Ichimura, Hidehiko; Newey, Whitney K.

Working Paper

The influence function of semiparametric estimators

cemmap working paper, No. CWP06/17

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Ichimura, Hidehiko; Newey, Whitney K. (2017) : The influence function of semiparametric estimators, cemmap working paper, No. CWP06/17, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2017.0617

This Version is available at:
https://hdl.handle.net/10419/189692

cemmap
centre for microdata methods and practice

The influence function of semiparametric estimators

Hidehiko Ichimura
Whitney K. Newey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP06/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

The Influence Function of Semiparametric Estimators*

Hidehiko Ichimura
University of Tokyo

Whitney K. Newey
MIT

July 2015
Revised January 2017

Abstract

There are many economic parameters that depend on nonparametric first steps. Examples include games, dynamic discrete choice, average consumer surplus, and treatment effects. Often estimators of these parameters are asymptotically equivalent to a sample average of an object referred to as the influence function. The influence function is useful in formulating regularity conditions for asymptotic normality, for bias reduction, in efficiency comparisons, and for analyzing robustness. We show that the influence function of a semiparametric estimator is the limit of a Gateaux derivative with respect to a smooth deviation as the deviation approaches a point mass. This result generalizes the classic Von Mises (1947) and Hampel (1974) calculation to apply to estimators that depend on smooth nonparametic first steps. We characterize the influence function of M and GMM-estimators. We apply the Gateaux derivative to derive the influence function with a first step nonparametric two stage least squares estimator based on orthogonality conditions. We also use the influence function to analyze high level and primitive regularity conditions for asymptotic normality. We give primitive regularity conditions for linear functionals of series regression that are the weakest known, except for a log term, when the regression function is smooth enough.

JEL Classification: C13, C14, C20, C26, C36
Keywords: Influence function, semiparametric estimation, NPIV.

*arXiv:1508.01378. The JSPS 15H05692 provided financial support, as did the NSF via Grant SES 1132399. We are grateful for comments by X. Chen, V. Chernozhukov, K. Kato, U. Mueller, J. Porter, D. Pouzo, A. Santos and participants at seminars at UC Berkeley, NYU, University of Kansas, and Yale.

1 Introduction

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

We show that the influence function is the limit of the Gateaux derivative with respect to a smooth deviation, as the deviation approaches a point mass. That is, let $\beta(F)$ denote the limit of the estimator $\hat{\beta}$ when F is the cumulative distribution function (CDF) of a single observation. Also, let $F_\tau = (1-\tau)F_0 + \tau G$, where F_0 is the true distribution and G some alternative distribution. We find that the influence function is the limit of $d\beta(F_\tau)/d\tau|_{\tau=0}$ as G converges to the CDF of a point. This calculation is an extension of the Von Mises (1947) and Hampel (1974) formula for the influence function, which is the Gateaux derivative when G is the CDF of a constant. That classic calculation does not apply to many parameters that depend on densities, conditional expectations, or identification conditions. Choosing G so that these objects exist for F_τ, i.e. so that $\beta(F_\tau)$ is well defined, allows the Gateaux derivative to be calculated for many more estimators. For example, G can be chosen to be a continuous distribution when $\beta(F)$ depends on a density or conditional expectation. We characterize the influence function of M and GMM-estimators.

In an extended example we use the Gateaux derivative limit to derive the influence function of semiparametric estimators with a first step nonparametric two stage least squares (NP2SLS) estimator. Here the first step estimates a function that satisfies an infinite set of orthogonality conditions and may satisfy linear restrictions, such as additivity. We derive the influence function under correct specification and under misspecification. We apply this calculation to obtain the influence function in new cases, including the average derivative of a quantile instrumental variables estimator. In considering orthogonality conditions other than conditional

moment restrictions, nonlinear residuals, and linear restrictions on the first step our results go beyond those of Ai and Chen (2007, 2012), Santos (2011), and Severini and Tripathi (2012). Also, although asymptotic variances of some estimators have previously been derived for linear conditional moment restrictions by Ai and Chen (2007) and Santos (2011), the influence function has not been. As discussed above, the form of the influence function is useful for many purposes. Furthermore, it is not generally possible to uniquely derive the influence function from the asymptotic variance (e.g. for least squares in a linear model with homoskedasticity). Thus, the NP2SLS influence function formulae of this paper appear to be novel and useful even for the well studied model of conditional moment restrictions that are linear in an unknown function.

