

# The Influence Function of Semiparametric Estimators\*

Hidehiko Ichimura  
University of Tokyo

Whitney K. Newey  
MIT

July 2015

## 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 comparisons, 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

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 for a number of purposes. Its variance is the asymptotic variance of the estimator and so it can be used for asymptotic efficiency comparisons. Also, the form of remainder terms follow from the form of the influence function so knowing the influence function should be a good starting point for finding regularity conditions. In addition, estimators of the influence function can be used to reduce bias of a semiparametric estimator. Furthermore, the influence function approximately gives the influence of a single observation on the estimator. Indeed this interpretation is where the influence function gets its name in the robust estimation literature, see Hampel (1968, 1974).

We show how the influence function of a semiparametric estimator can be calculated from the functional given by the limit of the semiparametric estimator. We show that the influence function is the limit of the Gateaux derivative of the functional with respect to a smooth deviation from the true distribution, as the deviation approaches a point mass. This calculation is similar to that of Hampel (1968, 1974), except that the deviation from the true distribution is restricted to be smooth. Smoothness of the deviation is necessary when the domain of the functional is restricted to smooth functions. As the deviation approaches a point mass the derivative with respect to it approaches the influence function. This calculation applies to many semiparametric estimators that are not defined for point mass deviations, such as those that depend on nonparametric estimators of densities and conditional expectations.

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 pathwise derivative equation. That approach has proven useful in many settings but does require solving a functional equation in some way. The approach of this paper corresponds to specifying a path so that the influence can be calculated directly

from the derivative. This approach eliminates the necessity of finding a solution to a functional equation.

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  $z_1, \dots, z_n$  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  $\beta_0$ , 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  $\beta$ , and a function  $\gamma$  that may be vector valued. Here  $\gamma$  represents some possible value of a nonparametric estimator. A GMM estimator can be based on a moment condition where  $\beta_0$  is the unique parameter vector satisfying

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

and  $\gamma_0$  is the true  $\gamma$ . Here it is assumed that this moment condition identifies  $\beta$ . Let  $\hat{\gamma}$  denote

some nonparametric estimator of  $\gamma_0$ . 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.

Some examples can help illustrate the results. One example is an estimator of a bound on average surplus of a price change when there are bounds on income effects, as formulated in Hausman and Newey (2016). Let  $y$  be quantity consumed of some good,  $x = (x_1, x_2)'$  where  $x_1$  is price,  $x_2$  is income, price is changing between  $\check{x}_1$  and  $\bar{x}_1$ ,  $B$  is a bound on the income effect,  $w_2(x_2)$  a weight function for income (such as an indicator for some interval), and  $\gamma_1(x)$  a possible value of the conditional expectation of quantity given price and income. The object of interest is a bound on the weighted average over income of equivalent variation given by

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

where  $w_1(x_1) = 1(\check{x}_1 \leq x_1 \leq \bar{x}_1) e^{-B(x_1-\check{x}_1)}$ . Here the moment function is

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

Another example is a weighted average derivative of a function  $\gamma_1(w)$  satisfying a conditional moment restriction  $E[y - \gamma_{10}(w_i) | x_i] = 0$ , a nonparametric instrumental variables (NPIV) moment condition as in Newey and Powell (1989, 2003). Here the object of interest is

$$\beta_0 = E[v(w_i) \partial \gamma_{10}(w_i) / \partial w],$$

where  $v(w)$  is some weight function. This object was also considered in Ai and Chen (2007, 2012). Here the moment function is

$$m(z, \beta, \gamma) = v(w) \frac{\partial \gamma(w)}{\partial w} - \beta.$$

In both of these examples the parameter of interest is an explicit functional of a data observation  $z$  and  $\gamma$ . In other interesting cases  $\beta$  is implicitly obtained as the solution to the moment equation (2.1). These other cases include various kinds of structural econometric models, as further discussed below.

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

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

The function  $\psi(z)$  is referred to as the influence function, following terminology of Hampel (1968,1974). 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. We will discuss the influence functions for the examples below.

### 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 CDF  $F$ . We denote this object by  $\beta(F)$ . 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 the real line,

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

In the average surplus example

$$\beta(F) = \int w(\tilde{x}_2)w_1(\tilde{x}_1)E_F[y_i|x_i = \tilde{x}]f(\tilde{x}_2)d\tilde{x},$$

where  $E_F[y_i|x_i]$  denotes the conditional expectation under distribution  $F$  and  $f(x_2)$  is the marginal pdf of  $x_{2i}$  corresponding to  $F$ . Below we will also consider the average derivative and GMM estimators.

We use how  $\beta(F)$  varies with  $F$  to calculate the influence function. Let  $G_z^h$  denote a CDF such that  $(1-t)F_0 + tG_z^h$  is in the domain  $\mathcal{F}$  of  $\beta(F)$  for small enough  $t$  and  $G_z^h$  approaches a point-mass at  $z$  as  $h \rightarrow 0$ . For example, if  $\mathcal{F}$  is restricted to continuous distributions then we could take  $G_z^h$  to be continuous with pdf  $g_z^h(\tilde{z}) = h^{-r}K((\tilde{z} - z)/h)$  for  $K(u)$  a bounded pdf with bounded support and  $\tilde{z}$  denoting a possible value of  $z \in \mathfrak{R}^r$ . Under regularity conditions given below the influence function can be calculated as

$$\psi(z) = \lim_{h \rightarrow 0} \left[ \frac{d}{dt} \beta((1-t) \cdot F_0 + t \cdot G_z^h) \Big|_{t=0} \right]. \quad (3.3)$$

The derivative in this expression is the Gateaux derivative of the functional  $\beta(F)$  with respect to "contamination"  $G_z^h$  to the true distribution  $F_0$ . Thus this formula says that the influence function is the limit of the Gateaux derivative of  $\beta(F)$  as the contamination distribution  $G_z^h$  approaches a point mass at  $z$ .

