The Influence Function of Semiparametric Estimators*

Hidehiko Ichimura
University of Tokyo

Whitney K. Newey
MIT

July 2016

Abstract

Often semiparametric estimators are asymptotically equivalent to a sample average. The object being averaged is referred to as the influence function. The influence function is useful in formulating primitive regularity conditions for asymptotic normality, in efficiency comparions, for bias reduction, and for analyzing robustness. We show that the influence function of a semiparametric estimator can be calculated as the limit of the Gateaux derivative of a parameter with respect to a smooth deviation as the deviation approaches a point mass. We also consider high level and primitive regularity conditions for validity of the influence function calculation. The conditions involve Frechet differentiability, nonparametric convergence rates, stochastic equicontinuity, and small bias conditions. We apply these results to examples.

JEL Classification: C14, C24, H31, H34, J22
Keywords: Influence function, semiparametric estimation, bias correction.

*The NSF and JSPS provided partial financial support. We are grateful for comments by V. Chernozhukov, K. Kato, U. Mueller, J. Porter and participants at seminars at UC Berkeley, NYU, University of Kansas, and Yale.

1 Introduction

There are many economic parameters that depend on nonparametric first steps. Examples include games, dynamic discrete choice, average consumer surplus, and treatment effects. Often those estimators are asymptotically equivalent to a sample average. The thing being averaged is referred to as the influence function. The influence function is useful for a number of purposes. It can be used to construct estimators with improved properties, Chernozhukov, Escanciano, Ichimura, and Newey (2016). Its variance is the asymptotic variance of the estimator and so it can be used for asymptotic variance estimation and asymptotic efficiency comparisons. Also, the form of remainder terms follow from the form of the influence function so knowing the influence function is a good starting point in formulating regularity conditions. Furthermore, the influence function approximately gives the influence of a single observation on the estimator, and so can be used for robustness comparisons. Indeed this interpretation is where the influence function gets its name in the robust estimation literature, see Hampel (1974).

BETTER We show how the influence function can be calculated as the limit of a derivative of the object the estimator converges to. The derivative is taken with respect to a weight on a general alternative to the true CDF. The limit is taken as the alternative approaches the CDF of constant. We impose restrictions on the alternative so that the object being differentiated is well defined. This calculation is similar to that of FIX Von Mises (1947) and Hampel (1974), of a derivative with respect to a CDF of a constant, but we specify the alternative so that it is smooth and also allow it to satisfy other restrictions. As a result the calculation given here applies generally to semiparametric estimator including those where the first step is a conditional expectation or density.

We also consider regularity conditions for validity of the influence function calculation. The conditions involve Frechet differentiability as well as convergence rates for nonparametric estimators. They also involve stochastic equicontinuity and small bias conditions. When estimators depend on nonparametric objects like conditional expectations and pdf's, the Frechet differentiability condition is generally satisfied for intuitive norms, e.g. as is well known from Goldstein and Messer (1992). The situation is different for functionals of the empirical distribution where Frechet differentiability is only known to hold under special norms, Dudley (1994). The asymptotic theory here also differs from functionals of the empirical distribution in other ways as will be discussed below.

Newey (1994) previously showed that the influence function of a semiparametric estimator can be obtained by solving a functional equation involving pathwise derivatives and scores of parametric models. That approach has proven useful in many settings but does require the

solution to a functional equation. The approach of this paper is an explicit calculation that does not require finding the solution to a functional equation. Here we simply calculate a derivative and find its limit. This calculation is accomplished by specifying a parametric model, i.e. a path, in the right way to obtain the influence function.

Regularity conditions for functionals of nonparametric estimators involving Frechet differentiability have previously been formulated by Ait-Sahalia (1991), Goldstein and Messer (1992), Newey and McFadden (1994), Newey (1994), Chen and Shen (??), Chen, Linton, and Keilegom (2003), and Ichimura and Lee (2010), among others. Newey (1994) gave stochastic equicontinuity and small bias conditions for functionals of series estimators. In this paper we update those using Belloni, Chernozhukov, Chetverikov, and Kato (2015). Bickel and Ritov (2003) formulated similar conditions for kernel estimators. Andrews (2004) gave stochastic equicontinuity conditions for the more general setting of GMM estimators that depend on nonparametric estimators.

In Section 2 we describe the estimators we consider. Section 3 presents the method for calculating the influence function. In Section 4 we outline some conditions for validity of the influence function calculation. Section 5 gives primitive conditions for linear functionals of kernel density and series regression estimators. Section 6 outlines additional conditions for semiparametric GMM estimators. Section 7 concludes.

2 Semiparametric Estimators

This paper is about estimators where parameters of interest depend on a first step nonparametric estimator. We refer to these estimators as semiparametric. We could also refer to them estimators where nonparametric first step estimators are "plugged in." This terminology seems awkward though, so we simply refer to them as semiparametric estimators. We denote such an estimator by $\hat{\beta}$, which is a function of the data $z_1, ..., z_n$ where n is the number of observations. Throughout the paper we will assume that the data observations z_i are i.i.d. We denote the object that $\hat{\beta}$ estimates as β_0, the subscript referring to the parameter value under the distribution that generated the data.

We adopt a general framework where the estimator of the parameter of interest is a generalized method of moments estimator depending on a nonparametric first step. To describe the type of estimator we consider let $m(z, \beta, \gamma)$ denote a vector of functions of the data observation z, parameters of interest $\beta \in \mathbb{R}^q$, and a function γ that may be vector valued. Here γ represents some possible value of a nonparametric estimator. A GMM estimator can be based on a

moment condition where β_0 is the unique parameter vector satisfying

$$E[m(z_i, \beta_0, \gamma_0)] = 0, \tag{2.1}$$

and γ_0 is the true γ. Here it is assumed that this moment condition identifies β. Let $\hat{\gamma}$ denote some nonparametric estimator of γ_0. Plugging in $\hat{\gamma}$ to obtain $m(z_i, \beta, \hat{\gamma})$ and averaging over z_i gives the estimated sample moments $\hat{m}(\beta) = \sum_{i=1}^n m(z_i, \beta, \hat{\gamma})/n$. For $\hat{W}$ a positive semi-definite weighting matrix a semiparametric GMM estimator is

$$\hat{\beta} = \arg\min_{\beta \in B} \hat{m}(\beta)^T \hat{W} \hat{m}(\beta).$$

We note that this class of estimators includes an explicit functional $\mu(F)$ of the distribution F of a single observations, where $m(z, \beta, \gamma) = \mu(F) - \beta$ and $F = \gamma$. Many other estimators are also included as special cases.

Examples can help illustrate the results. We consider one example here and more below. The first example is an estimator of a bound on average surplus of a price change when there are bounds on income effects, as in Hausman and Newey (2016a,b). Let y denote quantity consumed of some good, $x = (x_1, x_2)'$ where x_1 is price, x_2 is income, $w_2(x_2)$ be a weight function for income (such as an indicator for some interval), and $\gamma_1(x)$ a possible conditional expectation function $E[y_i|x_i = x]$. We assume that B is a uniform bound on the derivative of demand with respect to income, i.e. the income effect. The object of interest is a bound on the weighted average over income of equivalent variation for a price change from $\breve{x}_1$ to $\bar{x}_1$, given by

$$\beta_0 = E[w_2(x_{2i}) \int_{\breve{x}_1}^{\bar{x}_1} \gamma_{10}(u, x_{2i}) e^{-B(u - \breve{x}_1)} du] = E[w_2(x_{2i}) \int w_1(u) \gamma_{10}(u, x_{2i}) du],$$

where $w_1(x_1) = 1(\breve{x}_1 \leq x_1 \leq \bar{x}_1) e^{-B(x_1 - \breve{x}_1)}$. If B is an upper (lower) bound on income effects then β_0 is a lower (upper) bound on average equivalent variation over income and individual heterogenetiy of a price change from $\breve{x}_1$ to $\bar{x}_1$. This object is identified from the semiparametric moment function.

$$m(z, \beta, \gamma_1) = w_2(x_2) \int w_1(u) \gamma_1(u, x_2) du - \beta.$$

We will consider additional examples below.

The results of this paper apply generally to asymptotically linear estimators. An asymptotically linear estimator is one satisfying

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sum_{i=1}^n \psi(z_i)/\sqrt{n} + o_p(1), E[\psi(z_i)] = 0, E[\psi(z_i)^T \psi(z_i)] < \infty. \tag{2.2}$$

The function $\psi(z)$ is referred to as the influence function, following terminology of Hampel (1974). It gives the influence of a single observation in the leading term of the expansion in

equation (2.2). It also quantifies the effect of a small change in the distribution on the limit of $\hat{\beta}$ as we further explain below.

A semiparametric GMM estimator will be asymptotically linear under regularity conditions that are summarized below. Let

$$M = \left.\frac{\partial E[m(z_i, \beta, \gamma_0)]}{\partial \beta}\right|_{\beta=\beta_0}, W = plim(\hat{W}).$$

A standard expansion argument along with well understood properties of semiparametric estimators leads to an influence function of the form

$$\psi(z) = -(M^T W M)^{-1} M^T W [m(z, \beta_0, \gamma_0) + \phi(z)], \tag{2.3}$$

where $\phi(z)$ is an adjustment term for the estimator $\hat{\gamma}$ of γ_0, as discussed in Newey (1994). Here $m(z, \beta_0, \gamma_0) + \phi(z)$ will be the influence function of $\hat{m}(\beta_0)$. This formula for the influence function is valid under weak regularity conditions, that allow for $m(z, \beta, \gamma_0)$ to not be smooth in β, e.g. as in Chen, Linton, et al. (??).

3 Calculating the Influence Function

In this Section we provide a method for calculating the influence function. The key object on which the influence function depends is the limit of the estimator when z_i has a CDF F that is unrestricted except for regularity conditions. We denote this object by $\beta(F)$. One can think of $\beta(F)$ as the object that is estimated by $\hat{\beta}$ when misspecification is allowed. The idea is that every estimator converges to something under some regularity conditions. The function $\beta(F)$ is that something. It describes how the limit of the estimator varies as the distribution of a data observation varies. Formally, it is mapping from a set $\mathcal{F}$ of CDF's into real vectors,

$$\beta(\cdot) : \mathcal{F} \longrightarrow \Re.$$

In the surplus bound example

$$\beta(F) = \int w_2(\tilde{x}_2) w_1(\tilde{x}_1) E_F[y_i | x_i = \tilde{x}] d\tilde{x}_1 F_2(d\tilde{x}_2), \tag{3.4}$$

where $E_F[y_i|x_i]$ denotes the conditional expectation under distribution F and $F_2(x_2)$ is the marginal CDF of x_{2i}.

How $\beta(F)$ varies as F varies near the true distribution F_0 can be used to calculate the influence function. An important feature of $\beta(F)$ is that it may only be well defined when F is restricted in some way. In the average surplus example $\beta(F)$ will only be well defined when the uniform distribution on $(\breve{x}_1, \bar{x}_1)$ is absolutely continuous with respect to the distribution of

x_{1i}. In formal terms this feature means that the domain $\mathcal{F}$ of $\beta(\cdot)$ is restricted. To allow for a restricted domain we consider only variations in F that are contained in $\mathcal{F}$. The specific kind of variation we consider is a convex combination $F_\tau = (1-\tau)F_0 + \tau G_z^j$ of the true distribution F_0 with some other distribution G_z^j where $F_\tau \in \mathcal{F}$. The superscript j and subscript z designate G_z^j as a member of sequence of CDF's approaching the CDF of the constant z. Under conditions given below the influence function can be calculated as

$$\psi(z) = \lim_{j \longrightarrow \infty} \left[\frac{d}{d\tau} \beta((1-\tau) \cdot F_0 + \tau \cdot G_z^j) \right], \tag{3.5}$$

where all derivatives with respect to τ are right derivatives at $\tau = 0$. The derivative in this expression is the Gateaux derivative of the functional $\beta(F)$ with respect to a deviation $\tau[G_z^j - F_0(z)]$ from the true distribution F_0. This formula says that the influence function is the limit of this Gateaux derivative as G_z^j approaches the CDF of the constant z.

Equation (3.5) can be thought of as a generalization of the influence function calculation of Von Mises (1947) and Hampel (1974). That calculation is based on $G_z^j = \lambda_z$ where λ_z is the CDF of the constant z. If $(1-\tau) \cdot F_0 + \tau \cdot \lambda_z$ is in $\mathcal{F}$ then the influence function is given by the Gateaux derivative

$$\psi(z) = \frac{d}{d\tau} \beta((1-\tau) \cdot F_0 + \tau \cdot \lambda_z)$$

The problem with this formula is that $F_\tau = (1-\tau) \cdot F_0 + \tau \cdot \lambda_z$ will not be in the domain $\mathcal{F}$ for many semiparametric estimators. In many cases $F \in \mathcal{F}$ (i.e. $\beta(F)$ being well defined) reqiures that certain marginal distributions of F are continuous. The CDF $(1-\tau) \cdot F_0 + \tau \cdot \lambda_z$ does not satisfy that restriction. Equation (3.5) circumvents this problem by restricting F_τ to be in $\mathcal{F}$. The influence function is then obtained as the limit of a Gateaux derivative as $G_z^j \longrightarrow \lambda_z$ rather than the Gateaux derivative with respect to the CDF of a point. This generalization applies to most semiparametric estimators.

We can relate equation (3.5) to the pathwise derivative characterization of the influence function in Newey (1994). Denote one of a class of parametric models as F_θ, where θ denotes a vector of parameters, with $F_\theta \in \mathcal{F}$ equal to the true distribution F_0 at $\theta = 0$. Restrict each parametric model in the class to be regular in the sense used in the semiparametric efficiency bounds literature, so that F_θ has a score $S(z)$ (derivative of the log-likelihood in many cases, e.g. see Van der Vaart, 1998, p. 362) at $\theta = 0$ and possibly other conditions are satisfied. We assume that the set of scores over all regular parametric families has mean square closure that includes all functions with mean zero and finite variance. This assumption is the precise meaning of the statment that we are not restricting F except for regularity conditions. As shown by Newey (1994) the influence function $\psi(z)$ is then the unique solution to the derivative

equation of Van der Vaart (1991),

$$\frac{\partial \beta(F_\theta)}{\partial \theta} = E[\psi(z_i)S(z_i)], E[\psi(z_i)] = 0, \tag{3.6}$$

as the score $S(z)$ varies over those for regular parametric models.

The advantage of equation (3.5) is that it is a direct calculation while the outer product formula (3.6) is a functional equation that must be solved to find the influence function. It is true that Newey (1994), Hahn (1998), Hirano, Imbens, and Ridder (2003), Bajari, Hong, Krainer, and Nekipelov (2010), Bajari, Chernozhukov, Hong, and Nekipelov (2009), Hahn and Ridder (2013, 2016), and Ackerberg, Chen, Hahn, and Liao (2015) have solved equation (3.6) for important models. Generally though, these results have required the solution to a functional equation, such as a Riesz representation in Propositions 4 and 5 of Newey (1994) and in Ackerberg et. al. (2015). No such solution is required to apply equation (3.5). Instead, all that is required is an expression for $\partial \beta(F_\tau)/\partial \tau$ and for its limit. This advantage of equation (3.5) is highlighted in examples to follow. MOVE.

