CIRJE-F-985

The Influence Function of Semiparametric Estimators

Hidehiko Ichimura
University of Tokyo

Whitney K. Newey
Massachusetts Institute of Technology

August 2015

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 (1998), 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

The subject of this paper is estimators of parameters that depend on unknown functions such as probability densities or conditional expectations. We consider estimators of these parameters based on nonparametric estimates of the unknown functions. We refer to these estimators as semiparametric, with the understanding that they depend on nonparametric estimators. We could also refer to them as "plug in estimators" or more precisely as "plug in estimators that have an influence function." 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.

Some examples can help fix ideas. One example with a long history is the integrated squared density where $\beta_0 = \int f_0(z)^2 dz$, z_i has pdf $f_0(z)$, and z is r-dimensional. This object is useful in certain testing settings. A variety of different estimators of β_0 have been suggested. One estimator is based on a kernel estimator $\hat{f}(z)$ of the density given by

$$\hat{f}(z) = \frac{1}{nh^r} \sum_{i=1}^{n} K(\frac{z - z_i}{h}), \int K(u)du = 1,$$

where h is a bandwidth and $K(u)$ is a kernel. An estimator $\hat{\beta}$ can then be constructed by plugging in $\hat{f}$ in place of f_0 in the formula for β_0 as

$$\hat{\beta} = \int \hat{f}(z)^2 dz.$$

This estimator of β_0 and other estimators have been previously considered by many others. We use it as an example to help illustrate the results of this paper.

It is known that there are other estimators that are better than $\hat{\beta}$. One of these is

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(z_i), \hat{f}_{-i}(z) = \frac{1}{(n-1)h^r}\sum_{j\neq i} K(\frac{z - z_j}{h}),$$

where $K(u)$ is a symmetric kernel. Gine and Nickl (2008) showed that this estimator converges at optimal rates while it is well known that $\hat{\beta}$ does not. Our purpose in considering $\hat{\beta}$ is not to suggest it as the best estimator but instead to use it to illustrate the results of this paper.

Another example is based on the bound on average consumer surplus given in Hausman and Newey (2015). Here a data observation is $z = (q, p, y)$ where q is quantity of some good, p is price, and y is income. For $x = (p, y)$ the object of interest is

$$\beta_0 = \int W(x) d_0(x) dx, W(x) = w(y) 1(p^0 \leq p \leq p^1) e^{-b(p-p^0)}, d_0(x) = E[q|x].$$

From Hausman and Newey (2015) it follows that this object is a bound on the weighted average over income and individuals of average equivalent variation for a price change from p^0 to p^1 when there is general heterogeneity. It is an upper (or lower) bound for average surplus when b is a lower (or upper) bound for individual income effects. Here $w(y) \geq 0$ is a known weight function that is used to average across income levels.

One estimator of β_0 can be obtained by plugging-in a series nonparametric regression estimator of $d_0(x)$ in the formula for β_0. To describe a series estimator let $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))^T$ be a vector of approximating functions such as power series or regression splines. Also let $P = [p^K(x_1), ..., p^K(x_n)]^T$ and $Q = (q_1, ..., q_n)^T$ be the matrix and vector of observations on the approximating functions and on quantity. A series estimator of $d_0(x) = E[q|x]$ is given by

$$\hat{d}(x) = p^K(x)^T \hat{\gamma}, \hat{\gamma} = \hat{\Sigma}^{-1} P^T Q/n, \hat{\Sigma} = P^T P/n,$$

where $P^T P$ will be nonsingular with probability approaching one under conditions outlined below. We can then plug in this estimator to obtain

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

We use this estimator as a second example.

This paper is about estimators that have an influence function. We and others refer to these as asymptotically linear estimators. An asymptotically linear estimator is one satisfying

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} + o_p(1), E[\psi(z_i)] = 0, E[\psi(z_i)^T \psi(z_i)] < \infty. \tag{2.1}$$

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

In the integrated squared density example the influence function is well known to be

$$\psi(z) = 2[f_0(z) - \beta_0].$$

This formula holds for the estimators mentioned above and for all other asymptotically linear estimators of the integral of the square of an unrestricted pdf. In the consumer surplus example the influence function is

$$\psi(z) = \delta(x)[q - d_0(x)], \delta(x) = f_0(x)^{-1} W(x).$$

as will be shown 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,

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

In the integrated squared density example $\beta(F) = \int f(z)^2 dz$, where all elements of the domain $\mathcal{F}$ are restricted to be continuous distributions with pdfs that are square integrable. In the average surplus example $\beta(F) = \int W(x) E_F[q|x] dx$ where the domain is restricted to distributions where $E_F[q|x]$ and $\beta(F)$ exist and x is continuously distributed with pdf $f_0(x)$ that is positive where $W(x)$ is positive.