To illustrate, consider the average surplus bound. Let the contamination distribution  $G_z^h$  have a pdf  $g_z^h(\tilde{z}) = h^{-r}K((\tilde{z} - z)/h)$  for a bounded kernel  $K(u)$  with bounded support. Let  $f_0(x) > 0$  denote the marginal density of  $x_i$  and  $g_x^h(\tilde{x}) = \int g_z^h(\tilde{z})d\tilde{y}$  the marginal pdf of  $x$  for  $g_z^h(\tilde{z})$ . For the parametric model  $F_t$  with pdf  $(1 - t)f_0(\tilde{z}) + tg_z^h(\tilde{z})$  note that

$$E_{F_t}[y_i|x_i = \tilde{x}] = \int \tilde{y} \frac{(1 - t)f_0(\tilde{z}) + tg_z^h(\tilde{z})}{(1 - t)f_0(\tilde{x}) + tg_x^h(\tilde{x})} d\tilde{y}.$$

For  $S(z) = f_0(z)^{-1}g_z^h(z) - 1$  and  $S(x) = f_0(x)^{-1}g_x^h(x) - 1$ , integrating over  $y$  gives  $S(\tilde{x}) = E[S(z_i)|x_i = \tilde{x}]$ . Then differentiating the previous equation gives

$$\left. \frac{\partial E_{F_t}[y_i|x_i = \tilde{x}]}{\partial t} \right|_{t=0} = E[y_i S(z_i)|x_i = \tilde{x}] - \gamma_{10}(\tilde{x})S(\tilde{x}) = \Lambda(\tilde{x}), \quad (3.4)$$

where  $\gamma_{10}(x_i) = E[y_i|x_i]$ ,  $\rho_i = y_i - \gamma_{10}(x_i)$  and  $\Lambda(x_i) = E[\rho_i S(z_i)|x_i]$ . Note that for  $f_0(x_1|x_2) = f_0(x_1, x_2)/f_0(x_2)$  and  $\gamma_{20}(x) = w_1(x_1)w_2(x_2)/f_0(x_1|x_2)$ , iterated expectations implies that

$$\int w(\tilde{x}_2)w_1(\tilde{x}_2)\Lambda(\tilde{x})f_0(\tilde{x}_2)d\tilde{x} = E[\gamma_{20}(x_i)\rho_i S(z_i)]$$

Then applying the chain rule gives

$$\begin{aligned} \frac{d}{dt}\beta(F_t)|_{t=0} &= \int w_2(\tilde{x}_2) \left\{ \int w_1(\tilde{x}_1)\gamma_{10}(\tilde{x})d\tilde{x}_1 \right\} g_{x_2}^h(\tilde{x}_2)d\tilde{x}_2 - \beta_0 \\ &\quad + \int w_1(\tilde{x}_1)w_2(\tilde{x}_2)\Lambda(\tilde{x})f_0(\tilde{x}_2)d\tilde{x} \\ &= \int m(\tilde{z}, \beta_0, \gamma_{10})g_z^h(\tilde{z})d\tilde{z} + E[\gamma_{20}(x_i)\rho_i S(z_i)] \\ &= \int \{m(\tilde{z}, \beta_0, \gamma_{10}) + \gamma_{20}(\tilde{x})(\tilde{y} - E[y_i|x_i = \tilde{x}])\}g_z^h(\tilde{z})d\tilde{z}. \end{aligned}$$

Assuming that  $w_1(\tilde{x}_1)$ ,  $w_2(\tilde{x}_2)$ ,  $f^{-1}(\tilde{x}_1|\tilde{x}_2)$ , and  $E[y_i|x_i = \tilde{x}]$  are continuous at  $z$ , we have

$$\lim_{h \rightarrow 0} \left[ \frac{d}{dt}\beta(F_t)|_{t=0} \right] = m(z, \beta_0, \gamma_{10}) + \gamma_{20}(x)(y - E[y_i|x_i = x]).$$

The expression following the equality is the influence function of an estimator of the average surplus bound.

Equation (3.3) can be thought of as a generalization of the influence function calculation of Hampel (1968, 1974). That calculation is based on contamination  $\delta_z$  that puts probability one on  $z_i = z$ . If  $(1 - t) \cdot F_0 + t \cdot \delta_z$  is the domain  $\mathcal{F}$  of  $\beta(F)$  then the influence function is given by the Gateaux derivative

$$\psi(z) = \frac{d}{dt}\beta((1 - t) \cdot F_0 + t \cdot \delta_z)|_{t=0}.$$

The problem with this calculation is that  $(1 - t) \cdot F_0 + t \cdot \delta_z$  will not be in the domain  $\mathcal{F}$  for many semiparametric estimators. It is not defined for the integrated squared density, average

consumer surplus, nor for any other  $\beta(F)$  that is only well defined for continuous distributions. Equation (3.3) circumvents this problem by restricting the contamination to be in  $\mathcal{F}$ . The influence function is then obtained as the limit of a Gateaux derivative as the contamination approaches a point mass, rather than the Gateaux derivative with respect to a point mass. This generalization applies to most semiparametric estimators.