A characterization of the influence function as a solution to a functional equation is given in Newey (1994). That characterization has proven useful for finding the influence function of many important estimators, e.g. Newey (1994), Hahn (1998), Hirano, Imbens, and Ridder (2003), Bajari, Hong, Krainer, and Nekipelov (2010), Bajari, Chernozhukov, Hong, and Nekipelov (2009), Hahn and Ridder (2013, 2016), and Ackerberg, Chen, Hahn, and Liao (2015). The Gateaux derivative limit provides a way to calculate the solution to the functional equation, where the influence function emerges as the limit of the Gateaux derivative. In this way the Gateaux derivative limit provides a way to circumvent the need to try to guess the solution of the functional equation. We illustrate this calculation with examples, including NP2SLS. In addition, the mixture form of $F_\tau = (1-\tau)F_0 + \tau G$ can help in formulating important primitive conditions for existence of the influence function. For NP2SLS the form of F_τ allows us to show that $\beta(F_\tau)$ is well defined for $\tau \neq 0$, an essential condition. In this way NP2SLS provides a new example where the Gateaux derivative limit is essential for finding the influence function.

Knowing the influence function is useful for specifying regularity conditions, because the form of remainders is determined by the influence function. Regularity conditions that are sufficient for negligible remainders can then be specified. Here we give a remainder decomposition that leads to primitive conditions for asymptotic equivalence with the sample average of the influence function. The remainder decomposition has three terms, a stochastic equicontinuity term, a linearization term, and a linear functional of the first step. This decomposition is like that of Newey and McFadden (1994) except we linearize the effect of the first step on the expectation of the moments rather than the sample moments. We give conditions for the linear functional remainder that are more general in some respects than previously given by Andrews (1994), Newey (1994), Newey and McFadden (1994), Chen, Linton, and van Keilegom (2003), Bickel and Ritov (2003), and Ichimura and Lee (2010). These papers all build on pioneering work on the asymptotic distribution of specific estimators in Robinson (1988), Powell, Stock,

and Stoker (1989), Pakes and Pollard (1989), Ait-Sahalia (1991), Goldstein and Messer (1992), Ichimura (1993), Klein and Spady (1993), and Sherman (1993).

We give primitive regularity conditions for linear functionals of series nonparametric regression. These conditions apply to the linear remainder for any estimator with a first step series regression. We find that the stochastic size of the remainder is nearly as small as is known possible and the conditions on the number of series terms nearly as weak as possible when the regression function is smooth enough. These results have the same small bias structure as Newey (1994) but use recent results of Belloni, Chernozhukov, Chetverikov, and Kato (2015) to obtain weak rate conditions for root-n consistency.

After the first version of this paper appeared on ArXiv, Luedtke, Carone, and van der Laan (2015) and Carone, Luedtke, and van der Laan (2016) used the idea of smoothing the CDF of an indicator in constructing efficient estimators. This idea is clearly useful in that setting but we emphasize that we have a different goal, that is calculating the influence function of any semiparametric estimator.

Summarizing, the contributions of this paper are to i) introduce a Gateaux derivative limit formula for the influence function; ii) show that the formula is important for finding the influence function for a NP2SLS first step; iii) give a general remainder decomposition that imposes weaker regularity conditions than previously; iv) give primitive, weak regularity conditions for linear functionals of a first step series regression.

In Section 2 we describe the estimators we consider. Section 3 gives the Gateaux derivative limit formula and conditions for its validity. In Section 4 we derive the influence function when the first step is the NP2SLS estimator for linear, nonparametric conditional moment restrictions. Section 5 extends the results of Section 4 to orthogonality conditions, misspecification, nonlinear residuals, and a restricted first step. Section 6 gives a remainder decomposition and the series regression results. Section 7 concludes.

2 Semiparametric Estimators

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

The Gateaux derivative limit formula applies generally to asymptotically linear estimators. An asymptotically linear estimator is one satisfying

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

The asymptotic variance of $\hat{\beta}$ is then $E[\psi(z_i)\psi(z_i)^T]$. The function $\psi(z)$ is referred to as the influence function, following terminology of Hampel (1974). It gives the influence of a single observation in the leading term of the expansion in equation (2.1). It also quantifies the effect of a small change in the distribution on the limit of $\hat{\beta}$ as we further explain below. Nearly all root-n consistent semiparametric estimators we are aware of are asymptotically linear under sufficient regularity conditions, including M-estimators, Z-estimators, estimators based on U-statistics, and many others; see Bickel, Klaasen, Ritov, and Wellner (1993).