To use the derivative formula to calculate the influence function we need to specify G_z^j. Various kinds of restrictions on G_z^j may be needed to insure that $F_\tau \in \mathcal{F}$. In the surplus bound example the Lebesgue measure on $[\check{x}_1, \bar{x}_1]$ must be absolutely continuous with respect to the distribution of the price variable x_{1i} for $\beta(F)$ to be well defined. In other examples an identification condition may need to be satisfied. We are free to choose G_z^j in whatever way is convenient for imposing these restrictions and ensuring that equation (3.5) holds. Here we use

$$G_z^j(\tilde{z}) = E[1(z_i \leq \tilde{z})\delta(z_i)], \tag{3.7}$$

where $\delta(z_i)$ is a bounded nonnegative function with $E[\delta(z_i)] = 1$. The variable $\tilde{z}$ represents a possible value of the random variable z_i, and we supress a j superscript and z subscript on $\delta(z_i)$ for notational convenience. We will assume that $z \in \mathbb{R}^r$ and that F_0 is absolutely continuous with respect to a product measure μ on $\mathbb{R}^r$. This assumption allows for components of z_i to be continuously or discretely distributed, or some mixture of the two. The distribution $F_\tau = (1-\tau)F_0 + \tau G_z^j$ will have a pdf with respect to μ given by

$$f_\tau(\tilde{z}) = f_0(\tilde{z})[1 - \tau + \tau\delta(\tilde{z})] = f_0(\tilde{z})[1 + \tau S(\tilde{z})], S(\tilde{z}) = \delta(\tilde{z}) - 1.$$

Also, for any measureable function y_i and components x_i of z_i the marginal pdf $f_\tau(\tilde{x})$ of x_i, conditional expectation $E_\tau[y_i|x_i]$ of y_i given x_i, and its derivative are

$$\begin{aligned} f_\tau(\tilde{x}) &= f_0(\tilde{x})\{1 + \tau E[S(z_i)|x_i = \tilde{x}]\}, \\ E_\tau[y_i|x_i] &= \frac{E[y_i|x_i] + \tau E[y_i S(z_i)|x_i]}{1 + \tau E[S(z_i)|x_i]}, \frac{\partial E_\tau[y_i|x_i]}{\partial \tau} = E[\{y_i - E[y_i|x_i]\}S(z_i)], \end{aligned} \tag{3.8}$$

as shown in Lemma A1 of the Appendix. These formulae will be useful for calculating the influence function in many cases.

The characterization of the influence function in equation (3.6) justifies the influence function formula (3.5) for this choice of G_z^j . Consider $F_\tau = (1-\tau)\cdot F_0 + \tau \cdot G_z^j$ as a model with parameter τ passing through the truth at $\tau = 0$. This model is regular, as shown in the proof of Theorem 1 below. Also, the score is given by $\partial \ln f_\tau(\tilde{z})/\partial\tau = S(\tilde{z})$. Then equation (3.6) and $E[S(z_i)] = 0$ gives

$$\frac{\partial \beta(F_\tau)}{\partial \tau} = E[\psi(z_i)\delta(z_i)] = \int \psi(\tilde{z})\delta(\tilde{z})f_0(\tilde{z})d\mu.$$

At any point z where $\psi(\tilde{z})$ is continuous, equation (3.5) will follow by taking the limit as $\delta(\tilde{z})f_0(\tilde{z})$ approaches a spike at z, which corresponds to G_z^j approaching the CDF of a constant at z.

A particular choice of $\delta(z)$ will be useful in the calculations. Let $K(u)$ be a pdf that is symmetric around zero, has bounded support, and is continuously differentiable of all orders with bounded derivatives. The smoothness of K is useful for some of the examples. Also let $\bar{\mu}_\ell^j = j^{-1}\int K((z_\ell - \tilde{z}_\ell)/j)d\mu_\ell(\tilde{z}_\ell)$, and

$$g(\tilde{z}) = \Pi_{\ell=1}^r \kappa_\ell^j(\tilde{z}_\ell), \kappa_\ell^j(\tilde{z}_\ell) = \left(j\bar{\mu}_\ell^j\right)^{-1} K((z_\ell - \tilde{z}_\ell)/j). \tag{3.9}$$

We take $\delta(\tilde{z})$ to be the function

$$\delta(\tilde{z}) = g(\tilde{z})1(f_0(\tilde{z}) \geq 1/j)f_0(\tilde{z})^{-1}. \tag{3.10}$$

For this choice of $\delta(\tilde{z})$ equation (3.5) will hold when $\psi(\tilde{z})$ is continuous at z and $f_0(\tilde{z})$ is bounded away from zero on a set of $\tilde{z}$ that has full μ measure locally to z, as shown and further discussed below.

We can calculate the influence function for the surplus bound using this $\delta(z)$. We assume that the joint pdf $f_0(\tilde{y},\tilde{x})$ of (y_i, x_i) is bounded away from zero on a neighborhood of z so that $\delta(z) = g(z)/f_0(\tilde{z})$ where $g(z)$ is positive for large enough j. Now assume j is large enough so that holds. Then for any $a(y,x)$ we have $E[a(y_i,x_i)S(z_i)|x_i = \tilde{x}] = \int a(\tilde{y},\tilde{x})g(\tilde{z})d\tilde{y}/f_0(\tilde{x})$. Let

$$\begin{aligned}
\Delta &= \int w_2(\tilde{x}_2)w_1(\tilde{x}_1)E[\{y_i - \gamma_{10}(x_i)\}S(z_i)|x_i = \tilde{x}]d\tilde{x}_1 f_2(\tilde{x}_2)d\mu_{x_2} \\
&= \int \{w_2(\tilde{x}_2)w_1(\tilde{x}_1)[\tilde{y} - \gamma_{10}(\tilde{x})]f_2(\tilde{x}_2)/f_0(\tilde{x})\}g(\tilde{z})d\mu = \int \alpha(\tilde{x})[\tilde{y} - \gamma_1(\tilde{x})]g(\tilde{z})d\mu, \\
\alpha(\tilde{x}) &= w_2(\tilde{x}_2)w_1(\tilde{x}_1)f_2(\tilde{x}_2)/f_0(\tilde{x}),
\end{aligned}$$

where μ_{x_2} is the dominating measure for x_{2i} and $f_2(\tilde{x}_2)$ the pdf of x_{2i} with respect to this measure. Then plugging F_τ in equation (3.4) and applying the chain rule gives

$$\frac{d}{d\tau}\beta(F_\tau) = E[w_2(x_{2i})\left\{\int w_1(\tilde{x}_1)\gamma_{10}(\tilde{x}_1, x_{2i})d\tilde{x}_1\right\}S(z_i)] + \Delta = \int m(\tilde{z},\beta_0,\gamma_{10})g(\tilde{z})d\mu + \Delta,$$

where the second equality follows by iterated expecations, the definition of $m(z, \beta, \gamma_1)$ above, and by $E[S(z_i)] = 0$. Assume that $\gamma_{10}(\tilde{x})$, $w_1(\tilde{x}_1)$, $w_2(\tilde{x}_2)$, and $f_0(\tilde{x}_1|\tilde{x}_2) = f_{20}(\tilde{x}_2)/f_0(\tilde{x})$ are continuous at x, so that $\alpha(\tilde{x})[\tilde{y} - \gamma_{10}(\tilde{x})]$ is continuous at z. Also assume that $\gamma(\tilde{x}_1, \tilde{x}_2)$ is continuous at $(\tilde{x}_1, x_2)$ for all $\tilde{x}_1 \in (\underline{x}_1, \bar{x}_1)$, so that $\int w_1(\tilde{x}_1)\gamma_{10}(\tilde{x}_1, \tilde{x}_2)d\tilde{x}_1$ is continuous at x_2. Then $m(\tilde{z}, \gamma_{10}, \beta_0)$ is continuous at z. By the construction of $g(\tilde{z})$ we have $\int a(\tilde{z})g(\tilde{z})d\mu \longrightarrow a(z)$ for any $a(\tilde{z})$ function that is continuous at z. Therefore as $j \longrightarrow \infty$ we have

$$\frac{d}{d\tau}\beta(F_\tau) \longrightarrow m(z, \beta_0, \gamma_{10}) + \alpha(x)[y - \gamma_{10}(x)]. \tag{3.11}$$

We can also characterize the influence function in this example using Proposition 4 of Newey (1994). To do so we need to find the solution ot the Riesz representation in equation (4.4) of Newey (1994). Multiplying and dividing by $f_0(\tilde{x}_1|\tilde{x}_2)$ gives

$$E[w_2(x_{2i})\{\int w_1(\tilde{x}_1)\gamma_1(\tilde{x}_1, x_{2i})d\tilde{x}_1\}] = E[\alpha(x_i)\gamma_1(x_i)].$$

Here we see that $\alpha(x_i)$ is the Riesz representor in Proposition 4 of Newey (1994), so the conclusion of that result implies that the influence function of the surplus bound is the expression on the right of equation (3.11).

This example shows the advantage of the derivative formula in equation (3.5) over solving the functional equation (3.6). In the example the influence function was derived by a straighforward calculation of a limit. At no point in that calculation did we need to solve for the function $\alpha(x)$. Instead the expression for $\alpha(x)$ emerged from the derivative calculation. Thus this new example demonstrates how the influence function can be obtained from a derivative without solving a functional equation. We will show this approach is similarly useful in an even more challenging and original example in the next Section.

We give a precise theoretical justification for the formula in equation (3.5) by assuming that an estimator is asymptotically linear and then showing that equation (3.5) is satisfied under a few mild regularity conditions. One of the regularity conditions we use is local regularity of $\hat{\beta}$ along the path F_τ. This property is that for any $\tau_n = O(1/\sqrt{n})$, when $z_1, ..., z_n$ are i.i.d. with distribution F_{τ_n},

$$\sqrt{n}[\hat{\beta} - \beta(F_{\tau_n})] \xrightarrow{d} N(0, V), V = E[\psi(z_i)\psi(z_i)^T].$$

That is, under a sequence of local alternatives, when $\hat{\beta}$ is centered at $\beta(F_\tau)$, then $\hat{\beta}$ has the same limit in distribution as for F_0. This is a very mild regularity condition. Many semiparametric estimators could be shown to be uniformly asymptotically normal for τ in a neighborhood of 0, which would imply this condition. Furthermore, it turns out that asymptotic linearity of $\hat{\beta}$ and Gateaux differentiability of $\beta(F_\tau)$ at $\tau = 0$ with the correct derivative are sufficient for

local regularity. For these reasons we view local regularity as a mild condition for the influence function calculation.

We will prove that (3.5) is valid for G_z^j as specified in equation (3.7). It would be straightforward to extend this validity result to more general classes of G_z^j but the result we give should suffice for most cases.

THEOREM 1: *Suppose that $\hat{\beta}$ is asymptotically linear with influence function $\psi(z)$ and there is an open set $\mathcal{N}$ containing z such that a) there is $\bar{\mathcal{N}} \subseteq \mathcal{N}$ such that $\mu(\mathcal{N}) = \mu(\bar{\mathcal{N}})$ and $\psi(\tilde{z})$ is continuous at z for $\tilde{z} \in \bar{\mathcal{N}}$; b) there is $\varepsilon > 0$ such that*

$$\mu(\mathcal{N} \cap \{z : f_0(z) \geq \varepsilon\}) = \mu(\mathcal{N}).$$

If $\hat{\beta}$ is locally regular for the parametric model $(1-\tau)F_0 + \tau G_z^j$ for τ in a neighborhood of zero then $d\beta(F_\tau)/\partial\tau$ exists and satisfies equation (3.5).

PROOF

This result shows that if an estimator is asymptotically linear and locally regular then the influence function satisfies equation (*3.5*), justifying that calculation. This result is like Van der Vaart (1991) in having differentiability of $\beta(F_\tau)$ as a conclusion. It differs in restricting the paths to have the form $(1-\tau)F_0 + \tau G_z^j$. Such a restriction on the paths actually weakens the local regularity hypothesis because $\hat{\beta}$ only has to be locally regular for a particular kind of path rather than the general class of paths in Van der Vaart (1991). We view locally regularity for such paths as a very weak condition because the deviations are bounded smooth densities. We expect that these deviations are regular enough so that F_τ will generally satisfy whateever regularity conditions are needed for asymptotic linearity uniformly in τ near zero, so $\hat{\beta}$ will be locally regular. The conditions of Theorem 1 are stronger than Van der Vaart (1991) in assuming that the influence function is continuous at z and that the pdf of z_i is bounded away from zero on a neighborhood of z. We view this as a weak restriction that will be satisfied almost everywhere with respect to the dominating measure μ in many cases. We also note that this result allows for distributions to have a discrete component because the dominating measure μ may have atoms.

The weak nature of the local regularity condition highlights the strength of the asymptotic linearity as hypothesis. Primitive conditions for asymptotic linearity can be quite strong and complicated. For example, it is known that asymptotic linearity of estimators with a nonparametric first step generally requires some degree of smoothness in the functions being estimated, see Bickel and Ritov (1988). Our purpose here is to bypass those conditions in order to calculate the influence function, which result can then be used in all the important ways outlined in the

introduction, including as a starting point for formulating regularity conditions for asymptotic linearity.

It is interesting to note that the scores for the parametric families $(1-\tau)F_0 + \tau G_z^j$ with G_z^j as given in equation (3.7) all satisfy $S(z) \geq -1$. Thus to calculate the influence function we do not require a family of parametric models where the set of scores can approximate any random variable with zero mean and finite variance, as is required in Newey (1994). Also, apparently this restriction means that Theorem 1 is not a special case of the results of Van der Vaart (1998), where it is assumed that the set of scores is a cone. The proof of Theorem 1 does use some of Van der Vaart's (1991) reasoning to show differentiability of $\beta(F_\tau)$ but otherwise is very straightforward.

We want to emphasize that the purpose of Theorem 1 is quite different than Van der Vaart (1991, 1998) and other important contributions to the semiparametric efficiency literature. Here $\beta(F)$ is not a parameter of interest for some semiparametric model. Instead $\beta(F)$ is associated with an *estimator* $\hat{\beta}$, being the limit of that estimator when F is a distribution that is unrestricted except for regularity conditions, as formulated in Newey (1994). Our goal is to use $\beta(F)$ to calculate the influence function of $\hat{\beta}$ under the assumption that $\hat{\beta}$ is asymptotically linear. The purpose of Theorem 1 is to justify this calculation via equation (*3.5*). In contrast, the goal of the semiparametric efficiency literature is to find the efficient influence function for a parameter of interest when F belongs to a family of distributions.

4 Nonparametric Instrumental Variables

In this Section we derive the influence function for a semiparametric GMM estimator where the first step γ_0 is the nonparametric two stage least squares estimator (NP2SLS) of Newey and Powell (1989, 2003) and Newey (1991), abbreivated NP henceforth. As with consumer surplus, the form of the influence function emerges from the calculation of the derivative. Also, the limit of the NP2SLS estimator exists and is unique under deviations similar to those specified above as is essential for calculation of the influence function. The uniqueness and existence result is made possible by the specification $F_\tau = (1-\tau)F_0 + \tau G_z^j$ of deviations from the true distribution. In this way the approach of this paper provides the key intermediate step of existence and uniqueness of the limit of the NP2SLS estimator.

The first step will be based on a linear, nonparametric, instrumental variables model in NP where

$$y_i = \gamma_0(w_i) + \varepsilon_i, E[\varepsilon_i | x_i] = 0, \tag{4.12}$$

where w_i are right hand side variables that may be correlated with the disturbance ε_i and x_i are

instrumental variables. We begin with the case where w_i and x_i have the same dimension and are continuously distributed with rectangular supports. The identification condition for $\gamma_0(w_i)$ in this model is completeness of the conditional expectation given x_i, meaning $E[\Delta(w_i)|x_i] = 0$ implies $\Delta(w_i) = 0$. Under our conditions on w_i and x_i completeness holds generically, as shown by Andrews (2011) and Chen, Chernozhukov, Lee, and Newey (2014). Genericity justifies our assumption of completness, although as with other important generic conditions (e.g. existence of moments), completeness cannot be be tested (see Canay, Santos, Shaik, 2013).

The NP2SLS estimator minimizes the objective function $\hat{Q}(\gamma) = \sum_i \{y_i - \hat{E}[\gamma(\cdot)|x_i]\}^2/n$ over $\gamma \in \Gamma_n$ where $\hat{E}[\cdot|x_i]$ is a conditional expectation estimator and Γ_n imposes restrictions on the function, such as it being a linear combination of known functions. We first consider the case where Γ_n leaves γ unrestricted in the limit, except for having finite second moment. For fixed γ the limit of the objective function will be $Q(\gamma, F) = E_F[\{y_i - E_F[\gamma(w_i)|x_i]\}^2]$ where F denotes the distribution of a single observation. The objective function will also be the limit of other regularized objective functions such as Darolles, Fan, Florens, and Renault (2011), so we expect that the corresponding estimators converge to the same object. As usual for an extremum estimator the limit of the minimizer will be the minimizer of the limit under appropriate regularity conditions. Therefore the limit $\gamma(F)$ of the NP2SLS estimator will be

$$\gamma(F) = \arg\min_{\gamma} Q(\gamma, F) = \arg\min_{\gamma} E_F[\{y_i - E_F[\gamma(w_i)|x_i]\}^2].$$

A problem with this calculation is that $\gamma(F)$ need not exist nor be unique. This problem occurs because γ appears inside a conditional expectation. Our framework helps. Under our conditions $\gamma(F_\tau)$ does exist and is unique for $F_\tau = (1-\tau)F_0 + \tau G_z^j$ when τ is small enough. The use of deviations from the truth of the form we have considered allows us to specify G_z^j in such a way that $\gamma(F_\tau)$ exists and is unique.