We can relate the influence function calculation here to the pathwise derivative characterization of the influence function given in Van der Vaart (1991) and Newey (1994). Consider  $(1-t) \cdot F_0 + t \cdot G_z^h$  as a path with parameter  $t$  passing through the truth at  $t = 0$ . It turns out that this path is exactly the right one to get the influence function from the pathwise derivative. Suppose that  $F_0$  has pdf  $f_0$  and  $G_z^h$  has density  $g_z^h$  so that the likelihood corresponding to this path is  $(1-t) \cdot f_0 + t \cdot g_z^h$ . The derivative of the corresponding log-likelihood at zero, i.e. the score, is  $S(z_i) = g_z^h(z_i)/f_0(z_i) - 1$ , where we do not worry about finite second moment of the score for the moment. As shown by Van der Vaart (1991), the influence function will solve the equation

$$\begin{aligned} \frac{d}{dt} \beta((1-t) \cdot F_0 + t \cdot G_z^h)|_{t=0} &= E[\psi(z_i)S(z_i)] \\ &= \int \psi(\tilde{z}) \left[ \frac{g_z^h(\tilde{z})}{f_0(\tilde{z})} - 1 \right] f_0(\tilde{z}) d\tilde{z} = \int \psi(\tilde{z}) g_z^h(\tilde{z}) d\tilde{z}. \end{aligned}$$

Taking the limit as  $h \rightarrow 0$  then gives the formula (3.3) for the influence function when the influence function is continuous at  $z$ . In this way  $F_t = (1-t) \cdot F_0 + t \cdot G_z^h$  can be thought of as a path where the pathwise derivative converges to the influence function as  $g_z^h(z)$  approaches a point mass at  $z$ .

We give a theoretical justification for the formula in equation (3.3) by assuming that an estimator is asymptotically linear and then showing that equation (3.3) is satisfied under a few mild regularity conditions. One of the regularity conditions we use is local regularity of  $\hat{\beta}$  along the path  $F_t$ . This property is that for any  $t_n = O(1/\sqrt{n})$ , when  $z_1, \dots, z_n$  are i.i.d. with distribution  $F_{t_n}$ ,

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

That is, under a sequence of local alternatives, when  $\hat{\beta}$  is centered at  $\beta(F_t)$ , 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  $t$  in a neighborhood of 0, which would imply this condition. Furthermore, it turns out that asymptotic linearity of  $\hat{\beta}$  and Gateaux differentiability of  $\beta(F_t)$  at  $t = 0$  are sufficient for local regularity. For these reasons we view local regularity as a mild condition for the influence function calculation.

For simplicity we give a result for cases where  $F_0$  is a continuous distribution with pdf  $f_0$  and  $\mathcal{F}$  includes paths  $(1-t) \cdot F_0 + t \cdot G_z^h$  where  $G_z^h$  has pdf  $g_z^h(\tilde{z}) = h^{-r}K((z - \tilde{z})/h)$  and

$K(u)$  is a bounded pdf with bounded support. We also show below how this calculation can be generalized to cases where the deviation need not be a continuous distribution.

**THEOREM 1:** *Suppose that  $\hat{\beta}$  is asymptotically linear with influence function  $\psi(\tilde{z})$  that is continuous at  $z$  and  $z_i$  is continuously distributed with pdf  $f_0(\tilde{z})$  that is bounded away from zero on a neighborhood of  $z$ . If  $\hat{\beta}$  is locally regular for the path  $(1-t)F_0 + tG_z^h$  then equation (3.3) is satisfied. Furthermore, if  $\beta((1-t)F_0 + tG_z^h)$  is differentiable at  $t=0$  with derivative  $\int \psi(\tilde{z})g_z^h(\tilde{z})d\tilde{z}$  then  $\hat{\beta}$  is locally regular.*

This result shows that if an estimator is asymptotically linear and certain conditions are satisfied then the influence function satisfies equation (3.3), justifying the calculation of the influence function. Furthermore, the process of that calculation will generally show differentiability of  $\beta((1-t)F_0 + tG_z^h)$  and so imply local regularity of the estimator, confirming one of the hypotheses that is used to justify the formula. In this way this result provides a precise link between the influence function of an estimator and the formula in equation (3.3).

This result is like Van der Vaart (1991) in showing that an asymptotically linear estimator is regular if and only if its limit is pathwise differentiable. It differs in some of the regularity conditions and in restricting the paths to have the mixture form  $(1-t)F_0 + tG_z^h$  with kernel density contamination  $G_z^h$ . 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 a general class of paths.

Although Theorem 1 assumes  $z$  is continuously distributed the calculation of the influence function will work for combinations of discretely and continuously distributed variables. For such cases the calculation can proceed with a deviation that is a product of a point mass for the discrete variables and a kernel density for the continuous variables. More generally, only the variables that are restricted to be continuously distributed in the domain  $\mathcal{F}$  need be continuously distributed in the deviation.

Therefore, assuming that  $\delta(\tilde{x})$  is continuous at  $x$  we have

$$\psi(z) = \lim_{h \rightarrow 0} \left. \frac{\partial \beta(F_t)}{\partial t} \right|_{t=0} = \delta(x)(q - E[q|x]).$$

This result could also be derived using the results for conditional expectation estimators in Newey (1994).

The fact that local regularity is necessary and sufficient for equation (3.3) highlights the strength of the asymptotic linearity condition. Calculating the influence function is a good starting point for showing asymptotic linearity but primitive conditions for asymptotic linearity can be complicated and strong. For example, it is known that asymptotic linearity can require

some degree of smoothness in underlying nonparametric functions, see Bickel and Ritov (1988). We next discuss regularity conditions for asymptotic linearity.