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 $h \longrightarrow 0$. 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 $z \in \Re^r$. Under regularity conditions given below the influence function can be calculated as

$$\psi(z) = \lim_{h \longrightarrow 0} \left[\frac{d}{dt} \beta((1-t) \cdot F_0 + t \cdot G_z^h)|_{t=0} \right]. \tag{3.2}$$

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.

For example, consider the integrated squared density where we 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 $K(u)$. Then

$$\begin{aligned} \frac{d}{dt}\beta((1-t) \cdot F_0 + t \cdot G_z^h)|_{t=0} &= \frac{d}{dt}\left\{ \int [(1-t) \cdot f_0(\tilde{z}) + t \cdot g_z^h(\tilde{z})]^2 d\tilde{z} \right\}|_{t=0} \\ &= \int 2[f_0(\tilde{z}) - \beta_0] g_z^h(\tilde{z}) d\tilde{z}. \end{aligned}$$

Assuming that $f_0(\tilde{z})$ is continuous at z, the limit as $h \longrightarrow 0$ is given by

$$\lim_{h \longrightarrow 0} \left[\frac{d}{dt} \beta((1-t) \cdot F_0 + t \cdot G_z^h)|_{t=0} \right] = 2 \lim_{h \longrightarrow 0} \int f_0(\tilde{z}) g_z^h(\tilde{z}) d\tilde{z} - 2\beta_0 = 2[f_0(z) - \beta_0].$$

This function is the influence function at z of semiparametric estimators of the integrated squared density. Thus equation (3.2) holds in the example of an integrated squared density. As we show below, equation (3.2), including the Gateaux differentiability, holds for any asymptotically linear estimator satisfying certain mild regularity conditions.

Equation (3.2) can be thought of as a generalization of the influence function calculation of Hampel (1968, 1974). That calculation is based on contamination δ_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 (3.2) 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 $h \longrightarrow 0$ then gives the formula (3.2) 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 (3.2) by assuming that an estimator is asymptotically linear and then showing that equation (3.2) 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 $z_1, ..., z_n$ 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*

(3.2) 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 (*3.2*), 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 (*3.2*).

This result is like Van der Vaart (1991) in showing that an asymptotically linear estimator is regular if an 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.

We can illustrate using the consumer surplus example. Consider a deviation that is a product of a point mass δ_q at some q and a kernel density $g_x^h(\tilde{x}) = h^{-2}K((\tilde{x}-x)/h)$ centered at $x = (p, y)$. The corresponding path is

$$F_t = (1-t)F_0 + t\delta_q G_x^h,$$

where G_x^h is the distribution corresponding to $g_x^h(\tilde{x})$. Let $f_t(\tilde{x}) = (1-t)f_0(\tilde{x}) + tg^h(\tilde{x})$ be the marginal pdf for x along the path. Multiplying and dividing by $f_t(\tilde{x})$ and using iterated expectations we find that

$$\beta(F_t) = \int W(\tilde{x})E_{F_t}[q|\tilde{x}]d\tilde{x} = \int f_t(\tilde{x})^{-1}W(\tilde{x})E_{F_t}[q|\tilde{x}]f_t(\tilde{x})dx = E_{F_t}[f_t(x_i)^{-1}W(x_i)q_i].$$

Differentiating with respect to t gives

$$\begin{aligned}\left.\frac{\partial\beta(F_t)}{\partial t}\right|_{t=0} &= q\int \delta(\tilde{x})g^h(\tilde{x})d\tilde{x} - \beta_0 \\ &\quad + \int(-1)f_0(\tilde{x})^{-2}[g^h(\tilde{x}) - f_0(\tilde{x})]W(\tilde{x})E[q|\tilde{x}]f_0(\tilde{x})d\tilde{x} \\ &= q\int \delta(\tilde{x})g^h(\tilde{x})d\tilde{x} - \int \delta(\tilde{x})E[q|\tilde{x}]g^h(\tilde{x})d\tilde{x}.\end{aligned}$$

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

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

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 (*3.2*) 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 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 6.

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

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta_0) - \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} &= \sqrt{n}\hat{R}_1(\hat{F}) + \sqrt{n}R_2(\hat{F}), \qquad (4.3) \\ \hat{R}_1(F) &= \int \psi(z)F(dz) - \sum_{i=1}^{n} \psi(z_i)/n, \quad R_2(F) = \beta(F) - \beta(F_0) - \int \psi(z)F(dz). \end{aligned}$$

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 (4.3) 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

$$\begin{aligned} E[\hat{R}_1(\hat{F})] &= E[\psi(z_i,h)] = \int [\int \psi(z+hu)f_0(z)dz]K(u)du, \qquad (4.4) \\ \hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})] &= \frac{1}{n}\sum_{i=1}^n \{\psi(z_i,h) - E[\psi(z_i,h)] - \psi(z_i)\}. \end{aligned}$$

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

$$\begin{aligned} E[\hat{R}_1(\hat{F})|\chi_n] &= \frac{1}{n}\sum_{i=1}^n \hat{\delta}(x_i)d_0(x_i) - \beta_0, \\ \hat{R}_1(\hat{F}) - E[\hat{R}_1(\hat{F})|\chi_n] &= \frac{1}{n}\sum_{i=1}^n [\hat{\delta}(x_i) - \delta(x_i)][q_i - d_0(x_i)]. \end{aligned} \tag{4.5}$$

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 (1998), 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 Assumptions 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.

