of Limiting Distributions of Regular Estimates", *A. Wahrscheinlichkeitstheorie verw. Geb.*, 14, 323-330.
7. Hirano, K., Imbens G. and G. Ridder(2003) "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica* 71, 1161-1189.
8. Hirano, K. and Porter, J.R. (2003), "Asymptotic Efficiency in Parametric Structural Models with Parameter-Dependent Support", *Econometrica* 71 (5), 1307–1338.
9. Hahn, J.(1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315-332.
10. Ibragimov, I.A. and R.Z. Has'minskii(1981), *Statistical Estimation Asymptotic Theory*, New York, NY: Springer.
11. Horowitz, J.L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505-531.
12. Horowitz, J.L. (1993), "Optimal Rates of Convergence of Parameter Estimators in the Binary Response Model with Weak Distributional Assumption" *Econometric Theory*, 9, 1-18.
13. Koul H. and V. Susarla and J. Van Rysin (1981) "Regression Analysis with Randomly Right Censored Data", *Annals of Statistics*, 9, pp 1276 - 1288.
14. Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, New York, NY: Springer.
15. Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation With Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121.
16. Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145-177.

17. Manski, C.F. (1975), “Maximum Score Estimation of the Stochastic Utility Model of Choice”, *Journal of Econometrics*, 3, 205-228
18. Manski, C.F. (1985), “Semiparametric Analysis of Discrete Response: Asymptotic Properties of Maximum Score Estimation”, *Journal of Econometrics*, 27, 313-334
19. Newey, W.K. (1990), “Semiparametric Efficiency Bounds”, *Journal of Applied Econometrics*, 5, 99-135.
20. Newey, W.K. and D. McFadden (1994) “Estimation and Hypothesis Testing in Large Samples”, in Engle, R.F. and D. McFadden (eds.) , *Handbook of Econometrics, Vol. 4*, Amsterdam: North-Holland.
21. Pfanzangle, J, and W. Wefelmeyer (1982), *Contributions to a General Asymptotic Statistical Theory*, New York, NY: Springer.
22. Severini, T.A. and G. Tripathi(2003), “A Simplified Approach to Computing Bounds in Semiparametric Models”, *Journal of Econometrics*, 102, 23-66.
23. Stein, C. (1956), “Efficient Nonparametric Testing and Estimation”, *in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA: University of California Press.
24. van der Vaart, A.W.(1991), *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press.
25. van der Vaart, A.W. and J. Wellner(2000), *Weak Convergence and Empirical Processes*, New York, NY: Springer.

A Appendix: the Optimal Kernel Structure for the Smoothed Maximum Score

We consider for simplicity the cases where $k = 1, 2$. For $k = 1$

$$\beta_h = \beta_0 + \frac{h^2}{2} \frac{\mathcal{P}''(-\beta_0)}{\mathcal{P}'(-\beta_0)^3} \int K(u)u^2 du + o(h^2).$$

Similarly, for $k = 2$ if $K(\cdot)$ is a second-order kernel

$$\beta_h = \beta_0 + \frac{h^4}{24} \frac{10\mathcal{P}''''(-\beta_0)\mathcal{P}''(-\beta_0)\mathcal{P}'(-\beta_0) - \mathcal{P}''''(-\beta_0)\mathcal{P}''(-\beta_0)^2 - 15\mathcal{P}''(-\beta_0)^3}{\mathcal{P}'(-\beta_0)^7} \int K(u)u^4 du + o(h^2).$$

We find variance using the asymptotic approximation for $\hat{\beta}$ using the fact that h is small:

$$\begin{aligned}\text{Var}(\hat{\beta}) &= \frac{1}{Nh^2} \int \frac{K\left(\frac{P(x)-\frac{1}{2}}{h}\right)^2}{f_X(x)} \mathcal{P}'(x)^2 x^2 dx + o(N^{-1}h^{-2}) \\ &= \frac{1}{Nh} \int \frac{K^2(u)}{f_X(\mathcal{P}^{-1}(\frac{1}{2}+uh))} \mathcal{P}'(\mathcal{P}^{-1}(\frac{1}{2}+uh)) (\mathcal{P}^{-1}(\frac{1}{2}+uh))^2 du \\ &= \frac{1}{Nh} \beta_0^2 \frac{\mathcal{P}'(-\beta_0)}{f_X(-\beta_0)} \int K^2(u) du + o(N^{-1}h^{-1}).\end{aligned}$$

Relying on smoothness, we find that normalization $\alpha = 2/5$ for $k = 1$ and $\alpha = 4/9$ for $k = 2$. Then we can find the mean-squared errors. In case of $k = 1$

$$N^{4/5} \text{MSE}(\hat{\beta}) = \beta_0^{8/5} \frac{5}{4} \left(\frac{\mathcal{P}''(-\beta_0)}{\mathcal{P}'(-\beta_0) f_X^2(-\beta_0)} \right)^{2/5} \left[\int K^2(u) du \right]^{4/5} \left[\int u^2 K(u) du \right]^{2/5}.$$