## 4 Nonparametric Instrumental Variables

In this Section we derive the influence function for a semiparametric GMM estimator with first step  $\gamma_{10}$  that satisfies a nonparametric orthogonality condition

$$E[\rho(z_i, \gamma_{10})a(x_i)] = 0, \gamma_{10} \in \Gamma, a(x) \in \mathcal{A}, \quad (4.5)$$

where  $\rho(z, \gamma_1)$  is a function of the data observation  $z$  and first step  $\gamma_1$ ,  $x$  are instrumental variables,  $a(x)$  is a function of  $x$ , and  $\mathcal{A}$  is a linear set that is closed in mean square. When  $\mathcal{A}$  consists of all measurable functions of  $x$  with finite mean square this restriction is equivalent to

$$E[\rho(z_i, \gamma_{10})|x_i] = 0.$$

The conditional mean and  $\tau^{th}$  conditional quantile are special cases where  $\rho(z, \gamma_1) = y - \gamma_1(x)$  or  $\rho(z, \gamma_1) = 1(y < \gamma_1(x)) - \tau$  respectively, for  $0 < \tau < 1$ . Endogeneity can be allowed for by specifying that  $\rho(z, \gamma_1) = y - \gamma_1(w)$  or  $\rho(z, \gamma_1) = 1(y < \gamma_1(w)) - \tau$  and  $w$  is not equal to  $x$ , as in the nonparametric instrumental variables settings of Newey and Powell (1989, 2003, NP), Ai and Chen (2003), Chernozhukov and Hansen (2005) and Chernozhukov, Imbens, and Newey (2007).

This setup also allows for  $\Gamma$  and  $\mathcal{A}$  to be restricted. For example, when  $\Gamma$  and  $\mathcal{A}$  both consist of functions that are additive in individual components of  $x$  and  $\rho(z, \gamma_1) = y - \gamma_1(x)$  then  $\gamma_{10}$  will be the least squares projection of  $y$  on additive functions of  $x$  as in Newey (1994) and Ai and Chen (2007). A generalization that allows for endogeneity would allow  $\gamma_1$  to be additive in functions of  $w$  where  $w$  differs from  $x$ .

We will derive the influence function when the first step is has the same limit as the nonparametric two stage least squares (NP2SLS) estimator. To do so we require the limit  $\gamma_1(F)$  of the NP2SLS estimator when  $F$  is the distribution of each observation. Let  $\pi_F(x_i, \gamma_1)$  denote the least squares projection of  $\rho(z_i, \gamma_1)$  on  $\mathcal{A}$  when  $F$  holds. The predicted value of a regression of the residual  $\rho(z_i, \gamma_1)$  on a vector of approximating functions that belong to  $\mathcal{A}$  will approximate  $\pi_F(x_i, \gamma_1)$ . The NP2SLS objective function is average sum of squares of those predicted values, so the NP2SLS objective function will have limit  $E_F[\pi_F(x_i, \gamma_1)^2]$ . The NP2SLS estimator minimizes that objective function, so  $\gamma_1(F)$  will satisfy

$$\gamma_1(F) = \arg \min_{\gamma_1 \in \Gamma} E_F[\pi_F(x_i, \gamma_1)].$$

To characterize this object it is helpful to have a little more structure. We will assume that there is a linear set  $\mathcal{T}$  such that  $\gamma_1(F) + \tau\Delta \in -$  for a small enough scalar  $\tau$  and that  $\pi_F(x_i, \gamma_1(F) + \tau\Delta)$  is differentiable in  $\tau$  at  $\tau = 0$  with probability one. Let

$$d_F(x_i, \Delta) = \frac{\partial \pi_F(x_i, \gamma_1(F) + \tau\Delta)}{\partial \tau},$$

where derivatives are taken at  $\tau = 0$ . We assume that  $d_F(x_i, \Delta) \in \mathcal{A}$ . The first order conditions for  $\gamma_1(F)$  will be that

$$0 = 2E_F[\pi_F(x_i, \gamma_1(F))d_F(x_i, \Delta)] = 2E_F[\rho(z_i, \gamma_1(F))d_F(x_i, \Delta)] \text{ for all } \Delta \in \mathcal{T}, \quad (4.6)$$

where the second equality holds by the law of iterated projections. We use this first-order condition for  $\gamma_1(F)$  to derive the influence function. To do so we impose further conditions. Let  $\pi(x_i, \gamma_1)$  denote the projection,  $d(x_i, \Delta)$  the function  $d_F(x_i, \Delta)$  and an overbar represent the mean-square closure of a set under the true distribution.

ASSUMPTION 1: *i) With probability one  $d(x_i, \Delta)$  and  $\partial\pi(x_i, \gamma_1(F_t))/\partial t$  exist for every regular parametric model  $F_t$ , ii)  $\mathcal{D} = \{d(x, \Delta) : \Delta \in \mathcal{T}\}$  is linear, and iii)*

$$\bar{\mathcal{D}} = \overline{\{\partial\pi(x_i, \gamma_1(F_t))/\partial t : F_t \text{ is regular}\}} \quad (4.7)$$

This condition allows us to relate the pathwise derivative  $\partial\pi(x_i, \gamma_1(F_t))/\partial t$  to  $d(x_i, \Delta)$ , which helps in deriving the influence function from the first-order condition. To explain Assumption 1 iii), note that if parametric models with  $\gamma(F_t) = \gamma_{10} + t \cdot \Delta$  are among the regular parametric models then we will have  $\bar{\mathcal{D}} \subseteq \overline{\{\partial\pi(x_i, \gamma_1(F_t))/\partial t : F_t \text{ is regular}\}}$  automatically. Also, if there is  $\Delta$  with

$$\pi(x_i, \gamma_1(F_t)) - \pi(x_i, \gamma_{10} + t\Delta) = o(t)$$

in root mean square then the inclusion goes the other way. We expect both of these inclusions to generally hold under sufficient regularity conditions, so that Assumption 1 iii) should hold in general.