5 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 $E[\psi(z_i)] = 0$.

5.1 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 (4.4) 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 $\partial^\lambda f(z) = \partial^{|\lambda|} f(z)/\partial z_1^{\lambda_1} \cdot \cdots \partial z_r^{\lambda_r}$ 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 differentiable 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 $|\lambda| \leq s_\psi$, $|\lambda'| = s_\psi$, 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.

5.2 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 (4.5) 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 (4.5) satisfies

$$\frac{1}{n}\sum_{i=1}^n \hat{\delta}(x_i)d_0(x_i) = \Gamma^T\hat{\Sigma}^{-1}\sum_{i=1}^n p^K(x_i)[d_0(x_i) - p^K(x_i)^T\gamma]/n + E[\delta(x_i)\{p^K(x_i)^T\gamma - d_0(x_i)\}], \tag{5.6}$$

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

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

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

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

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.

6 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 $(\hat{M}^T \hat{W} \bar{M})^{-1} \hat{M}^T \hat{W} \xrightarrow{p} (M^T W M)^{-1} M^T W$. 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 and the space of function in which F lie is not too complex; 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 (3.2) 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 $\partial\hat{m}(\bar{\beta})/\partial\beta \xrightarrow{p} M$, $M^T W M$ 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 $-(M^T W M)^{-1} M^T W [m(z, \beta_0, F_0) + \varphi(z)]$.*

Alternatively, Assumption 7 can be used to show that the GMM estimator is asymptotically equivalent to the estimator studied in Section 4.

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.

7 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. We intend to take up elsewhere the use of the influence function in bias corrected semiparametric estimation. Shen (1995) considered optimal robust estimation among some types of semiparametric estimators. Further work on robustness of the kinds of estimators considered here may be possible. Other work on the influence function of semiparametric estimators may also be of interest.

8 Appendix A: Proofs

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

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

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 $\left[(1-t)f_0(\tilde{z}) + tg_z^h(\tilde{z})\right]^{1/2}$ 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 $g_z^h(\tilde{z}) \longrightarrow 0$ 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 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 $(z_1, ..., z_n)$ 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 $(z_1, ..., z_n)$ 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 $\int \|\psi(\tilde{z})\|^2 f_t(\tilde{z})d\tilde{z} = (1-t)\int \|\psi(\tilde{z})\|^2 f_t(\tilde{z})d\tilde{z} + t\int \|\psi(\tilde{z})\|^2 g_z^h(\tilde{z})d\tilde{z}$ 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 $(z_1, ..., z_n)$ 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,

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

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

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

so the Lindbergh-Feller condition for a central limit theorem is satisfied. Furthermore, it follows by similar calculations that $\int \breve{\psi}_n(\tilde{z})\breve{\psi}_n(\tilde{z})^T f_{t_n}(\tilde{z})d\tilde{z} \longrightarrow V$. 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

$$\sqrt{n}t_n\mu_z^h - \sqrt{n}(\beta_n - \beta_0) \longrightarrow 0. \tag{8.7}$$

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

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

by $\sqrt{n}t_n$ bounded. Next, we follow the proof of Theorem 2.1 of Van der Vaart (1991), and suppose that eq. (8.7) holds for all $t_n = O(1/\sqrt{n})$. 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 $t_n = r_m$ for $n = n_m$ and $t_n = n^{-1/2}$ for $n \notin \{n_1, n_2, ...\}$. By construction, $t_n = O(1/\sqrt{n})$, so that eq (8.7) 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 $\beta(t)$ is differentiable at $t = 0$ with derivative μ_z^h.

We have now shown that eq. (8.7) holds for all sequences $t_n = O(1/\sqrt{n})$ if and only if $\beta(t)$ is differentiable at $t = 0$ with derivative μ_z^h. Furthermore, as shown above eq. (8.7) 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 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: This follows as outlined in the text from Assumptions 1 and 2 and eq. (4.3) 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 $\left|\lambda'\right| \leq s_f$

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

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. (5.6). *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

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

9 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 X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica* 66, 289-314.

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," *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.

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 semiparametric 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.

Shen, L.Z. (1995): "On Optimal B-Robust Influence Functions in Semiparametric Models," *The Annals of Statistics* 23, 968–989.

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.