We modify slightly our choice of $\delta(z)$ in order to allow for weak conditions on the marginal pdf $f_0(\tilde{x})$ of x_i as outlined below. Let $g_y(y)$, $g_w(w)$, and $g_x(x)$ be as specified in equation (3.9) except that for $g_x(x)$ (and $g_y(y), g_w(w)$) the product is only taken over components of x (y, w). For example we let $\bar{\mu}_{x\ell}^j = j^{-1} \int K((x_\ell - \tilde{x}_\ell)/j) d\tilde{\mu}_\ell$,

$$g_x(\tilde{x}) = \Pi_{\ell=1}^r \kappa_\ell^j(\tilde{x}_\ell), \kappa_\ell^j(\tilde{x}_\ell) = \left(j\bar{\mu}_\ell^j\right)^{-1} K((x_\ell - \tilde{x}_\ell)/j). \tag{4.13}$$

We then choose $\delta(z)$ to be

$$\delta(z) = g_y(y) g_w(w) g_x(x) \left[f_0(y, w|x) \int f_0(\tilde{x}) g_x(\tilde{x}) d\mu_{\tilde{x}} \right]^{-1} 1(f(y, w|x) \geq 1/j).$$

With this $\delta(z)$ we can make a key assumption that we use to show existence and uniqueness of $\gamma(F_\tau)$.

ASSUMPTION 1: a) w_i and x_i *have the same dimension, b)* $E[\Delta(w_i)|x_i]$ *is complete, and c) for each* j *there is* $\Delta^j(w_i)$ *such that* $E\left[\Delta^j(w_i)|x_i\right] = g_x(x_i)/\int g_x(\tilde{x})f_0(\tilde{x})d\tilde{x}$.

Recall that $K(u)$ is continuously differentiable of all orders with bounded support so that $g_x(\tilde{x})$ is also continuously differentiable with bounded derivatives of all orders. Assumption 1 c) is then satisfied if $E\left[\Delta(w_i)|x_i\right]$ is a compact operator with singular values that do not decline too fast and other technical conditions hold, as discussed in the Appendix. If we had used the $\delta(z)$ from earlier we would need to replace Assumption 1 c) with existence of $\Delta^j(w_i)$ with $E\left[\Delta^j(w_i)|x_i\right] = g_x(x_i)/f_x(x_i)$, which would only hold if $f_x(\tilde{x})$ were very smooth near x. The choice of $\delta(z)$ allows us to avoid these smoothness conditions for $f_0(\tilde{x})$.

The following result shows that $\gamma(F_\tau)$ exists for small enough τ and gives a useful formula

LEMMA 2: *If Assumption 1 is satisfied then for each* j *and* τ *small enough,* $\gamma(F_\tau) = \arg\min_\gamma Q(\gamma, F_\tau)$ *exists and is unique and there is* $c(\tau)$ *with* $c(0) = 0$ *and*

$$\gamma(w_i, F_\tau) = \gamma_0(w_i) + c_j(\tau)\Delta^j(w_i), \frac{\partial c_j(\tau)}{\partial \tau} = \int [\tilde{y} - \gamma_0(\tilde{w})]g_y(\tilde{y})g_w(\tilde{w})d\mu_{\tilde{y}}d\mu_{\tilde{w}}$$

With this result in place we can derive the influence function for a variety of different estimators with NP2SLS first step. We begin with a plug in estimator of the form

$$\hat{\beta} = \sum_{i=1}^{n} v(w_i)\hat{\gamma}(w_i)/n, \tag{4.14}$$

where $v(w)$ is a known function. This $\hat{\beta}$ is an estimator of $\beta_0 = E[v(w_i)\gamma_0(w_i)]$. The limit of $\hat{\beta}$ will be

$$\beta(F) = E_F[v(w_i)\gamma(w_i, F)].$$

As shown by Severini and Tripathi (2012), a necessary condition for root-n consistent estimability here is that β_0 there exists $\alpha(x_i)$ such that

$$v(w_i) = E[\alpha(x_i)|w_i]. \tag{4.15}$$

We will assume that $\alpha(x)$ is unique, which is equivalent to completeness of $E[a(x_i)|w_i]$, as holds generically like discussed above.

To calculate the influence $\hat{\beta}$ note that by equation (4.15) and Assumption 1,

$$\begin{aligned} E[v(w_i)\Delta^j(w_i)] &= E[E[\alpha(x_i)|w_i]\Delta^j(w_i)] = E[\alpha(x_i)\Delta^j(w_i)] = E[\alpha(x_i)E[\Delta^j(w_i)|x_i]] \\ &= E[\alpha(x_i)\tilde{g}_x(x_i)], \tilde{g}_x(x_i) = g_x(x_i) / \int g_x(\tilde{x}) f_0(\tilde{x})d\tilde{x}. \end{aligned}$$

By the chain rule and Lemma 2

$$\begin{aligned}\frac{\partial\beta(F_\tau)}{\partial\tau} &= \frac{\partial E_{F_\tau}[v(w_i)\gamma(w_i,F_\tau)]}{\partial\tau} \\ &= E[v(w_i)\gamma_0(w_i)S(z_i)] + \frac{\partial c_j(\tau)}{\partial\tau}E[v(w_i)\Delta^j(w_i)] \\ &= E[\{v(w_i)\gamma_0(w_i)-\beta_0\}\delta(z_i)] \\ &\quad + \int[\tilde{y}-\gamma_0(\tilde{w})]g_y(\tilde{y})g_w(\tilde{w})d\mu_{\tilde{y}}d\mu_{\tilde{w}}E[\alpha(x_i)\tilde{g}_x(x_i)] \\ &= E[\psi(z_i)\delta(z_i)],\ \psi(z)=v(w)\gamma_0(w)-\beta_0+\alpha(x)[y-\gamma_0(w)].\end{aligned} \tag{4.16}$$

As $j\longrightarrow\infty$ we will have $\int[\tilde{y}-\gamma_0(\tilde{w})]g_y(\tilde{y})g_w(\tilde{w})d\mu_{\tilde{y}}d\mu_{\tilde{w}}\longrightarrow y-\gamma_0(w)$ for $\gamma_0(\tilde{w})$ continuous at w and $E[\alpha(x_i)\tilde{g}_x(x_i)]\longrightarrow\alpha(x)f_0(x)/f_0(x)=\alpha(x)$ for $\alpha(\tilde{x})$ and $f_0(\tilde{x})$ continuous at x, so we have

THEOREM 3: *If Assumption 1 is satisfied, there exists a unique solution $\alpha(x_i)$ to $v(w_i)=E[\alpha(x_i)|w_i]$, each of $\tilde{\alpha}(x)$, $f_0(\tilde{x}),v(\tilde{w})$, and $\gamma_0(\tilde{w})$ are continuous at (w,x), the conditional pdf $f(\tilde{y},\tilde{w}|\tilde{x})$ is bounded away from zero in a neighborhood of (y,w,x), and $f_0(x)>0$,then for NP2SLS*

$$\lim_{j\longrightarrow\infty}\frac{\partial\beta(F_\tau)}{\partial\tau}=\psi(z)=v(w)\gamma_0(w)-\beta_0+\alpha(x)[y-\gamma_0(w)].$$

Here we find that the influence function of $\hat{\beta}$ of equation (4.14) is $\psi(z)$ of Theorem 3. Like the consumer surplus example a nonparametric residual $y-\gamma_0(w)$ residual emerges in the calculation of $\psi(z)$. Unlike the surplus example the residual is from the structural equation (4.12) rather than a nonparametric regression. The function $\alpha(x)$ of the instrumental variables is also a key component of the influence function. Here $\alpha(x)$ is defined implicitly rather than having an explicit form. This implicit form seems inherent to the NP2SLS first step, with existence of $\alpha(x)$ solving equation (4.15) being required for root-n consistency of $\hat{\beta}$.

Note that $\alpha(x)$ is the solution of a "reverse" structural equation involving an expectation conditional on the endogenous variable w_i rather than the instrument x_i. An analogous "reverse" structural equation also appears in a linear instrumental variables (IV) setting. Let $\hat{d}=(\sum_{i=1}^n X_iW_i^T)^{-1}\sum_{i=1}^n X_iy_i$ be the linear IV estimator having limit $d_0=(E[X_iW_i^T])^{-1}E[X_iy_i]$. A linear IV analog of the structural function $\gamma_0(w)$ is w^Td_0 and of parameter β_0 is

$$b_0=E[v(W_i)(w_i^Td_0)].$$

A corresponding estimator of b_0 is $\hat{b}=\sum_{i=1}^n v(W_i)W_i^T\hat{d}/n$. It is straightforward to show that the influence function of $\hat{b}$ is

$$v(w)(w^Td_0)-b_0+a(x)[y-w^Td_0],a(x)=x^T\left(E[w_ix_i^T]\right)^{-1}E[w_iv(w_i)].$$

Here $a(x)$ is the function obtained from "reverse" IV where x is the right hand side variable and w the instrumental variable. The functional $\alpha(x)$ is a nonparametric analog of $a(x)$ where linear IV is replaced by the solution to a conditional expectation equation.

A solution $\alpha(x)$ to equation (4.15) will only exist when $v(w)$ satisfies certain conditions. When $E[v(w_i)^2] < \infty$ a function $\alpha(x)$ can only exist when $v(w_i)$ has Fourier coefficients, with respect to the singular value basis corresponding to the (assumed to be compact) operator $E[\cdot|w_i]$, that decline fast enough relative to the inverse of the singular values. This condition requires some "smoothness" of $v(w_i)$ and will rule out some functions, such as $v(w)$ that that have jumps (e.g. indicator functions).

The influence function of Theorem 3 is consistent with the semiparametric efficiency bound given in Severini and Tripathi (2012). WHAT ABOUT AI AND CHEN 2007.

There is a different way of estimating β_0 that is analogous to Santos (2014). By equation (4.15), the conditional moment restriction (4.12), and iterated expectations,

$$\beta_0 = E[E[\alpha(x_i)|w_i]\gamma_0(w_i)] = E[\alpha(x_i)\gamma_0(w_i)] = E[\alpha(x_i)y_i]. \tag{4.17}$$

Based on the last equality an estimator for β_0 could be constructed as $\tilde{\beta} = \sum_{i=1}^n \hat{\alpha}(x_i)y_i/n$ where $\hat{\alpha}$ is an estimator of the solution of equation (4.15). The influence function for this estimator is the same as in Theorem 3. This equality of influence functions occurs because equation (4.17) is satisfied for any F_τ where equation (4.15) holds and equation (4.15) will hold for the F_τ we are considering using arguments like those of Lemma 2. Thus $\tilde{\beta}$ will have the same limit as $\hat{\beta}$ for a general distribution and thus the same influence function.

This influence function calculation can be extended beyond the estimator of (4.14) to other semiparametric GMM estimators. This extension requires a corresponding extension of equation (4.15). The following condition provides such an extension:

ASSUMPTION 2: *There exists $\alpha(x_i)$ with $E[\alpha(x_i)^2] < \infty$ such that for all F_τ*

$$\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau} = E[\alpha(x_i)\frac{\partial E[\gamma(w_i, F_\tau)|x_i]}{\partial \tau}] \tag{4.18}$$

For $m(z, \beta, \gamma) = v(w)\gamma(w) - \beta$ Assumption 2 is equivalent to equation (4.15). If equation (4.15) is satisfied then

$$\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau} = E[\alpha(x_i)\frac{\partial \gamma(w_i, \tau)}{\partial \tau}] = E[\alpha(x_i)\frac{\partial E[\gamma(w_i, F_\tau)|x_i]}{\partial \tau}],$$

where the last equality follows by iterated expectations and interchanging the order of differentiation and integration. Also, if Assumption 2 is satisfied for all $\Delta_\tau(w_i) = \partial\gamma(w_i, F_\tau)/\partial\tau$ we have

$$E[v(w_i)\Delta_\tau(w_i)] = E[\alpha(x_i)\Delta_\tau(w_i)] = E[E[\alpha(x_i)|w_i]\Delta_\tau(w_i)],$$

by iterated expectations. The only way this equation can hold over all $\Delta_\tau(w_i)$ is if equation (4.15) is satisfied.

The existence of $\alpha(x_i)$ satisfying Assumption 2 will follow from other conditions. If the $\partial E[m(z_i, \beta_0, \gamma(F_\tau))]/\partial\tau$ is a continuous linear functional of $\partial E[\gamma(w_i, F_\tau)|x_i]/\partial\tau$ which is contained in a closed linear set L and that functional can be extended to be continuous on all of L then existence of $\alpha(x)$ satisfying Assumtpion 2 follows by the Riesz represenation theorem. Similar uses of the Riesz representation theorem are given in Newey (1994), Ai and Chen (2007), and Ackerberg, Chen, Hahn, and Liao (2014).

We can use Assumption 2 and Lemma 2 to calculate the influence function of a semiparametric GMM estimator with a NP2SLS first step. Recall from the discussion of equation (2.3) that the influence function of semiparametric GMM is determined by the correction term $\phi(z)$ for the first step and that $\phi(z)$is the influence function of $E[m(z_i, \beta, \gamma(F))]$. When Assumption 2 is satisfied it follows exactly as in equation (4.16) that

$$\begin{aligned}\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial\tau} &= E[\alpha(x_i)\frac{\partial c_j(\tau)}{\partial\tau}E[\Delta^j(w_i)|x_i]] \\ &= E[\phi(z_i)\delta(z_i)], \phi(z) = \alpha(x)[y - \gamma_0(w)],\end{aligned}$$

giving the following result:

THEOREM 4: *If Assumptions 1 and 2 are satisfied then for the NP2SLS first step*

$$\lim_{j\longrightarrow\infty}\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial\tau} = \phi(z) = \alpha(x)[y - \gamma_0(w)].$$

An interesting example of this result is the average derivative estimator of Ai and Chen (2007), where $m(z, \beta, \gamma) = \bar{v}(w)\partial\gamma(w)/\partial w - \beta$ for some known $\bar{v}(w)$. Let $v(w) = -f_0(w)^{-1}\partial[\bar{v}(w)f_0(w)]/\partial w$. Assume that equation (4.15) is satisfied for this $v(w)$, so that there exist $\alpha(x_i)$ with

$$-f_0(w_i)^{-1}\partial[\bar{v}(w_i)f_0(w_i)]/\partial w = E[\alpha(x_i)|w_i]. \tag{4.19}$$

Then integration by parts and iterated expectations gives

$$\begin{aligned}\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial\tau} &= \frac{\partial E[\bar{v}(w_i)\partial\gamma(w_i, F_\tau)/\partial w]}{\partial\tau} = \frac{\partial E[v(w_i)\gamma(w_i, F_\tau)]}{\partial\tau} \\ &= \frac{\partial E[\alpha(x_i)E[\gamma(w_i, F_\tau)|x_i]]}{\partial\tau} = E[\alpha(x_i)\frac{\partial E[\gamma(w_i, F_\tau)|x_i]}{\partial\tau}],\end{aligned}$$

so Assumption 2 is satisfied. It then follows by $m(z, \beta, \gamma) = \bar{v}(w)\partial\gamma(w)/\partial w - \beta$ and Theorem 4 that the influence function for the weighted average derivative is

$$\psi(z) = \bar{v}(w)\frac{\partial\gamma_0(w)}{\partial w} - \beta_0 + \alpha(x)[y - \gamma_0(w)].$$

Comparison with Ai and Chen ??.

ESTIMATOR; NP2SLS estimator of ??.