Many estimators have structure that facilitates derivation of the influence function. One type of estimator with such structure, that includes most asymptotically linear estimators, is a maximization (M) estimator where

$$\hat{\beta} = \arg\max_{\beta \in B} \hat{Q}(\beta),$$

for a function $\hat{Q}(\beta)$ that depends on the data and parameters. M estimators have long been studied. A more general type that is useful when $\hat{Q}(\beta)$ is not continuous has $\hat{Q}(\hat{\beta}) \geq \sup_{\beta \in B} \hat{Q}(\beta) - \hat{R}$, where the remainder $\hat{R}$ is small in large samples. The influence function will depend only on the limit of the objective function and so is not affected by whether $\hat{\beta}$ is an approximate or exact maximizer of $\hat{Q}(\beta)$.

Additional useful structure is present for generalized method of moments (GMM; Hansen, 1982) estimators where the moment functions depend on a first step nonparametric estimator. To describe this type of estimator let $m(z, \beta, \gamma)$ denote a vector of functions of the data observation z, β, and a function γ that may be vector valued. Here γ represents some first step function, i.e. a possible value of a nonparametric estimator. GMM is based on a moment condition $E[m(z_i, \beta, \gamma_0)] = 0$, assumed to be locally uniquely solved at $\beta = \beta_0$. Let $\hat{\gamma}_i$ denote some nonparametric estimator of γ_0, where $\hat{\gamma}_i$ can depend on i to allow sample splitting, such as a leave one out estimator, that is known to have good properties in some settings. Plugging in $\hat{\gamma}_i$ to obtain $m(z_i, \beta, \hat{\gamma}_i)$ and averaging over i gives the estimated sample moments $\hat{m}(\beta) = \sum_{i=1}^{n} m(z_i, \beta, \hat{\gamma}_i)/n$. For $\hat{W}$ a positive semi-definite weighting matrix a semiparametric GMM estimator is

$$\hat{\beta} = \arg\max_{\beta \in B} -\hat{m}(\beta)^T \hat{W} \hat{m}(\beta)/2.$$

A GMM estimator is an M-estimator with $\hat{Q}(\beta) = -\hat{m}(\beta)^T \hat{W} \hat{m}(\beta)/2$. We note that this type of estimator 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, e.g. see Newey (1994) and Chen and Liao (2015).

An M-estimator will be asymptotically linear under certain conditions. Suppose that $\hat{Q}(\beta)$ converges in probability to $Q(\beta)$ that is twice differentiable at β and let $H = \partial^2 Q(\beta_0)/\partial\beta\partial\beta^T$ be the Hessian of $Q(\beta)$ at β_0. Then the influence function will be

$$\psi(z) = -H^{-1}\xi(z),$$

where $\xi(z)$ will be described below. The GMM-estimator will be asymptotically linear with $\xi(z)$ having a specific form. Let

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

The limit of the GMM objective function will be $Q(\beta) = -E[m(z_i,\beta,\gamma_0)]^T W E[m(z_i,\beta,\gamma_0)]/2$. Under correct specification where $E[m(z_i,\beta_0,\gamma_0)] = 0$ the chain rule gives $H = -M^T W M$. Also, for GMM it will turn out that $\xi(z) = M^T W[m(z,\beta_0,\gamma_0) + \phi(z)]$ where $\phi(z)$ is an adjustment term for the estimator $\hat{\gamma}$ of γ_0. The influence function for GMM is then

$$\psi(z) = -(M^T W M)^{-1} M^T W[m(z,\beta_0,\gamma_0) + \phi(z)]. \tag{2.2}$$

Here $m(z,\beta_0,\gamma_0) + \phi(z)$ will be the influence function of $\hat{m}(\beta_0)$ and the adjustment term $\phi(z)$ will be the influence function of $\int m(z,\beta_0,\hat{\gamma})F_0(dz)$. These formulae for the influence function are valid under regularity conditions, that allow for $m(z,\beta,\gamma_0)$ to not be smooth in β, e.g. as in Chen, Linton, and van Keilegom, (2003) and Ichimura and Lee (2010).

One specific example is a bound on the average surplus (integrated over heterogeneity) of a price change when there are bounds on income effects, as in Hausman and Newey (2016a,b). Let y denote quantity consumed of some good, $x = (x_1, x_2)^T$ where x_1 is price, x_2 is income, $\omega_2(x_2)$ be a weight function for income (such as an indicator for some interval divided by the length of the interval), and $\gamma(x)$ a possible conditional expectation function $E[y_i|x_i = x]$. We assume that B is a uniform bound on the derivative of demand with respect to income, i.e. the income effect. We consider a price change from $\breve{x}_1$ to $\bar{x}_1$. Let $\omega_1(x_1) = 1(\breve{x}_1 \leq x_1 \leq \bar{x}_1)\exp(-B[x_1 - \breve{x}_1])$. The object of interest is a bound on the weighted average over income of average equivalent variation for a price change from $\breve{x}_1$ to $\bar{x}_1$ given by

$$\beta_0 = \int \omega(x)\gamma_0(x)dx, \omega(x) = \omega_2(x_2)\omega_1(x_1).$$

If B is an upper (lower) bound on income effects then β_0 is a lower (upper) bound on average equivalent variation over income and individual heterogeneity of a price change from $\breve{x}_1$ to $\bar{x}_1$.