ASSUMPTION 2:  *$\partial E[m(z_i, \beta_0, \gamma_1(F_t))]/\partial t$  is a mean-square continuous linear functional of  $\partial\pi(x_i, \gamma_1(F_t))/\partial t$  as  $F_t$  varies across regular parametric models.*

This condition allows us to use the Riesz representation theorem to characterize of the influence function. By Assumption 1 iii) and Assumption 2 it follows that there exists  $\gamma_{20}(x_i) \in \bar{\mathcal{D}}$  such that

$$\partial E[m(z_i, \beta_0, \gamma_1(F_t))]/\partial t = -E[\gamma_{20}(x_i)\partial\pi(x_i, \gamma_1(F_t))/\partial t]. \quad (4.8)$$

Also, note that by  $d_F(x_i, \Delta) \in \mathcal{A}$  and equation (4.6) we have  $E[\rho(z_i, \gamma_{10})d_{F_t}(x_i, \Delta)] = 0$  identically in  $t$ , so that

$$\frac{\partial E[\rho(z_i, \gamma_{10})d_{F_t}(x_i, \Delta)]}{\partial t} = 0.$$

The first order condition in equation (4.6) becomes an identity in  $t$  when evaluated at  $F = F_t$ . Differentiating this identity with respect to  $t$  and applying the chain rule it follows that for all  $\Delta$ ,

$$\begin{aligned} 0 &= \frac{\partial E_{F_t}[\rho(z_i, \gamma_1(F_t))d_{F_t}(x_i, \Delta)]}{\partial t} = \frac{\partial E_{F_t}[\rho(z_i, \gamma_{10})d(x_i, \Delta)]}{\partial t} + \frac{\partial E[\rho(z_i, \gamma_1(F_t))d(x_i, \Delta)]}{\partial t} \\ &= E[\rho(z_i, \gamma_{10})d(x_i, \Delta)S(z_i)] + \frac{\partial E[d(x_i, \Delta)\pi(x_i, \gamma_1(F_t))]}{\partial t} \\ &= E[\rho(z_i, \gamma_{10})d(x_i, \Delta)S(z_i)] + E[d(x_i, \Delta)\frac{\partial \pi(x_i, \gamma_1(F_t))}{\partial t}], \end{aligned}$$

where the third equality follows by interchanging the order of differentiation and integration, which we assume to be possible for any regular parametric model  $F_t$ . Let  $\Delta_j, j = 1, 2$ , be a sequence such that  $d(x_i, \Delta_j)$  converges to  $\gamma_{20}(x_i)$  in mean square. If  $E[\rho(z_i, \gamma_{10})^2|x_i]$  is bounded then  $\rho(z_i, \gamma_{10})d(x_i, \Delta_j)$  converges to  $\gamma_{20}(x_i)\rho(z_i, \gamma_{10})$  in mean square. Therefore evaluating the last equation at  $\Delta_j$  and taking the limit gives

$$0 = E[\gamma_{20}(x_i)\rho(z_i, \gamma_{10})S(z_i)] + E[\gamma_{20}(x_i)\frac{\partial \pi(x_i, \gamma_1(F_t))}{\partial t}].$$

Combining this equation with equation (4.8) gives

$$\frac{\partial E[m(z_i, \beta_0, \gamma_1(F_t))]}{\partial t} = E[\gamma_{20}(x_i)\rho(z_i, \gamma_{10})S(z_i)].$$

This equation and the chain rule give

$$\frac{\partial E_{F_t}[m(z_i, \beta_0, \gamma_1(F_t))]}{\partial t} = E[\psi(z_i)S(z_i)], \psi(z) = m(z, \beta_0, \gamma_{10}) + \gamma_{20}(x)\rho(z, \gamma_{10}).$$

Summarizing, the preceding argument shows the following result:

**PROPOSITION 2:** *If equation (4.5) and Assumptions 1 and 2 are satisfied and  $\hat{\gamma}_1$  has the same limit as the NP2SLS estimator, the influence function of  $\sum_{i=1}^n m(z_i, \beta_0, \hat{\gamma}_1)/n$  is*

$$\psi(z) = m(z, \beta_0, \gamma_{10}) + \gamma_{20}(x)\rho(z, \gamma_{10}).$$

The function  $\gamma_{20}(x)$  is key to the form of the influence function. To explain the role of  $\gamma_{20}(x)$  and to relate Proposition 2 to prior results from the literature it is helpful to consider some examples. An important example is the linear nonparametric instrumental variables model of

NP, where  $\rho(z, \gamma_1) = y - \gamma_1(w)$  and  $\mathcal{A}$  is the set of all measurable functions of  $x$  with finite mean square. This model corresponds to

$$y_i = \gamma_{10}(w_i) + \varepsilon_i, E[\varepsilon_i|x_i] = 0.$$

Consider also a simple example of  $m(z, \beta, \gamma_1)$  where

$$m(z, \beta, \gamma_1) = v(w)\gamma_1(w) - \beta.$$

for some known function  $v(w)$ . For this moment function the true parameter is  $\beta_0 = E[v(w_i)\gamma_{10}(w_i)]$ . Here there is an interesting interpretation of the function  $\gamma_{20}(x)$ . Let  $\Delta(w_i) = \partial\gamma_1(w_i, F_t)/\partial t$  for some parametric model  $F_t$ . Then equation (4.8) is

$$E[v(w_i)\Delta(w_i)] = -E[\gamma_{20}(x_i)E[\Delta(w_i)|x_i]] = E[E[-\gamma_{20}(x_i)|w_i]\Delta(w_i)].$$

Subtracting gives

$$E[\{v(w_i) + E[\gamma_{20}(x_i)|w_i]\}\Delta(w_i)] = 0.$$

Assuming that the set of derivatives spans in mean square the set of functions of  $w_i$  the previous equation implies

$$v(w_i) = E[-\gamma_{20}(x_i)|w_i].$$

Thus we see that Assumption 2 implies that there exists a function of  $x_i$  whose conditional expectation given  $w_i$  is  $v(w_i)$ . We can also work backwards to obtain the converse, so that we have the following result:

**PROPOSITION 3:** *If  $\rho(z, \gamma_1) = y - \gamma_1(w)$  and  $m(z, \beta, \gamma_1) = v(w)\gamma_1(w) - \beta$  then Assumption 2 is satisfied if and only if there is a function  $\alpha(x)$  with*

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

From Lemma 4.1 of Severini and Tripathi (2012) it is known that existence of such a function  $\alpha(x_i)$  is a necessary condition for root-n consistent estimability of  $\beta_0$ . Proposition 3 says that this necessary condition is equivalent to Assumption 2. In the exogenous case Newey (1994) and Newey and McFadden (1994) argued that Assumption 2 was a necessary condition for existence of a root-n consistent estimator. Proposition 3 extends that result to models with endogeneity.