We can also calculate the influence function for NP2SLS for a nonlinear, possibly misspecified residual $\rho(z,\gamma)$ where the NP2SLS estimatori is based on orthogonality of $\rho(z,\gamma_0)$ with a set $\mathcal{A}$ of functions of instrumental variables x. We will assume that there is a fixed countable basis $(a_1(x), a_2(x), ...)$ that spans $\mathcal{A}$ in mean square for each F_τ for τ small enough. NP2SLS is here based on the orthogonality condition

$$E[a_j(x_i)\rho(z_i,\gamma_0)] = 0 \text{ for all } j.$$

For example, if $\mathcal{A}$ is all functions of x_t with finite second moment this restriction is equivalent to $E[\rho(z_i,\gamma_0)|x_i] = 0$. Here $a_j(x)$ could be power series in a bounded one-to-one function of x or could be regression splines. More generally $a_j(x)$ could be functions of only a subvector of x or could be restricted to be additive in subvectors of x.

The NP2SLS estimator is like that given in NP. It minimizes over $\gamma \in \Gamma_n$ the sample second moment of the predicted values from the ordinary least squares regression of the residual $\rho(z_i,\gamma)$ on $p^K(x_i) = (p_{1K}(x_i), ..., p_{KK}(x_i))^T$, where each function $p_{kK}(x)$ is one of the basis functions $a_j(x)$ for some j. We assume that as $K \longrightarrow \infty$ any element of $\mathcal{A}$ may be approximated in mean square by a linear combination of $p^K(x_i)$. Let $b(z_i)$ denote any random variable with finite variance and $\pi_\tau(\rho(\gamma), x_i)$ and $\pi_\tau(b, x_i)$ denote the population orthogonal projection of $\rho(z_i,\gamma)$ and $b(z_i)$ on $\mathcal{A}$ when the true distribution is F_τ. Also let $E_\tau[\cdot]$ denote the expectation under F_τ. The limit of the 2SLS objective function will be

$$Q_\tau(\gamma) = E_\tau[\pi_\tau(\rho(\gamma), x_i)^2].$$

We will assume that $\gamma(w)$ is restricted to belong to a linear set Γ and that $\Gamma_n \subset \Gamma$ for all n. For example, $\gamma(w)$ might be restricted to be additive in subvectors of w. It then follows as usual for extemum estimators that the limit of the NP2SLS estimator will be

$$\gamma_\tau = \arg\min_{\Gamma} Q_\tau(\gamma).$$

Here we will just assume that γ_τ exists and is unique. We do this, rather than prove existence and uniqueness, because it is difficult to show uniqueness of a minimum when $\rho(z,\gamma)$ is nonlinear in γ and when there are more instrumental variables than endogenous variables. Given the previous results of this Section and Chen, Chernozhukov, Lee, and Newey (2014) we conjecture that conditions for local existence and uniquenss could be formulated, but we leave that formulation to future work.

To find the form of the adjustment term in this setting we need an extended version of Assumption 2. The following condition provides that extension.

ASSUMPTION 3: *There exists $\alpha(x_i) \in \mathcal{A}$ with $E[\alpha(x_i)^2] < \infty$ such that for all F_τ*

$$\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau} = -E[\alpha(x_i)\frac{\partial \pi_0(\rho(\gamma_\tau), x_i)}{\partial \tau}] \tag{4.20}$$

To derive the adjustment term for NP2SLS it is helpful to consider first order conditions for γ_τ. Let ζ denote a scalar and $\Delta(w)$ some function of the endogenous variables w such that $\Delta(w_i) \in \Gamma$. Then we have $\gamma_\tau(w_i) + \zeta\Delta(w_i) \in \Gamma$ for any ζ by Γ linear. We will impose the following condition:

ASSUMPTION 4: *For all τ small enough $\pi_\tau(\rho(\gamma_\tau + \zeta\Delta), x_i)$ is differentiable in at $\zeta = 0$ and there is $d_\tau(z_i)$ such that*

$$\left.\frac{\partial \pi_\tau(\rho(\gamma_\tau + \zeta\Delta), x_i)}{\partial \zeta}\right|_{\zeta=0} = \pi_\tau(d_\tau\Delta, x_i).$$

To illustrate consider the endogenous quantile model of Chernozhukov and Hansen (2004) and Chernozhukov, Imbens, and Newey (2007) where $\rho(z, \gamma) = 1(y < \gamma(w)) - \eta$ for a scalar η with $0 < \eta < 1$. Suppose that for τ small enough the distribution of y_i conditional on x_i and w_i is continuous in a neighborhood of $\gamma_\tau(w_i)$ with conditional pdf $f_\tau(y|w, x)$. Let derivatives with respect to ζ be evaluated at $\zeta = 0$. Then

$$\begin{aligned} \frac{\partial E_\tau[\rho(z_i, \gamma_\tau + \zeta\Delta)|w_i, x_i]}{\partial \zeta} &= -f_\tau(\gamma_\tau(w_i)|w_i, x_i)\Delta(w_i) = -d_\tau(w_i, x_i)\Delta(w_i), \\ d_\tau(w_i, x_i) &= f_\tau(\gamma_\tau(w_i)|w_i, x_i). \end{aligned} \tag{4.21}$$

Assuming that the order of differentiation and projection can be interchanged, it follows by iterated projections that

$$\frac{\partial \pi_\tau(\rho(\gamma_\tau + \zeta\Delta), x_i)}{\partial \zeta} = \pi_\tau\left(\frac{\partial E[\rho(z_i, \gamma_\tau + \zeta\Delta)|w_i, x_i]}{\partial \zeta}, x_i\right) = \pi_\tau(d_\tau\Delta, x_i),$$

so that Assumption 3 is satisfied. More generally Assumption 3 will hold if $\rho(z, \gamma) = \rho(z, \gamma(w))$ with $d_\tau(w, x) = \partial E[\rho(z_i, \gamma_\tau(w) + \zeta)|w_i, x_i]/\partial\zeta$.

By calculus of variations and Assumption 3 the first order conditions for γ_τ are that for any $\Delta(w_i) \in \Gamma$,

$$0 = E_\tau[\pi_\tau(\rho(\gamma_\tau), x_i)\frac{\partial \pi_\tau(\rho(\gamma_\tau + \zeta\Delta), x_i)}{\partial \zeta}] = E_\tau[\pi_\tau(\rho(\gamma_\tau), x_i)\pi_\tau(d_\tau\Delta, x_i)].$$

This is an identity in τ. Differentiating this identity in τ at $\tau = 0$ and applying the chain rule gives

$$\begin{aligned} 0 &= E[\pi_0(d_0\Delta, x_i)\frac{\partial \pi_0(\rho(\gamma_\tau), x_i)}{\partial \tau}] + E[\pi_0(\rho(\gamma_0), x_i)\pi_0(d_0\Delta, x_i)S(z_i)] \\ &+E[\pi_0(d_0\Delta, x_i)\frac{\partial \pi_\tau(\rho(\gamma_0), x_i)}{\partial \tau}] + E[\pi_0(\rho(\gamma_0), x_i)\frac{\partial \pi_\tau(d_0\Delta, x_i)}{\partial \tau}] \\ &+E[\pi_0(\rho(\gamma_0), x_i)\frac{\partial \pi_0(d_\tau\Delta, x_i)}{\partial \tau}] \end{aligned} \tag{4.22}$$

The following result helps us evaluate the third and fourth terms in this equation.

LEMMA 5: For any $a(x_i) \in \mathcal{A}$ and $b(z_i)$ with finite variance

$$\frac{\partial E[a(x_i)\pi_\tau(b|x_i)]}{\partial \tau} = E[a(x_i)\{b(z_i) - \pi_0(b|x_i)\}S(z_i)].$$

Proof of Lemma 5: Note that for each j the definition of the projection implies that

$$E_\tau[a_j(x_i)b(z_i)] = E_\tau[a_j(x_i)\pi_\tau(b|x_i)]$$

identically in τ. Differentiating both sides at $\tau = 0$ and applying the chain rule and ?? gives

$$E[a_j(x_i)b(z_i)S(z_i)] = E[a_j(x_i)\pi_0(b|x_i)S(z_i)] + \frac{\partial E[a_j(x_i)\pi_\tau(b|x_i)]}{\partial \tau}.$$

Solving it follows that for each j,

$$\frac{\partial E[a(x_i)\pi_\tau(b|x_i)]}{\partial \tau} = E[a_j(x_i)\{b(z_i) - \pi_0(b|x_i)\}S(z_i)].$$

Consider $\lambda = (\lambda_1, ..., \lambda_J)^T$ such that $(a_1(x_i), ..., a_J(x_i))\lambda \longrightarrow a(x_i)$ in mean square. The conclusion then follows by $S(z)$ bounded. Q.E.D..

We can apply Lemma 5 to the third and fourth terms of equation (4.22) and solve for the first term to obtain

$$E[\pi_0(d_0\Delta, x_i)\frac{\partial \pi_0(\rho(\gamma_\tau), x_i)}{\partial \tau}] = -E[\phi_\Delta(z_i)S(z_i)] - -E[\pi_0(\rho(\gamma_0), x_i)\frac{\partial \pi_0(d_\tau\Delta, x_i)}{\partial \tau}],$$

$$\begin{aligned} \phi_\Delta(z) &= \pi_0(\rho(\gamma_0), x)[d_0(w, x)\Delta(w) - \pi_0(d\Delta, x)] + \pi_0(d\Delta, x)[\rho(z, \gamma_0) - \pi_0(\rho(\gamma_0), x)] \\ &+\pi_0(\rho(\gamma_0), x)\pi_0(d\Delta, x) - E[\pi_0(\rho(\gamma_0), x_i)\pi_0(d\Delta, x_i)]. \end{aligned}$$

This result can be combined with Assumption 3 to obtain the adjustment term when the first step has is the NP2SLS estimator. We state this result as a Proposition, similarly to Newey

(1994), because its derivation uses formal caclulations without specifying a sufficient set of regularity conditions.

PROPOSITION 6: *If the model is correctly specified, so* $\pi_0(\rho(\gamma_0), x_i) = 0$*, and there is a sequence* $\Delta_j(w)$ *such that* $\pi_0(d_0\Delta_j, x_i) \longrightarrow \alpha(x_i)$ *in mean square then the adjustment term is*

$$\phi(z) = \alpha(x)\rho(z, \gamma_0).$$

If the model is misspecified with $\pi_0(\rho(\gamma_0), x_i) \neq 0$*,* $\partial\pi_0(d_\tau\Delta, x_i)/\partial\tau = 0$*, and there exists* $\Delta(w)$ *such that* $\alpha(x_i) = \pi_0(d\Delta, x_i)$ *then the adjustment term is* $\phi(z) = \phi_\Delta(z)$*.*

Note here that the result with misspecification assumes that $\partial\pi_0(d_\tau\Delta, x_i)/\partial\tau = 0$. We do not know if an influence function exists when this condition does not hold. The problem is that d_τ may be a nonparametric object evaluated at a point and hence the derivative of the projection of $d_\tau\Delta$ on $\mathcal{A}$ may not have a representation as an expected product with the score. For example, for quantile IV $d_\tau(w_i, x_i) = f_\tau(\gamma_\tau(w_i)|w_i, x_i)$ which is nonparametric conditional density evaluated at a point. In such cases it may be the case that the influence function does not exist.

The existence of $\Delta(w)$ with $\alpha(x_i) = \pi_0(d\Delta, x_i)$ in the misspecified case is restrictive. This condition requires $\alpha(x)$ be smooth in a way similar to $v(w)$ being smooth because of $v(w_i) = E[\alpha(x_i)|w_i]$. Because these conditions are similar that the one for $v(w)$ is necessary for root-n consistent estimability under correct specification it may be that existence of $\Delta(w)$ with $\alpha(x_i) = \pi_0(d\Delta, x_i)$ is necessary for root-n consistent estimability under misspecification. The condition under correct specification, that there is a sequence $\Delta_j(w)$ such that $\pi_0(d_0\Delta_j, x_i) \longrightarrow \alpha(x_i)$, is a much weaker condition. For example, for a model where d_0 is constant, $\mathcal{A}$ is the set of all functions of x_i with finite variance, and the dimension of w_i is equal to the dimension of z_i this condition automatically holds.

To illustrate we can apply Proposition 6 to obtain the adjustment term when the first step is quantile IV, so that $\rho(z, \gamma) = 1(y < \gamma(w)) - \tau$, and the model is correctly specified. Similar to the above discussion of equation (4.21) assume that

$$\frac{\partial\pi_0(\rho(\gamma_\tau), x_i)}{\partial\tau} = -\pi_0(d_0\frac{\partial\gamma_\tau(\cdot)}{\partial\tau}, x_i).$$

Suppose also that there is some $v(w_i)$ such that

$$\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial\tau} = E[v(w_i)\frac{\partial\gamma_\tau(w_i)}{\partial\tau}].$$

Then by $\alpha(x_i) \in \mathcal{A}$ and $E[\alpha(x_i)\pi_0(b, x_i)] = E[\alpha(x_i)b(z_i)]$ for any $b(z_i)$ with finite variance Assumption 3 becomes

$$E[v(w_i)\frac{\partial\gamma_\tau(w_i)}{\partial\tau}] = E[\alpha(x_i)d_0(w_i, x_i)\frac{\partial\gamma_\tau(w_i)}{\partial\tau}].$$

This equation will hold if

$$v(w_i) = E[\alpha(x_i)d_0(w_i, x_i)|w_i].$$

Applying Proposition 6 then gives

PROPOSITION 7: *If* $\rho(z,\gamma) = 1(y < \gamma(w)) - \eta$, $\pi_0(\rho(\gamma_0), x_i) = 0$, *and i) there there exists* $\alpha(x_i)$ *such that* $v(w_i) = E[d_0(w_i, x_i)\alpha(x_i)|w_i]$ *and ii) there exists a sequence* $\Delta_j(w)\pi_0(d_0\Delta_j, x_i) \longrightarrow \alpha(x_i)$ *in mean square then the adjustment term is*

$$\phi(z) = \alpha(x)\rho(z, \gamma_0).$$

For example consider the average derivative for quantile IV where $\beta_0 = E[\bar{v}(w_i)\partial\gamma_0(w_i)/\partial w]$ for known $\bar{v}(w)$. Here condition i) of Proposition 7 is existence of $\alpha(x)$ such that

$$v(w_i) = -f_0(w_i)^{-1}\partial[\bar{v}(w_i)f_0(w_i)]/\partial w = E[\alpha(x_i)d_0(w_i, x_i)|w_i].$$

This is a weighted (by $d_0(w_i, x_i)$) modification of equation (4.19) that will only hold when the function on the left satisfies certain restrictions, similar to the above discussion. Condition ii) may place some additional restrictions on $v(w_i)$. These restrictions will be weaker the richer is the instrumental variable set $\mathcal{A}$. Consider the case where $\mathcal{A}$ is the set of all functions of x_i so that $\pi_0(d_0\Delta_j, x_i) = E[d_0(w_i, x_i)\Delta_j(w_i)|x_i]$. If the operator $E[d_0(w_i, x_i)\Delta_j(w_i)|x_i]$ is compact then condition ii) will hold by standard arguments as in ??. More generally if $\mathcal{A}$ were restricted that would impose corresponding restrictions on $\alpha(x)$, and hence on $v(w)$. When both conditions i) and ii) are satisfied the conclusion of Proposition 7 implies that the influence function of the weighted average derivative for quantile IV is

$$\psi(z) = \bar{v}(w)\frac{\partial\gamma_0(w)}{\partial w} - \beta_0 + \alpha(x)[1(y < \gamma_0(w)) - \eta].$$

5 Sufficient Conditions for Asymptotic Linearity

One of the important uses of the influence function is to help specify regularity conditions for asymptotic linearity. The idea is that an formula for $\psi(z)$ determines the remainder terms that can then be analyzed in order to formulate primitive regularity conditions. In this Section we formulate such regularity conditions using a functional expansion approach that applies quite broadly. It may be possible to formulate regularity conditions for particular estimators that are weaker than we consider.

In this section we consider estimators that are functionals of a nonparametric estimator taking the form

$$\hat{\beta} = \beta(\hat{F}),$$

where $\hat{F}$ is some nonparametric estimator of the distribution of z_i. Both the integrated squared density and the average consumer surplus estimators have this form, as discussed below. We consider a more general class of estimators in Section 7.