In case where $k = 2$

$$N^{8/9} \text{MSE}(\hat{\beta}) = \beta_0^{16/9} \frac{3^{20/9}}{4} A^{1/9} \left[\int K^2(u) du \right]^{8/9} \left[\int u^4 K(u) du \right]^{1/9},$$

where

$$A = \frac{10\mathcal{P}'''(-\beta_0) \mathcal{P}''(-\beta_0) \mathcal{P}'(-\beta_0) - \mathcal{P}''''(-\beta_0) \mathcal{P}''(-\beta_0)^2 - 15\mathcal{P}'''(-\beta_0)^3}{f_X^8(-\beta_0)} \mathcal{P}'(-\beta_0).$$

We need to minimize over possible $K(\cdot)$ simultaneously with h . Suppose that $K(\cdot)$ is s times differentiable with support on $[-1, 1]$ where $s > 2k$. We look for the solution in the form of expansion in terms of normalized Legendre polynomials

$$K(x) = \sum_{i=0}^s a_i L_i(x).$$

For $k = 1$ we will solve the problem by noticing that $1 = \sqrt{2}L_0(x)$ and $x^2 = \frac{\sqrt{2}}{3}L_0(x) + \frac{2\sqrt{10}}{15}L_2(x)$. Then, omitting the terms not related to the kernel, we can see that the problem is equivalent to the standard constrained optimization for

$$\left[\int K^2(u) du \right]^{4/5} \left[\int u^2 K(u) du \right]^{2/5} = \left(\sum_{i=0}^s a_i^2 \right)^{4/5} \left(\frac{\sqrt{2}}{3} a_0 + \frac{2\sqrt{10}}{15} a_2 \right)^{2/5},$$

which minimized with respect to a_0, \dots, a_K with restriction

$$\int K(u) du = \int L_0(u) \cdot 1 du = \sqrt{2} a_0 = 1.$$

In these expressions we used orthonormality of the polynomial series. This immediately leads to the result $a_0 = \frac{1}{\sqrt{2}}$. Substituting this to the objective function we can write the first-order condition to the form

$$2\sqrt{10} \left(a_2 + \frac{1}{\sqrt{10}} \right)^2 \left(\frac{1}{2} + a_2^2 \right)^{1/5} \left(\frac{1}{3} + \frac{2\sqrt{10}}{15} a_2 \right)^{3/5} = 0.$$

The objective has a unique minimum $a_2 = -\frac{1}{\sqrt{10}}$. Therefore, the optimal kernel has Epanechnikov structure

$$K(u) = \frac{1}{\sqrt{2}}L_0(u) - \frac{1}{\sqrt{10}}L_2(u) = \frac{3}{4}(1 - u^2).$$

As a result, the minimal mean-squared error is

$$N^{4/5}\text{MSE}(\hat{\beta}) = \frac{5}{4} \left(\frac{9\beta_0^4}{80} \right)^{2/5} \left(\frac{\mathcal{P}''(-\beta_0)}{\mathcal{P}'(-\beta_0) f_X^2(-\beta_0)} \right)^{2/5}.$$

To extend this result to the case where $k = 2$ we note that

$$x^4 = \frac{7\sqrt{2}}{35}L_0(x) + \frac{4\sqrt{10}}{35}L_2(x) + \frac{8\sqrt{2}}{105}L_4(x).$$

Then the objective function can be written as

$$\left[\int K^2(u) du \right]^{8/9} \left[\int u^4 K(u) du \right]^{1/9} = \left(\sum_{i=0}^s a_i^2 \right)^{8/9} \left(\frac{7\sqrt{2}}{35}a_0 + \frac{4\sqrt{10}}{35}a_2 + \frac{8\sqrt{2}}{105}a_4 \right)^{1/9}.$$

The constraints are provided by the fact that we consider a second-order kernel, so that

$$\int K(u) du = \int L_0(u) \cdot 1 du = \sqrt{2}a_0 = 1,$$

and

$$\int u^2 K(u) du = \frac{\sqrt{2}}{3}a_0 + \frac{2\sqrt{10}}{15}a_2 = 0.$$

From the constraints it immediately follows that $a_0 = \frac{1}{\sqrt{2}}$ and that $a_2 = -\frac{\sqrt{10}}{4}$. Substituting the numbers into the objective function and solving the first-order condition we can see that a unique minimum is attained at $a_4 = \frac{36\sqrt{2}-3\sqrt{38}}{136}$.