Proposition 3 can be generalized to other functionals such as the average derivative of Ai and Chen (2007), as in the following result:

PROPOSITION 3: If  $\rho(z, \gamma_1) = y - \gamma_1(w)$  and there is  $v(w)$  with  $E[m(z, \beta_0, \gamma_1)] = E[v(w)\gamma_1(w)] - \beta_0$  then Assumption 2 is satisfied if and only if there is a function  $\alpha(x)$  with

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

In particular if  $m(z, \beta, \gamma_1) = \partial\gamma_1(w)/\partial w - \beta$  and the density  $f_0(w)$  of  $w$  is differentiable on the boundary of its support then Assumption 2 is satisfied if and only if there is  $\alpha(x)$  with

$$\partial \ln f_0(w_i)/\partial w = E[\alpha(x_i)|w_i].$$

Let  $E_{F_t}[\cdot|x_i]$  and  $E_{F_t}[\cdot]$  be the associated conditional and unconditional expectations respectively for the parametric model  $F_t$ ,  $\rho_i = y_i - \gamma_{10}(w_i)$ , and  $\bar{\rho}_i = E[\rho_i|x_i]$ . Here we leave open the possibility that  $\bar{\rho}_i$  is not zero due to misspecification where the conditional moment restriction does not hold. Then as in Section ??,

$$\begin{aligned} \frac{\partial}{\partial t} E_{F_t} [\rho_i E [\Delta (w_i) | x_i]] &= E[\rho_i E [\Delta (w_i) | x_i] S(z_i)], \\ \frac{\partial}{\partial t} E [\rho_i E_{F_t} [\Delta (w_i) | x_i]] &= \frac{\partial}{\partial t} E[\bar{\rho}_i E_{F_t} [\Delta (w_i) | x_i]] \\ &= E[\bar{\rho}_i E [\{\Delta (w_i) - E[\Delta(w_i)|x_i]\} S(z_i) | x_i]] \\ &= E[\bar{\rho}_i \{\Delta (w_i) - E[\Delta(w_i)|x_i]\} S(z_i)]. \end{aligned}$$

Then applying the chain rule in differentiating the first order conditions in equation (??) with  $F = F_t$  gives

$$\frac{\partial E[E[\Delta(w_i)|x_i]\gamma_1(w_i, F_t)]}{\partial t} = E[\{\rho_i E[\Delta(w_i)|x_i] + \bar{\rho}_i \{\Delta(w_i) - E[\Delta(w_i)|x_i]\}\} S(z_i)]$$

for all  $\Delta(w_i)$  with  $E[\Delta(w_i)^2] < \infty$ . If there is  $\Delta(w_i)$  such that  $\gamma_{20}(x_i) = E[\Delta(w_i)|x_i]$  then plugging  $\gamma_{20}(x_i)$  in the previous equation and using  $S(z) = f_0(\tilde{z})^{-1} g_z^h(\tilde{z}) - 1$  we have

PROPOSITION 2: If Assumptions 1 and 2 are satisfied and there is  $\Delta(w_i)$  with  $\gamma_{20}(x_i) = E[\Delta(w_i)|x_i]$  and  $E[\Delta(w_i)^2] < \infty$  then the influence function of  $\hat{\beta}$  is

$$\psi(z_i) = m(z_i, \beta_0, \gamma_1) + \gamma_{20}(x_i)\rho_i + \bar{\rho}_i[\Delta(w_i) - \gamma_{20}(x_i)].$$

When the conditional moment conditions are correctly specified so that  $\bar{\rho}_i = 0$  then the last term in the influence function drops out. Also, it is possible to drop the Assumption that  $\gamma_{20}(x_i) = E[\Delta(w_i)|x_i]$  for some  $\Delta(w_i)$ . Note that there exists  $\Delta_j(w_i)$  such that  $\gamma_j(x_i) = E[\Delta_j(w_i)|x_i] \rightarrow \gamma_{20}(x_i)$  in mean square. Assume that  $E[\rho_i^2|x_i]$  is bounded, so that  $\gamma_j(x_i)\rho_i \rightarrow$

$\gamma_{20}(x_i)\rho_i$  in mean square. Then plugging in the above equation and taking limits as  $j \rightarrow \infty$  gives

PROPOSITION 3: *If Assumptions 1 and 2 are satisfied,  $E[\rho_i|x_i] = 0$ , and  $E[\rho_i^2|x_i]$  is bounded then the influence function of  $\hat{\beta}$  is*

$$\psi(z_i) = m(z_i, \beta_0, \gamma_1) + \gamma_{20}(x_i)\rho_i.$$

In the misspecified case where  $\bar{\rho}_i$  it would be interesting to know more about the importance of existence of  $\Delta(w_i)$  with  $E[\Delta(w_i)|x_i] = \gamma_{20}(x_i)$ . That condition may be required for root-n consistent estimation in the misspecified case though it is beyond the scope of this paper to check that.

## 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 once  $\psi(z)$  has been calculated we know what the remainder term for asymptotic linearity must be. The remainder term can then be analyzed in order to formulate conditions for it to be small and hence the estimator be asymptotically linear. In this section we give one way to specify conditions for the remainder term to be small. It is true that this formulation may not lead to the weakest possible conditions for asymptotic linearity of a particular estimator. It is only meant to provide a useful way to formulate conditions for asymptotic linearity.