Since $\beta_0 = \beta(F_0)$, adding and subtracting the term $\int \psi(z)\hat{F}(dz)$ gives

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta_0) - \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} &= \sqrt{n}\hat{R}_1(\hat{F}) + \sqrt{n}R_2(\hat{F}), \qquad (5.23) \\ \hat{R}_1(F) &= \int \psi(z)F(dz) - \sum_{i=1}^{n} \psi(z_i)/n, \quad R_2(F) = \beta(F) - \beta(F_0) - \int \psi(z)F(dz). \end{aligned}$$

If $\sqrt{n}\hat{R}_1(\hat{F})$ and $\sqrt{n}R_2(\hat{F})$ both converge in probability to zero then $\hat{\beta}$ will be asymptotically linear. To the best of our knowledge little is gained in terms of clarity or relaxing conditions by considering $\hat{R}_1(F) + R_2(F)$ rather than $\hat{R}_1(F)$ and $R_2(F)$ separately, so we focus on the individual remainders.

The form of the remainders $\hat{R}_1(F)$ and $R_2(F)$ are motivated by $\psi(z)$ being a derivative of $\beta(F)$ with respect to F. The derivative interpretation of $\psi(z)$ suggests a linear approximation of the form

$$\beta(F) \approx \beta(F_0) + \int \psi(z)(F - F_0)(dz) = \beta(F_0) + \int \psi(z)F(dz),$$

where the equality follows by $E[\psi(z_i)] = 0$. Plugging in $\hat{F}$ in this approximation gives $\int \psi(z)\hat{F}(dz)$ as a linear approximation to $\hat{\beta} - \beta_0$. The term $R_2(\hat{F})$ is then the remainder from linearizing $\hat{\beta} = \beta(\hat{F})$ around F_0. The term $\hat{R}_1(\hat{F})$ is the difference between the linear approximation $\int \psi(z)F(dz)$ evaluated at the nonparametric estimator $\hat{F}$ and at the empirical distribution $\tilde{F}$, with $\int \psi(z)\tilde{F}(dz) = \sum_{i=1}^{n} \psi(z_i)/n$.

It is easy to fit the kernel estimator of the integrated squared density into this framework. We let $\hat{F}$ be the CDF corresponding to a kernel density estimator $\hat{f}(z)$. Then for $\beta(F) = \int f(z)^2 dz$, the fact that $\hat{f}^2 - f^2 = (\hat{f} - f)^2 + 2f(\hat{f} - f)$ gives an expansion as in equation (5.23) with

$$\hat{R}_1(\hat{F}) = \int \psi(z)\hat{f}(z)dz - \sum_{i=1}^{n} \psi(z_i)/n, R_2(\hat{F}) = \int [\hat{f}(z) - f_0(z)]^2 dz.$$

Applying this framework to a series regression estimator requires formulating that as an estimator of a distribution F. One way to do that is to specify a conditional expectation operator conditional on x and a marginal distribution for x, since a conditional expectation operator implies a conditional distribution. For a series estimator we can take $\hat{F}$ to have a conditional expectation operator such that

$$E_{\hat{F}}[a(q,x)|x] = \frac{1}{n}\sum_{i=1}^{n} a(q_i,x)p^K(x_i)^T\hat{\Sigma}^{-1}p^K(x).$$

Then it will be the case such that

$$\beta(\hat{F}) = \int W(x) E_{\hat{F}}[q|x] dx = \int W(x) \hat{d}(x) dx = \hat{\beta},$$

which only depends on the conditional expectation operator, leaving us free to specify any marginal distribution for x that is convenient. Taking $\hat{F}$ to have a marginal distribution which is the true distribution of the data we see that

$$\beta(\hat{F}) - \beta_0 = \int E_{\hat{F}}[W(x)\{q - d_0(x)\}|x] dx = \int E_{\hat{F}}[\psi(z)|x] f_0(x) dx = \int \psi(z) \hat{F}(dz).$$

In this case $R_2(F) = 0$ and

$$\hat{R}_1(\hat{F}) = \int E_{\hat{F}}[\psi(z)|x] f_0(x) dx - \frac{1}{n} \sum_{i=1}^{n} \psi(z_i).$$

Next we consider conditions for both of the remainder terms $\hat{R}_1(\hat{F})$ and $R_2(\hat{F})$ to be small enough so that $\hat{\beta}$ is asymptotically linear. The remainder term $\hat{R}_1(\hat{F}) = \int \psi(z)(\hat{F} - \tilde{F})(dz)$ is the difference between a linear functional of the nonparametric estimator $\hat{F}$ and the same linear functional of the empirical distribution $\tilde{F}$. It will shrink with the sample size due to $\hat{F}$ and $\tilde{F}$ being nonparametric estimators of the distribution of z_i, meaning that they both converge to F_0 as the sample size grows. This remainder will be the only one when $\beta(F)$ is a linear functional of $\hat{F}$.

This remainder often has an important expectation component that is related to the bias of $\hat{\beta}$. Often $\hat{F}$ can be thought of as a result of some smoothing operation applied to the empirical distribution. The $\hat{F}$ corresponding to a kernel density estimator is of course an example of this. An expectation of $\hat{R}_1(\hat{F})$ can then be thought of as a smoothing bias for $\hat{\beta}$, or more precisely a smoothing bias in the linear approximation term for $\hat{\beta}$. Consequently, requiring that $\sqrt{n}\hat{R}_1(\hat{F}) \xrightarrow{p} 0$ will include a requirement that $\sqrt{n}$ times this smoothing bias in $\hat{\beta}$ goes to zero.

Also $\sqrt{n}$ times the deviation of $\hat{R}_1(\hat{F})$ from an expectation will need to go zero in order for $\sqrt{n}\hat{R}_1(\hat{F}) \xrightarrow{p} 0$. Subtracting an expectation from $\sqrt{n}\hat{R}_1(F)$ will generally result in a stochastic equicontinuity remainder, which is bounded in probability for fixed F and converges to zero as F approaches the empirical distribution. In the examples the resulting remainder goes to zero under quite weak conditions.

To formulate a high level condition we will consider an expectation conditional on some sigma algebra χ_n that can depend on all of the observations. This set up gives flexibility in the specification of the stochastic equicontinuity condition.

ASSUMPTION 1: $E[\hat{R}_1(\hat{F})|\chi_n] = o_p(n^{-1/2})$ and $\hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})|\chi_n] = o_p(n^{-1/2})$.

We illustrate this condition with the examples. For the integrated square density let χ_n be a constant so that the conditional expectation in Assumption 1 is the unconditional expectation. Let $\psi(z,h) = \int \psi(z+hu)K(u)du$ and note that by a change of variables $u = (z-z_i)/h$ we have $\int \psi(z)\hat{f}(z)dz = n^{-1}h^{-r}\sum_{i=1}^n \int \psi(z)K((z-z_i)/h)dz = \sum_{i=1}^n \psi(z_i,h)/n$. Then

$$\begin{aligned} E[\hat{R}_1(\hat{F})] &= E[\psi(z_i,h)] = \int [\int \psi(z+hu)f_0(z)dz]K(u)du, \qquad (5.24) \\ \hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})] &= \frac{1}{n}\sum_{i=1}^n \{\psi(z_i,h) - E[\psi(z_i,h)] - \psi(z_i)\}. \end{aligned}$$

Here $E[\hat{R}_1(\hat{F})]$ is the kernel bias for the convolution $\rho(t) = \int \psi(z+t)f_0(z)dz$ of the influence function and the true pdf. It will be $o(n^{-1/2})$ under smoothness, kernel, and bandwidth conditions that are further discussed below. The term $\hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})]$ is evidently a stochastic equicontinuity term that is $o_p(n^{-1/2})$ as long as $\lim_{h\longrightarrow 0} E[\{\psi(z_i,h) - \psi(z_i)\}^2] = 0$.

For the series estimator for consumer surplus let $\hat{\delta}(x) = [\int W(x)p^K(x)dx]^T\hat{\Sigma}^{-1}p^K(x)$ and note that $\hat{\beta} = \sum_{i=1}^n \hat{\delta}(x_i)q_i/n$. Here we take $\chi_n = \{x_1, ..., x_n\}$. Then we have

$$\begin{aligned} E[\hat{R}_1(\hat{F})|\chi_n] &= \frac{1}{n}\sum_{i=1}^n \hat{\delta}(x_i)d_0(x_i) - \beta_0, \qquad (5.25) \\ \hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})|\chi_n] &= \frac{1}{n}\sum_{i=1}^n [\hat{\delta}(x_i) - \delta(x_i)][q_i - d_0(x_i)]. \end{aligned}$$

Here $E[\hat{R}_1(\hat{F})|\chi_n]$ is a series bias term that will be $o_p(n^{-1/2})$ under conditions discussed below. The term $\hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})|\chi_n]$ is a stochastic equicontinuity term that will be $o_p(n^{-1/2})$ as $\hat{\delta}(x)$ gets close to $\delta(x)$. In particular, since $\hat{\delta}(x)$ depends only on $x_1, ..., x_n$, the expected square of this term conditional on χ_n will be $n^{-2}\sum_{i=1}^n [\hat{\delta}(x_i) - \delta(x_i)]^2 Var(q_i|x_i)$, which is $o_p(n^{-1})$ when $Var(q_i|x_i)$ is bounded and $n^{-1}\sum_{i=1}^n [\hat{\delta}(x_i) - \delta(x_i)]^2 = o_p(1)$.

Turning now to the other remainder $R_2(F)$, we note that this remainder results from linearizing around F_0. The size of this remainder is related to the smoothness properties of $\beta(F)$. We previously used Gateaux differentiability of $\beta(F)$ along certain directions to calculate the influence function. We need a stronger smoothness condition to make the remainder $R_2(\hat{F})$ small. Frechet differentiability is one helpful condition. If the functional $\beta(F)$ is Frechet differentiable at F_0 then we will have

$$R_2(F) = o(\|F - F_0\|),$$

for some norm $\|\cdot\|$. Unfortunately Frechet differentiability is generally not enough for $R_2(\hat{F}) = o_p(n^{-1/2})$. This problem occurs because $\beta(F)$ and hence $\|F - F_0\|$ may depend on features of F which cannot be estimated at a rate of $1/\sqrt{n}$. For the integrated squared error $\|F - F_0\| =$

$\left\{\int [f(z) - f_0(z)]^2 dz\right\}^{1/2}$ is the root integrated squared error. Consequently $\sqrt{n}\left\|\hat{F} - F_0\right\|$ is not bounded in probability and so $\sqrt{n}R_2(\hat{F})$ does not converge in probability to zero.

This problem can be addressed by specifying that $\left\|\hat{F} - F_0\right\|$ converges at some rate and that $\beta(F)$ satisfies a stronger condition than Frechet differentiability. One condition that is commonly used is that $R_2(F) = O(\|F - F_0\|^2)$. This condition will be satisfied if $\beta(F)$ is twice continuously differentiable at F_0 or if the first Frechet derivative is Lipschitz. If it is also assumed that $\hat{F}$ converges faster than $n^{-1/4}$ then Assumption A1 will be satisfied. A more general condition that allows for larger $R_2(F)$ is given in the following hypothesis.

ASSUMPTION 2: For some $1 < \zeta \leq 2$, $R_2(F) = O(\|F - F_0\|^\zeta)$ *and* $\left\|\hat{F} - F_0\right\| = o_p(n^{-1/2\zeta})$.

This condition separates nicely into two parts, one about the properties of the functional and another about a convergence rate for $\hat{F}$. For the case $\zeta = 2$ Assumption 2 has been previously been used to prove asymptotic linearity, e.g. by Ait-Sahalia (1991), Andrews (1994), Newey (1994), Newey and McFadden (1994), Chen and Shen (1997), Chen, Linton, and Keilegom (2003), and Ichimura and Lee (2010) among others.

In the example of the integrated squared density $R_2(F) = \int [f(z) - f_0(z)]^2 dz = O(\|F - F_0\|^2)$ for $\|F - F_0\| = \{\int [f(z) - f_0(z)]^2 dz\}^{1/2}$. Thus Assumption 2 will be satisfied with $\zeta = 2$ when $\hat{f}$ converges to f_0 faster than $n^{-1/4}$ in the integrated squared error norm.

The following result formalizes the observation that Assumption 1 and 2 are sufficient for asymptotic linearity of $\hat{\beta}$.

THEOREM 2: *If Assumptions 1 and 2 are satisfied then $\hat{\beta}$ is asymptotically linear with influence function $\psi(z)$.*

An alternative set of conditions for asymptotic normality of $\sqrt{n}(\hat{\beta} - \beta_0)$ was given by Ait-Sahalia (1991). Instead of using Assumption 1 Ait-Sahalia used the condition that $\sqrt{n}(\hat{F} - F_0)$ converged weakly as a stochastic process to the same limit as the empirical process. Asymptotic normality of $\sqrt{n}\int \psi(z)\hat{F}(dz)$ then follows immediately by the functional delta method. This approach is a more direct way to obtain asymptotic normality of the linear term in the expansion. However weak convergence of $\sqrt{n}(\hat{F} - F_0)$ requires stronger conditions on the nonparametric bias than does the approach adopted here. Also, Ait-Sahalia's (1991) approach does not deliver asymptotic linearity, though it does give asymptotic normality.

These conditions for asymptotic linearity of semiparametric estimators are more complicated than the functional delta method outlined in Reeds (1976), Gill (1989), and Van der Vaart and Wellner (1996). The functional delta method gives asymptotic normality of a functional of the empirical distribution or other root-n consistent distribution estimator under just two

conditions, Hadamard differentiability of the functional and weak convergence of the empirical process. That approach is based on a nice separation of conditions into smoothness conditions on the functional and statistical conditions on the estimated distribution. It does not appear to be possible to have such simple conditions for semiparametric estimators. One reason is that they are only differentiable in norms where $\sqrt{n}\left\|\hat{F}-F_0\right\|$ is not bounded in probability. In addition the smoothing inherent in $\hat{F}$ introduces a bias that depends on the functional and so the weakest conditions are only attainable by accounting for interactions between the functional and the form of $\hat{F}$. In the next Section we discuss this bias issue.

6 Linear Functionals

In this Section we consider primitive conditions for Assumption 1 to be satisfied for kernel density and series estimators. We focus on Assumption 1 because it is substantially more complicated than Assumption 2. Assumption 2 will generally be satisfied when $\beta(F)$ is sufficiently smooth and $\hat{F}$ converges at a fast enough rate in a norm. Such conditions are quite well understood. Assumption 1 is more complicated because it involves both bias and stochastic equicontinuity terms. The behavior of these terms seems to be less well understood than the behavior of the nonlinear terms.

Assumption 1 being satisfied is equivalent to the linear functional $\beta(F)=\int\psi(z)F(dz)$ being an asymptotically linear estimator. Thus conditions for linear functionals to be asymptotically linear are also conditions for Assumption 1. For that reason it suffices to confine attention to linear functionals in this Section. Also, for any linear functional of the form $\beta(F)=\int\zeta(z)F(dz)$ we can renormalize so that $\beta(F)-\beta_0=\int\psi(z)F(dz)$ for $\psi(z)=\zeta(z)-E[\zeta(z_i)]$. Then without loss of generality we can restrict attention to functionals $\beta(F)=\int\psi(z)F(dz)$ with $E[\psi(z_i)]=0$.

6.1 Kernel Density Estimators

Conditions for a linear functional of a kernel density estimator to be asymptotically linear were stated though (apparently) not proven in Bickel and Ritov (2003). Here we give a brief exposition of those conditions and a result. Let z be an $r\times 1$ vector and $\hat{F}$ have pdf $\hat{f}(z)=n^{-1}h^{-r}\sum_i K((z-z_i)/h)$. As previously noted, for $\psi(z,h)=\int\psi(z+hu)K(u)du$ we have $\hat{\beta}=n^{-1}\sum_{i=1}^n\psi(z_i,h)$. To make sure that the stochastic equicontinuity condition holds we assume:

ASSUMPTION 3: *$K(u)$ is bounded with bounded support, $\int K(u)du=1$, $\psi(z)$ is continuous almost everywhere, and for some $\varepsilon>0$, $E[\sup_{|t|\leq\varepsilon}\psi(z_i+t)^2]<\infty$.*

From Bickel and Ritov (2003, pp. 1035-1037) we know that the kernel bias for linear functionals is that of a convolution. From equation (5.24) we see that

$$E[\hat{\beta}] - \beta_0 = \int \rho(hu)K(u)du, \rho(t) = \int \psi(z+t)f_0(z)dz = \int \psi(\tilde{z})f_0(\tilde{z}-t)d\tilde{z}.$$

Since $\rho(0) = 0$ the bias in $\hat{\beta}$ is the kernel bias for the convolution $\rho(t)$. A convolution is smoother than the individual functions involved. Under quite general conditions the number of derivatives of $\rho(t)$ that exist will equal the sum of the number of derivatives s_f of $f_0(z)$ that exist and the number of derivatives s_ψ of $\psi(z)$ that exist. The idea is that we can differentiate the first expression for $\rho(t)$ with respect to t up to s_ψ times, do a change of variables $\tilde{z} = z + t$, and then differentiate s_f more times with respect to t to see that $\rho(t)$ is $s_\psi + s_f$ times differentiable. Consequently, the kernel smoothing bias for $\hat{\beta}$ behaves like the kernel bias for a function that is $s_\psi + s_f$ times differentiable. If a kernel of order $s_f + s_\psi$ is used the bias of $\hat{\beta}$ will be of order $h^{s_\psi + s_f}$ that is smaller than the bias order h^{s_f} for the density. Intuitively, the integration inherent in a linear function is a smoothing operation and so leads to bias that is smaller order than in estimation of the density.

Some papers have used asymptotics for kernel based semiparametric estimators based on the supposition that the bias of the semiparametric estimator is the same order as the bias of the nonparametric estimator. Instead the order of the bias of $\hat{\beta}$ is the product of the order of kernel bias for $f_0(z)$ and $\psi(z)$ when the kernel is high enough order. This observations is made in Bickel and Ritov (2003). Newey, Hsieh, and Robins (2004) also showed this result for a twicing kernel, but a twicing kernel is not needed, just any kernel of appropriate order.

As discussed in Bickel and Ritov (2003) a bandwidth that is optimal for estimation of f_0 may also give asymptotic linearity. To see this note that the optimal bandwidth for estimation of f_0 is $n^{-1/(r+2s_f)}$. Plugging in this bandwidth to a bias order of $h^{s_\psi+s_f}$ gives a bias in $\hat{\beta}$ that goes to zero like $n^{-(s_\psi+s_f)/(r+2s_f)}$. This bias will be smaller than $n^{-1/2}$ for $s_\psi > r/2$. Thus, root-n consistency of $\hat{\beta}$ is possible with optimal bandwidth for $\hat{f}$ when the number of derivatives of $\psi(z)$ is more than half the dimension of z. Such a bandwidth will require use of a $s_\psi + s_f$ order kernel, which is higher order than is needed for optimal estimation of f_0. Bickel and Ritov (2003) refer to nonparametric estimators that both converge at optimal rates and for which linear functionals are root-n consistent as plug in estimators, and stated $s_\psi > r/2$ as a condition for existence of a kernel based plug in estimator.

We now give a precise smoothness condition appropriate for kernel estimators. Let $\lambda = (\lambda_1, ..., \lambda_r)^T$ denote a vector of nonnegative integers and $|\lambda| = \sum_{j=1}^r \lambda_j$. Let $\partial^\lambda f(z) = \partial^{|\lambda|} f(z)/\partial z_1^{\lambda_1} \cdot \cdot \cdot \partial z_r^{\lambda_r}$ denote the λ^{th} partial derivative of $f(z)$ with respect to the components of z.

ASSUMPTION 4: *$f_0(z)$ is continuously differentiable of order s_f, $\psi(z)$ is continuously dif-*

ferentiable of order s_ψ, $K(u)$ *is a kernel of order* $s_f + s_\psi$, $\sqrt{n}h^{s_f+s_\psi} \longrightarrow 0$, *and there is* $\varepsilon > 0$ *such that for all* $\lambda, \lambda', \lambda''$ with $|\lambda| \leq s_\psi$, $\left|\lambda'\right| = s_\psi$, and $\left|\lambda'\right| \leq s_f$

$$\int \sup_{|t|\leq\varepsilon} \left|\partial^\lambda \psi(z+t)\right| f_0(z)dz < \infty, \int \left|\partial^{\lambda'}\psi(z)\right| \sup_{|t|\leq\varepsilon} \left|\partial^{\lambda''} f(z+t)\right| dz < \infty$$

Here is a result on asymptotic linearity of kernel estimators of linear functionals.

THEOREM 3: *If Assumptions 3 and 4 are satisfied then* $\int \psi(z)\hat{F}(dz) = \sum_{i=1}^n \psi(z_i)/n + o_p(n^{-1/2})$.

There are many previous results on asymptotic linearity of linear functionals of kernel density estimators. Newey and McFadden (1994) survey some of these. Theorem 3 differs from many of these previous results in Assumption 4 and the way the convolution form of the bias is handled. We follow Bickel and Ritov (2003) in this.

6.2 Series Regression Estimators

Conditions for a linear functional of series regression estimator to be asymptotically linear were given in Newey (1994). It was shown there that the bias of a linear functional of a series estimator is of smaller order than the bias of the series estimator. Here we provide an update to those previous conditions using Belloni, Chernozhukov, Chetverikov, and Kato (2015) on asymptotic properties of series estimators. We give conditions for asymptotic linearity of a linear functional of a series regression estimator of the form

$$\hat{\beta} = \int W(x)\hat{d}(x)dx.$$

We give primitive conditions for the stochastic equicontinuity and bias terms from equation (5.25) to be small.

Let $\hat{\delta}(x) = [\int W(x)p^K(x)dx]^T\hat{\Sigma}^{-1}p^K(x) = E[\delta(x)p^K(x)^T]\hat{\Sigma}^{-1}p^K(x)$ and $\delta(x) = f_0(x)^{-1}W(x)$ as described earlier. The stochastic equicontinuity term will be small if $\sum_{i=1}^n[\hat{\delta}(x_i)-\delta(x_i)]^2/n \xrightarrow{p} 0$. Let $\Sigma = E[p^K(x_i)p^K(x_i)^T]$ and $\gamma = \Sigma^{-1}E[p^K(x_i)d_0(x_i)]$ be the coefficients of the population regression of $d_0(x_i)$ on $p^K(x_i)$. Then the bias term from equation (5.25) satisfies

$$\frac{1}{n}\sum_{i=1}^n \hat{\delta}(x_i)d_0(x_i) = \Gamma^T\hat{\Sigma}^{-1}\sum_{i=1}^n p^K(x_i)[d_0(x_i) - p^K(x_i)^T\gamma]/n + E[\delta(x_i)\{p^K(x_i)^T\gamma - d_0(x_i)\}], \tag{6.26}$$

The first term following the equality is a stochastic bias term that will be $o_p(n^{-1/2})$ under relatively mild conditions from Belloni et. al. (2015). For the coefficients $\gamma_\delta = \Sigma^{-1}E[p^K(x_i)\delta(x_i)]$ of the population projection of $\delta(x_i)$ on $p^K(x_i)$ the second term satisfies

$$E[\delta(x_i)\{p^K(x_i)^T\gamma - d_0(x_i)\}] = -E[\{\delta(x_i) - \gamma_\delta^T p^K(x_i)\}\{d_0(x_i) - p^K(x_i)^T\gamma\}]$$

where the equality holds by $d_0(x_i) - p^K(x_i)^T\gamma$ being orthogonal to $p^K(x_i)$ in the population. As pointed out in Newey (1994), the size of this bias term is determined by the product of series approximation errors to $\delta(x_i)$ and to $d_0(x_i)$. Thus, the bias of a series semiparametric estimator will generally be smaller than the nonparametric bias for a series estimate of $d_0(x)$. For example, for power series if $d_0(x)$ and $\delta(x)$ are continuously differentiable of order s_d and s_δ respectively, x is r-dimensional, and the support of x is compact then by standard approximation theory ,

$$\left|E[\{\delta(x) - \gamma_\delta^T p^K(x)\}\{d_0(x) - p^K(x)^T\gamma\}]\right| \leq CK^{-(s_d+s_\delta)/r}$$

As discussed in Newey (1994) it may be possible to use a K that is optimal for estimation of d_0 and also results in asymptotic linearity. If $s_\delta > r/2$ and K is chosen to be optimal for estimation of d_0 then $\sqrt{n}K^{-(s_d+s_\delta)/r} \longrightarrow 0$. Thus, root-n consistency of $\hat{\beta}$ is possible with optimal number of terms for d_0 when the number of derivatives of $\delta(x)$ is more than half the dimension of z.

Turning now to the regularity conditions for asymptotic linearity, we follow Belloni et. al. (2015) and impose the following assumption that takes care of the stochastic equicontinuity condition and the random bias term.:

ASSUMPTION 5: *$var(q_i|x_i)$ is bounded, $E[\delta(x_i)^2] < \infty$, the eigenvalues of $\Sigma = E[p^K(x_i)p^K(x_i)^T]$ are bounded and bounded away from zero uniformly in K, there is a set χ with $\Pr(x_i \in \chi) = 1$ and c_K and ℓ_K such that $\sqrt{E[\{d_0(x_i) - p^K(x_i)^T\gamma\}^2]} \leq c_K$, $\sup_{x\in\chi}|d_0(x) - p^K(x)^T\gamma| \leq \ell_K c_K$, and for $\xi_K = \sup_{x\in\chi}\left\|p^K(x)\right\|$, we have $K/n + \sqrt{\xi_K^2 \left(\ln K\right)/n}(1 + \sqrt{K}\ell_K c_K) + \ell_K c_K \longrightarrow 0$.*

The next condition takes care of the nonrandom bias term.

ASSUMPTION 6: *$\sqrt{E[\{\delta(x_i) - p^K(x_i)^T\gamma_\delta\}^2]} \leq c_K^\delta$, $c_K^\delta \longrightarrow 0$, and $\sqrt{n}c_K^\delta c_K \longrightarrow 0$.*

Belloni et. al. (2015) give an extensive discussion of the size of c_K, ℓ_K, and ξ_K for various kinds of series approximations and distributions for x_i. For power series Assumptions 5 and 6 are satisfied with $c_K = CK^{-s_d/r}$, $c_K^\delta = CK^{-s_\delta/r}$, $\ell_K = K$, $\xi_K = K$, and

$$\sqrt{K^2 \left(\ln K\right)/n}(1 + K^{3/2}K^{-s_d/r}) + K^{1-(s_d/r)} \longrightarrow 0, \sqrt{n}K^{-(s_d+s_\delta)/r} \longrightarrow 0.$$

For tensor product splines of order o, Assumptions 5 and 6 are satisfied with $c_K = CK^{-\min\{s_d,o\}/r}$, $c_K^\delta = CK^{-\min\{s_\delta,o\}/r}$, $\ell_K = C$, $\xi_K = \sqrt{K}$, and

$$\sqrt{K \left(\ln K\right)/n}(1 + \sqrt{K}K^{-\min\{s_d,o\}/r}) \longrightarrow 0, \sqrt{n}K^{-(\min\{s_d,o\}+\min\{s_\delta,o\})/r} \longrightarrow 0.$$

THEOREM 4: *If Assumptions 5 and 6 are satisfied then for $\psi(z) = \delta(x)[q - d_0(x)]$ we have $\int W(x)\hat{d}(x) = \sum_{i=1}^n \psi(z_i)/n + o_p(n^{-1/2})$.*

Turning now to the consumer surplus bound example, note that in this case $W(x)$ is not even continuous so that $\delta(x)$ is not continuous. This generally means that one cannot assume a rate at which c_K^δ goes to zero. As long as $p^K(x)$ can provide arbitrarily good mean-square approximation to any square integrable function, then $c_K^\delta \longrightarrow 0$ as K grows. Then Assumption 6 will require that $\sqrt{n}c_K$ is bounded. Therefore for power series it suffices for asymptotic linearity of the series estimator of the bound that

$$\sqrt{K^2 (\ln K) /n}(1 + K^{3/2}K^{-s_d/2}) + K^{1-(s_d/2)} \longrightarrow 0, \sqrt{n}K^{-s_d/2} \leq C.$$

For this condition to hold it suffices that $d_0(x)$ is three times differentiable, $K^2 \ln(K)/n \longrightarrow 0$, and K^3/n is bounded away from zero. For regression splines it suffices that

$$\sqrt{K (\ln K) /n}(1 + \sqrt{K}K^{-\min\{s_d,o\}/2}) \longrightarrow 0, \sqrt{n}K^{-\min\{s_d,o\}/2} \leq C.$$

For this condition to hold it suffices that the splines are of order at least 2, $d_0(x)$ is twice differentiable, $K \ln(K)/n \longrightarrow 0$ and K^2/n is bounded away from zero. Here we find weaker sufficient conditions for a spline based estimator to be asymptotically linear than for a power series estimator.

7 Semiparametric GMM Estimators

A more general class of semiparametric estimators that has many applications is the class of generalized method of moment (GMM) estimators that depend on nonparametric estimators. Let $m(z, \beta, F)$ denote a vector of functions of the data observation z, parameters of interest β, and a distribution F. A GMM estimator can be based on a moment condition where β_0 is the unique parameter vector satisfying

$$E[m(z_i, \beta, F_0)] = 0.$$

That is we assume that this moment condition identifies β.

Semiparametric single index estimation provides examples. For the conditional mean restriction, the model assumes the conditional mean function to only depend on the index, so that $E(y|x) = \phi(x^T\theta_0)$. With normalization imposed, first regressor coefficient is 1 so that $\theta_0 = (1, \beta_0^T)^T$. Let $\theta = (1, \beta^T)^T$. Ichimura (1993) showed that under some regularity conditions,

$$\min_\beta E\{[y - E(y|x^T\theta)]^2\}$$

identifies β_0. Thus in this case, $z = (x, y)$ and

$$m(z, \beta, F) = \frac{\partial\{[y - E_F(y|x^T\theta)]^2\}}{\partial\beta}.$$

For the conditional median restriction, the model assumes the conditional median function $M(y|x)$ to only depend on the index, so that $M(y|x) = \phi(x^T\theta_0)$. Ichimura and Lee (2010) showed that under some regularity conditions,

$$\min_{\beta} E\{|y - M(y|x^T\theta)|\}$$

identifies β_0. Thus in this case,

$$m(z, \beta, F) = \frac{\partial\{|y - M_F(y|x^T\theta)|\}}{\partial\beta}.$$

Let $x = (x_1, \tilde{x}^T)^T$. Note that at $\beta = \beta_0$, the derivative of $E(y|x^T\theta)$ with respect to β equals

$$\phi'(x^T\theta_0)[\tilde{x} - E(\tilde{x}|x^T\theta_0)].$$

Thus the target parameter β_0 satisfies the first order condition

$$0 = E\{\phi'(x^T\theta_0)[\tilde{x} - E(\tilde{x}|x^T\theta_0)][y - E(y|x^T\theta_0)]\}.$$

Analogously, at $\beta = \beta_0$, the derivative of $M(Y|X^T\theta)$ with respect to β equals

$$\phi'(x^T\theta_0)[\tilde{x} - E(\tilde{x}|x^T\beta)]/f_{y|x}(M(y|x^T\theta_0)|x).$$

Thus the target parameter β_0 satisfies the first order condition

$$0 = E\{\phi'(x^T\theta_0)[\tilde{x} - E(\tilde{x}|x^T\theta_0)][2 \cdot 1\{y < M(y|x^T\theta_0)\} - 1]/f_{y|x}(M(y|x^T\theta_0)|x)\}.$$

Estimators of β_0 can often be viewed as choosing $\hat{\beta}$ to minimize a quadratic form in sample moments evaluated at some estimator $\hat{F}$ of F_0. For $\hat{m}(\beta) = \sum_{i=1}^n m(z_i, \beta, \hat{F})/n$ and $\hat{W}$ a positive semi-definite weighting matrix the GMM estimator is given by

$$\hat{\beta} = \arg\min_{\beta \in B} \hat{m}(\beta)^T \hat{W} \hat{m}(\beta).$$

In this Section we discuss conditions for asymptotic linearity of this estimator.