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

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta_0) - \sum_{i=1}^n \psi(z_i)/\sqrt{n} &= \sqrt{n}\hat{R}_1(\hat{F}) + \sqrt{n}R_2(\hat{F}), \\ \hat{R}_1(F) &= \int \psi(z)F(dz) - \sum_{i=1}^n \psi(z_i)/n, \quad R_2(F) = \beta(F) - \beta(F_0) - \int \psi(z)F(dz). \end{aligned} \tag{5.9}$$

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 (5.9) 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  $\sqrt{n}\hat{R}_1(\hat{F})$  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  $\chi_n$  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  $\chi_n$  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

$$E[\hat{R}_1(\hat{F})] = E[\psi(z_i, h)] = \int \left[ \int \psi(z + hu)f_0(z)dz \right] K(u)du, \quad (5.10)$$

$$\hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})] = \frac{1}{n} \sum_{i=1}^n \{ \psi(z_i, h) - E[\psi(z_i, h)] - \psi(z_i) \}.$$

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  $\lim_{h \rightarrow 0} E[\{ \psi(z_i, h) - \psi(z_i) \}^2] = 0$ .

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  $\chi_n = \{x_1, \dots, x_n\}$ . Then we have

$$\begin{aligned} E[\hat{R}_1(\hat{F})|\chi_n] &= \frac{1}{n} \sum_{i=1}^n \hat{\delta}(x_i)d_0(x_i) - \beta_0, \\ \hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})|\chi_n] &= \frac{1}{n} \sum_{i=1}^n [\hat{\delta}(x_i) - \delta(x_i)][q_i - d_0(x_i)]. \end{aligned} \quad (5.11)$$

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  $x_1, \dots, x_n$ , the expected square of this term conditional on  $\chi_n$  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  $\|F - F_0\| = \{\int [f(z) - f_0(z)]^2 dz\}^{1/2}$  is the root integrated squared error. Consequently  $\sqrt{n} \|\hat{F} - F_0\|$  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  $\|\hat{F} - F_0\|$  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  $\|\hat{F} - F_0\| = o_p(n^{-1/2\zeta})$ .

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 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 non-parametric 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  $\sqrt{n} \|\hat{F} - F_0\|$  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 (5.10) 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  $s_\psi$  of  $\psi(z)$  that exist. The idea is that we can differentiate the first expression for  $\rho(t)$  with respect to  $t$  up to  $s_\psi$  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  $\lambda = (\lambda_1, \dots, \lambda_r)^T$  denote a vector of nonnegative integers and  $|\lambda| = \sum_{j=1}^r \lambda_j$ . Let  $\partial^\lambda f(z) = \partial^{|\lambda|} f(z) / \partial z_1^{\lambda_1} \dots \partial z_r^{\lambda_r}$  denote the  $\lambda^{th}$  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 differentiable of order  $s_\psi$ ,  $K(u)$  is a kernel of order  $s_f + s_\psi$ ,  $\sqrt{nh}^{s_f+s_\psi} \rightarrow 0$ , and there is  $\varepsilon > 0$  such that for all  $\lambda, \lambda', \lambda''$  with  $|\lambda| \leq s_\psi$ ,  $|\lambda'| = s_\psi$ , and  $|\lambda''| \leq s_f$

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

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 (5.11) 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 (5.11) satisfies

$$\frac{1}{n} \sum_{i=1}^n \hat{\delta}(x_i)d_0(x_i) = \Gamma^T \hat{\Sigma}^{-1} \sum_{i=1}^n p^K(x_i)[d_0(x_i) - p^K(x_i)^T \gamma]/n + E[\delta(x_i)\{p^K(x_i)^T \gamma - d_0(x_i)\}], \quad (6.12)$$

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  $s_\delta$  respectively,  $x$  is  $r$ -dimensional, and the support of  $x$  is compact then by standard approximation theory ,

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

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  $\sqrt{n}K^{-(s_d+s_\delta)/r} \rightarrow 0$ . 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  $\chi$  with  $\Pr(x_i \in \chi) = 1$  and  $c_K$  and  $\ell_K$  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  $\xi_K = \sup_{x \in \chi} \|p^K(x)\|$ , we have  $K/n + \sqrt{\xi_K^2 (\ln K) / n(1 + \sqrt{K} \ell_K c_K)} + \ell_K c_K \rightarrow 0$ .

The next condition takes care of the nonrandom bias term.

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

Belloni et. al. (2015) give an extensive discussion of the size of  $c_K$ ,  $\ell_K$ , and  $\xi_K$  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

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

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

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

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  $c_K^\delta$  goes to zero. As long as  $p^K(x)$  can provide arbitrarily good mean-square approximation to any square integrable function, then  $c_K^\delta \rightarrow 0$  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

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

For this condition to hold it suffices that  $d_0(x)$  is three times differentiable,  $K^2 \ln(K) / n \rightarrow 0$ , 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  $\beta$ , and a distribution  $F$ . A GMM estimator can be based on a moment condition where  $\beta_0$  is the unique parameter vector satisfying

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

That is we assume that this moment condition identifies  $\beta$ .