For this type of nonlinear estimator showing consistency generally precedes showing asymptotic linearity. Conditions for consistency are well understood. For differentiable $\hat{m}(\beta)$ asymptotic linearity of $\hat{\beta}$ will follow from an expansion of $\hat{m}(\hat{\beta})$ around β_0 in the first order conditions. This gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = -(\hat{M}^T\hat{W}\bar{M})^{-1}\hat{M}^T\hat{W}\sqrt{n}\hat{m}(\beta_0),$$

with probability approaching one, where $\hat{M} = \partial\hat{m}(\hat{\beta})/\partial\beta$, $\bar{M} = \partial\hat{m}(\bar{\beta})/\partial\beta$, and $\bar{\beta}$ is a mean value that actually differs from row to row of $\bar{M}$. Assuming that $\hat{W} \xrightarrow{p} W$ for positive semi-definite W, and that $\hat{M} \xrightarrow{p} M = E[\partial m(z_i, \beta_0, F_0)/\partial\beta]$ and $\bar{M} \xrightarrow{p} M$, it will follow that

$(\hat{M}^T\hat{W}\bar{M})^{-1}\hat{M}^T\hat{W} \xrightarrow{p} (M^TWM)^{-1}M^TW$. Then asymptotic linearity of $\hat{\beta}$ will follow from asymptotic linearity of $\hat{m}(\beta_0)$.

With an additional stochastic equicontinuity condition like that of Andrews (1994), asymptotic linearity of $\hat{m}(\beta_0)$ will follow from asymptotic linearity of functionals of $\hat{F}$. For $F \in \mathcal{F}$ let $\mu(F) = E[m(z_i, \beta_0, F)]$ and

$$\hat{R}_3(F) = \frac{1}{n}\sum_{i=1}^{n}\{m(z_i, \beta_0, F) - m(z_i, \beta_0, F_0) - \mu(F)\}$$

Note that $\sqrt{n}\hat{R}_3(F)$ is the difference of two objects that are bounded in probability (by $E[m(z_i, \beta_0, F_0)] = 0$) and differ only when F is different than F_0. Assuming that $m(z_i, \beta_0, F)$ is continuous in F in an appropriate sense we would expect that $\sqrt{n}\hat{R}_3(F)$ should be close to zero when F is close to F_0. As long as $\hat{F}$ is close to F_0 in large samples in that sense, i.e. is consistent in the right way, then we expect that the following condition holds.

ASSUMPTION 7: $\sqrt{n}\hat{R}_3(\hat{F}) \xrightarrow{p} 0$.

This condition will generally be satisfied when the nonparametrically estimated functions are sufficiently smooth with enough derivatives that are uniformly bounded, see Andrews (1994) and Van der Vaart and Wellner (1996). Under Assumption 7 asymptotic linearity of $\mu(\hat{F})$ will suffice for asymptotic linearity of $\sqrt{n}\hat{m}(\beta_0)$. To see this suppose that $\mu(\hat{F})$ is asymptotically linear with influence function $\varphi(z)$. Then under Assumption 7 and by $\mu(F_0) = E[m(z_i, \beta_0, F_0)] = 0$,

$$\sqrt{n}\hat{m}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i, \beta_0, F_0) + \sqrt{n}\mu(\hat{F}) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(z_i, \beta_0, F_0) + \varphi(z_i)] + o_p(1).$$

Thus Assumption 7 and asymptotic linearity of $\mu(\hat{F})$ suffice for asymptotic linearity of $\hat{m}(\beta_0)$ with influence function $m(z, \beta_0, F_0) + \varphi(z)$. In turn these conditions and others will imply that $\hat{\beta}$ is asymptotically linear with influence function

$$\psi(z) = -(M^TWM)^{-1}M^TW[m(z, \beta_0, F_0) + \varphi(z)].$$

The influence function $\varphi(z)$ of $\mu(F) = E[m(z_i, \beta_0, F)]$ can be viewed as a correction term for estimation of F_0. It can be calculated from equation (3.5) applied to the functional $\mu(F)$. Assumptions 1 and 2 can be applied with $\beta(F) = \mu(F)$ for regularity conditions for asymptotic linearity of $\mu(\hat{F})$. Here is a result doing so

THEOREM 5: *If $\hat{\beta} \xrightarrow{p} \beta_0$, $\hat{W} \xrightarrow{p} W$, $\hat{m}(\beta)$ is continuously differentiable in a neighborhood of β_0 with probability approaching* 1, *for any $\bar{\beta} \xrightarrow{p} \beta_0$ we have $\partial\hat{m}(\bar{\beta})/\partial\beta \xrightarrow{p} M$, M^TWM is nonsingular, Assumptions 1 and 2 are satisfied for $\beta(F) = E[m(z_i, \beta_0, F)]$ and $\psi(z) =$*

$\varphi(z)$, *and Assumption 7 is satisfied then* $\hat{\beta}$ *is asymptotically linear with influence function* $-(M^TWM)^{-1}M^TW[m(z,\beta_0,F_0)+\varphi(z)]$.

For brevity we do not give a full set of primitive regularity conditions for the general GMM setting. They can be formulated using the results above for linear functionals as well as Frechet differentiability, convergence rates, and primitive conditions for Assumption 7.

8 Conclusion

In this paper we have given a method for calculating the influence function of a semiparametric estimator. We have also considered ways to use that calculation to formulate regularity conditions for asymptotic linearity. Consideration of other uses of the influence function are outside the scope of this paper.

9 Appendix A: Proofs

We first give the formulas for the marginal pdf $f_\tau(\tilde{a})$ of a measureable function $a(z_i)$ conditional expectation $E_\tau[b(z_i)|a(z_i)]$ when the expectation is $E_\tau[b(z_i)] = E[b(z_i)\{1+\tau S(z_i)\}]$.

LEMMA A1: *For* $f_\tau(\tilde{z}) = f_0(\tilde{z})[1-\tau+\tau\delta(z)]$ *and* $S(z)=\delta(z)-1$ *the marginal pdf of any measureable function* $a(z_i)$ *is* $f_\tau(\tilde{a}) = f_0(\tilde{a})\{1+\tau E[S(z_i)|a(z_i)=\tilde{a}]\}$ *and for any* $b(z_i)$ *with* $E[|b(z_i)|]<\infty$,

$$E_\tau[b(z_i)|a(z_i)] = \frac{E[b(z_i)|a(z_i)]+\tau E[b(z_i)S(z_i)|a(z_i)]}{1+\tau E[S(z_i)|a(z_i)]}.$$

Proof: Let $1_i = 1(a(z_i)\in\mathcal{A})$ for any measureable set $\mathcal{A}$. By iterated expectations,τ

$$\begin{aligned}
\int 1(\tilde{a} \in \mathcal{A}) f_\tau(a)d\mu &= E[1_i]+\tau E[1_i E[S(z_i)|a_i]] = E[1_i]+\tau E[1_i S(z_i)] = E_\tau[1_i], \\
& E_\tau[1_i \bullet \frac{E[b(z_i)|a(z_i)]+\tau E[b(z_i)S(z_i)|a(z_i)]}{1+\tau E[S(z_i)|a(z_i)]}] \\
&= E[1_i\{E[b(z_i)|a(z_i)]+\tau E[b(z_i)S(z_i)|a(z_i)]\}] = E[1_i b(z_i)]+\tau E[1_i b(z_i)S(z_i)] \\
&= E[1_i b(z_i)\{1+\tau S(z_i)\}] = E[1_i b(z_i)]. Q.E.D.
\end{aligned}$$

Proof of Theorem 1: Note that in a neighborhood of $\tau=0$, $\left[(1-\tau)f_0(\tilde{z})+\tau g_z^h(\tilde{z})\right]^{1/2}$ is continuously differentiable and we have

$$s_\tau(\tilde{z}) = \frac{\partial}{\partial\tau}\left[(1-\tau)f_0(\tilde{z})+\tau g_z^h(\tilde{z})\right]^{1/2} = \frac{1}{2}\frac{g_z^h(\tilde{z})-f_0(\tilde{z})}{\left[\tau g_z^h(\tilde{z})+(1-\tau)f_0(\tilde{z})\right]^{1/2}} \leq C\frac{g_z^h(\tilde{z})+f_0(\tilde{z})}{f_0(\tilde{z})^{1/2}}.$$

By $f_0(\tilde{z})$ bounded away from zero on a neighborhood of z and the support of $g_z^h(\tilde{z})$ shrinking to zero as $h\longrightarrow 0$ it follows that there is a bounded set B with $g_z^h(\tilde{z})/f_0(\tilde{z})^{1/2}\leq C1(\tilde{z}\in B)$ for

h small enough. Therefore, it follows that

$$\int \frac{g_z^h(\tilde{z}) + f_0(\tilde{z})}{f_0(\tilde{z})^{1/2}} d\mu \leq C \int 1(\tilde{z} \in B) d\tilde{z} + 1 < \infty.$$

Then by the dominated convergence theorem $\left[(1-\tau) f_0(\tilde{z}) + \tau g_z^h(\tilde{z})\right]^{1/2}$ is mean-square differentiable and $I(\tau) = \int s_\tau(\tilde{z})^2 d\tilde{z}$ is continuous in τ on a neighborhood of zero for all h small enough. Also, by $g_z^h(\tilde{z}) \longrightarrow 0$ for all $\tilde{z} \neq z$ and $f_0(\tilde{z}) > 0$ on a neighborhood of it follows that $g_z^h(\tilde{z}) \neq f_0(\tilde{z})$ for all τ and h small enough and hence $I(\tau) > 0$. Then by Theorem 7.2 and Example 6.5 of Van der Vaart (1998) it follows that for any $\tau_n = O(1/\sqrt{n})$ a vector of n observations $(z_1, ..., z_n)$ that is i.i.d. with pdf $f_{\tau_n}(\tilde{z}) = (1-\tau_n) f_0(\tilde{z}) + \tau_n g_z^h(\tilde{z})$ is contiguous to a vector of n observations with pdf $f_0(\tilde{z})$. Therefore,

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

holds when $(z_1, ..., z_n)$ are i.i.d. with pdf $f_{\tau_n}(\tilde{z})$.

Next by $\psi(\tilde{z})$ continuous at z, $\psi(\tilde{z})$ is bounded on a neighborhood of z. Therefore for small enough h, $\int \|\psi(\tilde{z})\|^2 g_z^h(\tilde{z}) d\tilde{z} < \infty$, and hence $\int \|\psi(\tilde{z})\|^2 f_\tau(\tilde{z}) d\tilde{z} = (1-\tau) \int \|\psi(\tilde{z})\|^2 f_\tau(\tilde{z}) d\tilde{z} + \tau \int \|\psi(\tilde{z})\|^2 g_z^h(\tilde{z}) d\tilde{z}$ is continuous in τ in a neighborhood of $\tau = 0$. Also, for $\mu_z^h = \int \psi(\tilde{z}) g_z^h(\tilde{z}) d\tilde{z}$ note that $\int \psi(\tilde{z}) f_\tau(\tilde{z}) d\tilde{z} = \tau \mu_z^h$.

Suppose $(z_1, ..., z_n)$ are i.i.d. with pdf $f_{\tau_n}(\tilde{z})$. Let $\beta(\tau) = \beta((1-\tau)F_0 + \tau G_z^h)$ and $\beta_n = \beta(\tau_n)$. Adding and subtracting terms,

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta} - \beta_n\right) &= \sqrt{n}(\hat{\beta} - \beta_0) - \sqrt{n}(\beta_n - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + o_p(1) - \sqrt{n}(\beta_n - \beta_0) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \breve{\psi}_n(z_i) + o_p(1) + \sqrt{n}\tau_n \mu_z^h - \sqrt{n}(\beta_n - \beta_0), \breve{\psi}_n(z_i) = \psi(z_i) - \tau_n \mu_z^h.
\end{aligned}$$

Note that $\int \breve{\psi}_n(\tilde{z}) f_{\tau_n}(\tilde{z}) d\tilde{z} = 0$. Also, for large enough n,

$$\lim_{M \longrightarrow \infty} \int 1(\left\|\breve{\psi}_n(\tilde{z})\right\| \geq M) \left\|\breve{\psi}_n(\tilde{z})\right\|^2 f_{\tau_n}(\tilde{z}) d\tilde{z} \leq \lim_{M \longrightarrow \infty} C \int 1(\|\psi(\tilde{z})\| \geq M/2)(\|\psi(\tilde{z})\|^2 + C) f_0(\tilde{z}) d\tilde{z} \longrightarrow 0,$$

so the Lindbergh-Feller condition for a central limit theorem is satisfied. Furthermore, it follows by similar calculations that $\int \breve{\psi}_n(\tilde{z}) \breve{\psi}_n(\tilde{z})^T f_{\tau_n}(\tilde{z}) d\tilde{z} \longrightarrow V$. Therefore, by the Lindbergh-Feller central limit theorem, $\sum_{i=1}^n \breve{\psi}_n(z_i) \xrightarrow{d} N(0, V)$. Therefore we have $\sqrt{n}(\hat{\beta} - \beta_n) \xrightarrow{d} N(0, V)$ if and only if

$$\sqrt{n}\tau_n \mu_z^h - \sqrt{n}(\beta_n - \beta_0) \longrightarrow 0. \tag{9.27}$$

Suppose that $\beta(\tau)$ is differentiable at $\tau = 0$ with derivative μ_z^h. Then

$$\sqrt{n}(\beta_n - \beta_0) - \sqrt{n}\tau_n \mu_z^h = \sqrt{n} o(\tau_n) = \sqrt{n}\tau_n o(1) \longrightarrow 0$$

by $\sqrt{n}\tau_n$ bounded. Next, we follow the proof of Theorem 2.1 of Van der Vaart (1991), and suppose that eq. (9.27) holds for all $\tau_n = O(1/\sqrt{n})$. Consider any sequence $r_m \longrightarrow 0$. Let n_m be the subsequence such that

$$(1 + n_m)^{-1/2} < r_m \leq n_m^{-1/2}.$$

Let $\tau_n = r_m$ for $n = n_m$ and $\tau_n = n^{-1/2}$ for $n \notin \{n_1, n_2, ...\}$. By construction, $\tau_n = O(1/\sqrt{n})$, so that eq (9.27) holds. Therefore it also holds along the subsequence n_m, so that

$$\sqrt{n_m} r_m \left\{ \mu_z^h - \frac{\beta(r_m) - \beta_0}{r_m} \right\} = \sqrt{n_m} r_m \mu_z^h - \sqrt{n_m}[\beta(r_m) - \beta_0] \longrightarrow 0.$$

By construction $\sqrt{n_m} r_m$ is bounded away from zero, so that $\mu_z^h - [\beta(r_m) - \beta_0]/r_m \longrightarrow 0$. Since r_m is any sequence converging to zero it follows that $\beta(\tau)$ is differentiable at $\tau = 0$ with derivative μ_z^h.

We have now shown that eq. (9.27) holds for all sequences $\tau_n = O(1/\sqrt{n})$ if and only if $\beta(\tau)$ is differentiable at $\tau = 0$ with derivative μ_z^h. Furthermore, as shown above eq. (9.27) holds if and only if $\hat{\beta}$ is regular. Thus we have shown that $\hat{\beta}$ is regular if and only if $\beta(\tau)$ is differentiable at $\tau = 0$ with derivative μ_z^h.

Finally note that as $h \longrightarrow 0$ it follows from continuity of $\psi(\tilde{z})$ at z, $K(u)$ bounded with bounded support, and the dominated convergence theorem that

$$\mu_z^h = \int \psi(\tilde{z}) g_z^h(\tilde{z}) d\tilde{z} = h^{-r} \int \psi(\tilde{z}) K((\tilde{z} - z)/h) d\tilde{z} = \int \psi(z + hu) K(u) du. Q.E.D.$$

Proof of Theorem 2: We will first prove that $E_\tau[\Delta(w_i)|x_i]$ is complete as a function of $\Delta(w_i)$ for all τ small enough. Consider $\tau \in [0, \bar{\tau})$ for $\bar{\tau} = \min\{1/|1 - E_\kappa[\Delta^*(w)]|, 1\}$. Consider any $\Delta(w_i) \neq 0$. Note that for large enough j by $f(y, w|x)$ bounded away from zero,

$$E[\delta(z_i)\Delta(w_i)|x_i] = E_\kappa[\Delta(w_i)]\delta_x(x_i), E_\kappa[\Delta(\tilde{w})] = \int \Delta(w_i)\kappa_w(\tilde{w}) d\mu_{\tilde{w}}.$$

From equation (3.8), $E_\tau[\Delta(w_i)|x_i] = 0$ only if

$$0 = (1 - \tau) E[\Delta(w_i)|x_i] + \tau E_\kappa[\Delta(w_i)] \delta_x(x_i).$$

If $E_\kappa[\Delta(w_i)] = 0$ note that $(1 - \tau)E[\Delta(w_i)|x_i] \neq 0$ by hypothesis i), so that $E_\tau[\Delta(w_i)|x_i] \neq 0$. If $E_\kappa[\Delta(w_i)] \neq 0$ then $E[\Delta(w_i)|x_i] = C\delta_x(x_i)$ for $C = \tau E_\kappa[\Delta(w_i)]/(1 - \tau) \neq 0$. Note that $E[C\Delta^*(w_i)|x_i] = C\delta_x(x_i)$ by hypothesis ii) so by hypothesis i), $\Delta(w_i) = C\Delta^*(w_i)$. Substituting this back in the above equation it follows that

$$0 = \{(1 - \tau)C + \tau C E_\kappa[\Delta^*(w_i)]\}\delta_x(x_i).$$

By $C \neq 0$ and $\delta_x(x_i)$ positive with positive probability this equation implies $0 = (1-\tau) + \tau E_\kappa\left[\Delta^*(w_i)\right]$, which does not hold by $\tau \in [0,\tau)$. Therefore $E_\tau\left[\Delta(w_i)|x_i\right] \neq 0$ for all τ small enough.

Next consider $\gamma(w_i, C) = \gamma_0(w_i) + C\Delta^*(w_i)$ for a constant C. Note that $E_\tau[\gamma(w_i,C)|x_i] = E_\tau[y_i|x_i]$ if and only if

$$E[\gamma(w_i,C)|x_i] + \tau E[\delta(z_i)\gamma(w_i,C)|x_i] = E[y_i|x_i] + \tau E[\delta(z_i)y_i|x_i].$$

Noting that $E[\gamma_0(w_i)|x_i] = E[y_i|x_i]$ and $E[\Delta^*(w_i)|x_i] = \delta_x(x_i)$, this equation holds if any only if

$$C\delta_x(x_i) + \tau(E_\kappa[\gamma_0(w_i)] + CE_\kappa[\Delta^*(\tilde{w})])\delta_x(x_i) = \tau E_\kappa[\tilde{y}]\delta_x(x_i).$$

Since $\delta_x(x_i)$ is positive with positive probablity this equation holds if and only if

$$C + \tau(E_\kappa[\gamma_0(\tilde{w})] + CE_\kappa[\Delta^*(\tilde{w})]) = \tau E_\kappa[\tilde{y}].$$

Solve for $C = c(\tau)$ to obtain

$$c(\tau) = \frac{\tau E_\kappa[\tilde{y} - \gamma_0(\tilde{x})]}{1 + \tau E_\kappa[\Delta^*(w_i)]},$$

for all τ small enough. Then by construction and by differentiating $c(\tau)$ at $\tau = 0$ the conclusion holds. *Q.E.D.*

Proof of Theorem 2: This follows as outlined in the text from Assumptions 1 and 2 and eq. (5.23) and the fact that if several random variables converge in probability to zero then so does their sum. Q.E.D.

Proof of Theorem 3: By the first dominance condition of Assumption 4, $\int \psi(z+t)f(z)dz$ is continuously differentiable with respect t up to order s_ζ in a neighborhood of zero and for all λ with $|\lambda| \leq s_\zeta$,

$$\partial^\lambda \int \psi(z+t)f_0(z)dz = \int \partial^\lambda \psi(z+t)f_0(z)dz.$$

For any λ with $|\lambda| = s_\zeta$ it follows by a change of variables $\tilde{z} = z + t$ and the second dominance condition that

$$\int \partial^\lambda \psi(z+t)f_0(z)dz = \int \partial^\lambda \psi(\tilde{z})f_0(\tilde{z}-t)d\tilde{z}$$

is continuously differentiable in t up to order s_f in a neighborhood of zero and that for any λ' with $|\lambda'| \leq s_f$

$$\partial^{\lambda'} \int \partial^\lambda \psi(\tilde{z})f_0(\tilde{z}-t)d\tilde{z} = \int \partial^\lambda \psi(\tilde{z})\partial^{\lambda'} f_0(\tilde{z}-t)d\tilde{z}.$$

Therefore $\rho(t) = \int \psi(z+t) f_0(z) dz$ is continuously differentiable of order $s_\zeta + s_f$ in a neighborhood of zero. Since $\rho(0) = 0$ and $K(u)$ has bounded support and is order $s_\zeta + s_f$ the usual expansion for kernel bias gives

$$E[\hat{\beta}] - \beta_0 = \int \rho(hu) K(u) du = O(h^{s_\zeta + s_f}).$$

Therefore, $E[\sqrt{n}\hat{R}_1(\hat{F})] \longrightarrow 0$.

Next, by continuity almost everywhere of $\psi(z)$ in Assumption 3 it follows that $\psi(z_i + hu) \longrightarrow \psi(z_i)$ as $h \longrightarrow 0$ with probability one (w.p.1). Also, by Assumption 3 $\sup_{|t| \leq \varepsilon} |\psi(z_i + t)|$ is finite w.p.1, so that by $K(u)$ having bounded support and the dominated convergence theorem, w.p.1,

$$\psi(z_i, h) = \int \psi(z_i + hu) K(u) du \longrightarrow \psi(z_i).$$

Furthermore, for h small enough

$$\psi(z_i, h)^2 \leq C \sup_{|t| \leq \varepsilon} \psi(z_i + t)^2,$$

so it follows by the dominated convergence theorem that $E[\{\psi(z_i, h) - \psi(z_i)\}^2] \longrightarrow 0$ as $h \longrightarrow 0$. Therefore,

$$Var(\sqrt{n}\hat{R}_1(\hat{F})) = Var(n^{-1/2} \sum_{i=1}^{n} \{\psi(z_i, h) - \psi(z_i)\}) \leq E[\{\psi(z_i, h) - \psi(z_i)\}^2] \longrightarrow 0.$$

Since the expectation and variance of $\sqrt{n}\hat{R}_1(\hat{F})$ converges to zero it follows that Assumption 1 is satisfied. Assumption 2 is satisfied because $\beta(F)$ is a linear functional, so the conclusion follows by Theorem 2. *Q.E.D.*

Proof of Theorem 4: Since everything in the remainders is invariant to nonsingular linear transformations of $p^K(x)$ it can be assumed without loss of generality that $\Sigma = E[p^K(x_i) p^K(x_i)^T] = I$. Let $\tilde{\delta}(x_i) = \Gamma^T p^K(x_i) = \gamma_\delta' p^K(x_i)$ so that by Assumption 6, $E[\{\tilde{\delta}(x_i) - \delta(x_i)\}^2] \longrightarrow 0$. Note that by $Var(q_i|x_i)$ bounded and the Markov inequality,

$$\begin{aligned}
\sum_{i=1}^{n} \{\hat{\delta}(x_i) - \delta(x_i)\}^2 Var(q_i|x_i)/n &\leq C \sum_{i=1}^{n} \{\hat{\delta}(x_i) - \delta(x_i)\}^2/n \\
&\leq C \sum_{i=1}^{n} \{\tilde{\delta}(x_i) - \delta(x_i)\}^2/n + C \sum_{i=1}^{n} \{\Gamma^T(\hat{\Sigma}^{-1} - I) p^K(x_i)\}^2/n \\
&\leq o_p(1) + \Gamma^T(\hat{\Sigma}^{-1} - I)\hat{\Sigma}(\hat{\Sigma}^{-1} - I)\Gamma = o_p(1),
\end{aligned}$$

where the last equality follows as in Step 1 of the proof of Lemma 4.1 of Belloni et. al. (2015). We also have

$$\Gamma^T \Gamma = E[\delta(x) p^K(x_i)^T] \Sigma^{-1} E[\delta(x) p^K(x_i)] = E[\{\gamma_\delta^T p^K(x_i)\}^2].$$

By $c_K \longrightarrow 0$ it follows that $E[\{\gamma_\delta^T p^K(x_i)\}^2] \longrightarrow E[\delta(x_i)^2] > 0$, so that $\Gamma \neq 0$. Let $\bar{\Gamma} = \Gamma/(\Gamma^T\Gamma)^{1/2}$, so that $\bar{\Gamma}^T\bar{\Gamma} = 1$. Note that

$$\bar{\Gamma}^T\hat{\Sigma}^{-1}\sum_{i=1}^n p^K(x_i)[d_0(x_i) - p^K(x_i)^T\gamma]/n = \bar{\Gamma}^T(\tilde{\gamma} - \gamma), \tilde{\gamma} = \hat{\Sigma}^{-1}\sum_{i=1}^n p^K(x_i)d_0(x_i)/n$$

Let $R_{1n}(\Gamma)$ and $R_{2n}(\Gamma)$ be defined by the equations

$$\sqrt{n}\bar{\Gamma}^T(\tilde{\gamma} - \gamma) = \bar{\Gamma}^T\sum_{i=1}^n p^K(x_i)[d_0(x_i) - p^K(x_i)^T\gamma]/\sqrt{n} + R_{1n}(\bar{\Gamma}) = R_{1n}(\Gamma) + R_{2n}(\bar{\Gamma}).$$

By eqs. (4.12) and (4.14) of Lemma 4.1 of Belloni et. al. (2015) and by Assumption 5 we have

$$R_{1n}(\bar{\Gamma}) = O_p(\sqrt{\xi_K^2\,(\ln K)\,/n}(1 + \sqrt{K}\ell_K c_K)) \xrightarrow{p} 0, R_{2n}(\bar{\Gamma}) = O_p(\ell_K c_K) \xrightarrow{p} 0.$$

Noting that $\Gamma^T\Gamma \leq E[\delta(x_i)^2] = O(1)$, we have

$$\Gamma^T\hat{\Sigma}^{-1}\sum_{i=1}^n p^K(x_i)[d_0(x_i) - p^K(x_i)^T\gamma]/n = (\Gamma^T\Gamma)^{1/2}\bar{\Gamma}^T(\gamma - \gamma) = O(1)o_p(1) \xrightarrow{p} 0.$$

Also, note that $E[p^K(x_i)\{d_0(x_i) - p^K(x_i)^T\gamma\}] = 0$, so that by the Cauchy-Schwarz inequality,

$$\sqrt{n}\left|E[\delta(x_i)\{d_0(x_i) - p^K(x_i)^T\gamma\}]\right| = \sqrt{n}\left|E[\{\delta(x_i) - p^K(x_i)^T\gamma_\delta\}\{d_0(x_i) - p^K(x_i)^T\gamma\}]\right| \leq \sqrt{n}c_K^\delta c_K \longrightarrow 0.$$

Then the conclusion follows by the triangle inequality and eq. (6.26). *Q.E.D.*

Proof of Theorem 5: As discussed in the text it suffices to prove that $\hat{m}(\beta_0)$ is asymptotically linear with influence function $m(z, \beta_0, F_0) + \alpha(z)$. By Assumption 7 it follows that

$$\hat{m}(\beta_0) = \frac{1}{n}\sum_{i=1}^n m(z_i, \beta_0, F_0) + \mu(\hat{F}) + o_p(n^{-1/2}).$$

Also, by the conclusion of Theorem 1 and $\mu(F_0) = 0$ we have

$$\mu(\hat{F}) = \frac{1}{n}\sum_{i=1}^n \varphi(z_i) + o_p(n^{-1/2}).$$

By the triangle inequality it follows that

$$\hat{m}(\beta_0) = \frac{1}{n}\sum_{i=1}^n [m(z_i, \beta_0, F_0) + \varphi(z_i)] + o_p(n^{-1/2}).Q.E.D.$$

10 References

Ackerberg, D., X. Chen, J. Hahn and Z. Liao (2014): "Asymptotic Efficiency of Semiparametric Two-step GMM" *Review of Economic Studies* 81, 919-943.

Ait-Sahalia, Y. (1991): "Nonparametric Functional Estimation with Applications to Financial Models," MIT Economics Ph. D. Thesis.

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica* 62, 43–72.

Andrews, D.W.K. (2011): "Examples of L2-Complete and Boundedly-Complete Distributions," Cowles Foundation Discussion Paper No. 1801, Yale University.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010): "Estimating Static Models of Strategic Interactions," *Journal of Business and Economic Statistics* 28, 469-482.

Belloni, A., V. Chernozhukov, D. Chetverikov, K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics* 186, 345–366.

Bickel, P. J. and Y. Ritov (1988): "Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates," *Sankhya: The Indian Journal of Statistics, Series A* 50, 381–393.

Bickel, P. J. and Y. Ritov (2003): "Nonparametric Estimators That Can Be Plugged In," *The Annals of Statistics* 31, 1033–1053.

Canay, I.A., A. Santos, and A.H. Shaik (2013): "On the Testability of Identification in Some Nonparametric Models with Endogeneity," *Econometrica* 81, 2535–2559.

Chen, X., and ?? Shen (??).

Chen, X., O. Linton, and I. van Keilegom, (2003): "Estimation of Semiparametric Models When the Criterion Function is not Smooth," *Econometrica* 71, 1591–1608.

Chen, X., V. Chernozhukov, S. Lee, and W. Newey (2014): "Local Identification of Nonparametric and Semiparametric Models," *Econometrica* 82, 785-809.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica* 79, 1541–1565.

Dudley, R. M. (1994): "The Order of the Remainder in Derivatives of Composition and Inverse Operators for p-Variation Norms," *Annals of Statistics* 22, 1–20.

Gill, R. D. (1989): "Non- and Semi-Parametric Maximum Likelihood Estimators and the Von-Mises Method," *Scandanavian Journal of Statistics* 16, 97–128.

Gine, E. and R. Nickl (2008): "A Simple Adaptive Estimator of the Integrated Square of a Density," *Bernoulli* 14, 47–61.

Goldstein, L. and K. Messer (1992): "Optimal Plug-in Estimators for Nonparametric Functional Estimation," *Annals of Statistics* 20, 1306–1328.

Hahn, J., (1998): "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects" *Econometrica* 66, 315-332.

Hahn, J. and G. Ridder (2013): "The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors," *Econometrica* 81, 315-340.

Hahn, J. and G. Ridder (2016): ??.

Hampel, F. R. (1974): "The Influence Curve and Its Role In Robust Estimation," *Journal of the American Statistical Association* 69, 383–393.

Hausman, J. A. and W. K. Newey (2016a): "Individual Heterogeneity and Average Welfare," Econometrica 84,

Hausman, J.A. and W.K. Newey (2016b): "Nonparametric Welfare Analysis," working paper.

Hirano, K., G.W. Imbens, G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Propensity Score," Econometrica 71, 1161-1189.

Ichimura, H. and S. Lee (2010): "Characterization of the asymptotic distribution of semiparametric M-estimators," *Journal of Econometrics* 159, 252–266.

Newey, W.K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica* 59, 1161-1167.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349–1382.

Newey, W. K. and D. L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Vol. 4, Amsterdam, North-Holland, 2113–2245.

Newey, W. K., F. Hsieh, and J. Robins, (2004): "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica* 72, 947–962.

Newey, W.K., and J.L. Powell (1989) "Instrumental Variable Estimation of Nonparametric Models," presented at Econometric Society winter meetings, 1989.

Newey, W.K., and J.L. Powell (2003) "Instrumental Variable Estimation of Nonparametric Models," *Econometrica* 71, 1565-1578.

Reeds, J. A. (1976): "On the Definition of Von Mises Functionals," Ph. D. Thesis, Department of Statistics, Harvard University, Cambridge, MA.

Severini, T. and G. Tripathi (2012): "Efficiency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors," *Journal of Econometrics* 170, 491-498.

Van der Vaart, A. W. (1991): "On Differentiable Functionals," *Annals of Statistics* 19, 178–204.

Van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge, England: Cambridge University Press.

Von Mises, R. (1947): "On the Asymptotic Distribution of Differentiable Statistical Functions," *Annals of Mathematical Statistics* 18, 309-348.