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  $\beta_0$ . 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  $\beta_0$ . 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  $\beta$  equals

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

Thus the target parameter  $\beta_0$  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  $\beta$  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  $\beta_0$  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  $\beta_0$  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  $\beta_0$  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  $(\hat{M}^T \hat{W} \bar{M})^{-1} \hat{M}^T \hat{W} \xrightarrow{p} (M^T W M)^{-1} M^T W$ . 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

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

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 (3.3) 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  $\beta_0$  with probability approaching 1, for any  $\bar{\beta} \xrightarrow{p} \beta_0$  we have  $\partial \hat{m}(\bar{\beta}) / \partial \beta \xrightarrow{p} M$ ,  $M^T W M$  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  $-(M^T W M)^{-1} M^T W [m(z, \beta_0, F_0) + \varphi(z)]$ .*

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

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

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

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  $h \rightarrow 0$  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  $[(1-t)f_0(\tilde{z}) + tg_z^h(\tilde{z})]^{1/2}$  is mean-square differentiable and  $I(t) = \int s_t(\tilde{z})^2 d\tilde{z}$  is continuous in  $t$  on a neighborhood of zero for all  $h$  small enough. Also, by  $g_z^h(\tilde{z}) \rightarrow 0$  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  $t$  and  $h$  small enough and hence  $I(t) > 0$ . Then by Theorem 7.2 and Example 6.5 of Van der Vaart (1998) it follows that for any  $t_n = O(1/\sqrt{n})$  a vector of  $n$  observations  $(z_1, \dots, z_n)$  that is i.i.d. with pdf  $f_{t_n}(\tilde{z}) = (1-t_n)f_0(\tilde{z}) + t_n g_z^h(\tilde{z})$  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  $(z_1, \dots, z_n)$  are i.i.d. with pdf  $f_{t_n}(\tilde{z})$ .

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  $\int \|\psi(\tilde{z})\|^2 f_t(\tilde{z}) d\tilde{z} = (1-t) \int \|\psi(\tilde{z})\|^2 f_0(\tilde{z}) d\tilde{z} + t \int \|\psi(\tilde{z})\|^2 g_z^h(\tilde{z}) d\tilde{z}$  is continuous in  $t$  in a neighborhood of  $t = 0$ . Also, for  $\mu_z^h = \int \psi(\tilde{z}) g_z^h(\tilde{z}) d\tilde{z}$  note that  $\int \psi(\tilde{z}) f_t(\tilde{z}) d\tilde{z} = t\mu_z^h$ .

Suppose  $(z_1, \dots, z_n)$  are i.i.d. with pdf  $f_{t_n}(\tilde{z})$ . Let  $\beta(t) = \beta((1-t)F_0 + tG_z^h)$  and  $\beta_n = \beta(t_n)$ .

Adding and subtracting terms,

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

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

$$\lim_{M \rightarrow \infty} \int 1(\|\check{\psi}_n(\tilde{z})\| \geq M) \|\check{\psi}_n(\tilde{z})\|^2 f_{t_n}(\tilde{z})d\tilde{z} \leq \lim_{M \rightarrow \infty} C \int 1(\|\psi(\tilde{z})\| \geq M/2)(\|\psi(\tilde{z})\|^2 + C)f_0(\tilde{z})d\tilde{z} \rightarrow 0,$$

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

$$\sqrt{nt_n}\mu_z^h - \sqrt{n}(\beta_n - \beta_0) \rightarrow 0. \quad (9.13)$$

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

$$\sqrt{n}(\beta_n - \beta_0) - \sqrt{nt_n}\mu_z^h = \sqrt{no}(t_n) = \sqrt{nt_n}o(1) \rightarrow 0$$

by  $\sqrt{nt_n}$  bounded. Next, we follow the proof of Theorem 2.1 of Van der Vaart (1991), and suppose that eq. (9.13) holds for all  $t_n = O(1/\sqrt{n})$ . Consider any sequence  $r_m \rightarrow 0$ . Let  $n_m$  be the subsequence such that

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

Let  $t_n = r_m$  for  $n = n_m$  and  $t_n = n^{-1/2}$  for  $n \notin \{n_1, n_2, \dots\}$ . By construction,  $t_n = O(1/\sqrt{n})$ , so that eq (9.13) holds. Therefore it also holds along the subsequence  $n_m$ , so that

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

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

We have now shown that eq. (9.13) holds for all sequences  $t_n = O(1/\sqrt{n})$  if and only if  $\beta(t)$  is differentiable at  $t = 0$  with derivative  $\mu_z^h$ . Furthermore, as shown above eq. (9.13) holds if and only if  $\hat{\beta}$  is regular. Thus we have shown that  $\hat{\beta}$  is regular if and only if  $\beta(t)$  is differentiable at  $t = 0$  with derivative  $\mu_z^h$ .

Finally note that as  $h \rightarrow 0$  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:** This follows as outlined in the text from Assumptions 1 and 2 and eq. (5.9) 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  $s_\zeta$  in a neighborhood of zero and for all  $\lambda$  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  $\lambda$  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  $\lambda'$  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,  $E[\sqrt{n}\hat{R}_1(\hat{F})] \rightarrow 0$ .

Next, by continuity almost everywhere of  $\psi(z)$  in Assumption 3 it follows that  $\psi(z_i + hu) \rightarrow \psi(z_i)$  as  $h \rightarrow 0$  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,

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

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  $E[\{\psi(z_i, h) - \psi(z_i)\}^2] \rightarrow 0$  as  $h \rightarrow 0$ . Therefore,

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

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,  $E[\{\tilde{\delta}(x_i) - \delta(x_i)\}^2] \rightarrow 0$ . Note that by  $\text{Var}(q_i|x_i)$  bounded and the Markov inequality,

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

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  $c_K \rightarrow 0$  it follows that  $E[\{\gamma_\delta^T p^K(x_i)\}^2] \rightarrow E[\delta(x_i)^2] > 0$ , 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,

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

Then the conclusion follows by the triangle inequality and eq. (6.12). *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

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

## 10 References

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.

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.

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.

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,” *Scandinavian 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.

Hampel, F. R. (1968): *Contributions to the Theory of Robust Estimation*, Ph. D. Thesis, Univ. California, Berkeley.

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 (2015): “Individual Heterogeneity and Average Welfare,” working paper, MIT.

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

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.

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.