This object is identified from the semiparametric moment function

$$m(z, \beta, \gamma) = \int \omega(x)\gamma_0(x)dx - \beta.$$

We will derive the influence function for this and other objects below.

3 Calculating the Influence Function

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

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

For M-estimators, $\beta(F) = \arg\max_{\beta \in B} Q(\beta, F)$ where $Q(\beta, F)$ is the probability limit of $\hat{Q}(\beta)$ when z_i has CDF F, under well known regularity conditions that allow interchange of the limit and argmax operations. In the surplus bound example

$$\beta(F) = \int \omega(\tilde{x})E_F[y_i|x_i = \tilde{x}]d\tilde{x}, \tag{3.3}$$

where $E_F[y_i|x_i]$ denotes the conditional expectation under distribution F.

An important feature of $\beta(F)$ is that it may only be well defined when F is restricted in some way. In the average surplus example $\beta(F)$ will be well defined when $E[y_i|x_i]$ is well defined where $\omega(x_i) > 0$. In formal terms this feature means that the domain $\mathcal{F}$ of $\beta(\cdot)$ is restricted. To allow for a restricted domain we consider only variations in F that are contained in $\mathcal{F}$. The specific kind of variation we consider is a convex combination $F_\tau = (1 - \tau)F_0 + \tau G_z^j$ of the true distribution F_0 with some other function G_z^j where $F_\tau \in \mathcal{F}$ for all small enough τ. The superscript j and subscript z designate G_z^j as a member of sequence of functions approaching the CDF of the constant z. Under regularity conditions given below the influence function can be calculated as

$$\psi(z) = \lim_{j \longrightarrow \infty} \left[\frac{d}{d\tau} \beta((1 - \tau) \cdot F_0 + \tau \cdot G_z^j) \right], \tag{3.4}$$

where all derivatives with respect to τ are right derivatives at $\tau = 0$ unless otherwise stated. The derivative in this expression is the Gateaux derivative of the functional $\beta(F)$ with respect

to a deviation $G_z^j - F_0(z)$ from the true distribution F_0. This formula says that the influence function is the limit of this Gateaux derivative as G_z^j approaches the CDF of the constant z.

Equation (3.4) can be thought of as a generalization of the influence function calculation of Von Mises (1947) and Hampel (1974). That calculation is based on $G_z^j = \lambda_z$ where λ_z is the CDF of the constant z. If $(1-\tau)\cdot F_0 + \tau \cdot \lambda_z$ is in $\mathcal{F}$ then the influence function is given by the Gateaux derivative

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

The problem with this formula is that $F_\tau = (1-\tau)\cdot F_0 + \tau \cdot \lambda_z$ will not be in the domain $\mathcal{F}$ for many semiparametric estimators. In many cases $F \in \mathcal{F}$ (i.e. $\beta(F)$ being well defined) requires that certain marginal distributions of F are continuous. The CDF $(1-\tau)\cdot F_0 + \tau \cdot \lambda_z$ does not satisfy that restriction. Equation (3.4) circumvents this problem by allowing us to restrict F_τ to be in $\mathcal{F}$ through a choice of G_z^j. The influence function is then obtained as the limit of a Gateaux derivative as $G_z^j \longrightarrow \lambda_z$ rather than the Gateaux derivative with respect to δ_z. This generalization applies to most semiparametric estimators.

We can relate equation (3.4) to the pathwise derivative characterization of the influence function in Newey (1994). Denote a parametric model as F_θ, where θ denotes a vector of parameters, with $F_\theta \in \mathcal{F}$ equal to the true distribution F_0 at $\theta = 0$. Impose that each parametric model is regular in the sense used in the efficiency bounds literature, so that F_θ has a score $S(z)$ (derivative of the log-likelihood in many cases, e.g. see Van der Vaart, 1998, p. 362) at $\theta = 0$ and possibly other conditions are satisfied. Suppose that the set of scores over all the regular parametric families has mean square closure that includes all functions with mean zero and finite variance. This assumption is the precise meaning of the statement that we are not restricting F except for regularity conditions. As shown by Newey (1994) the influence function $\psi(z)$ is then the unique solution to the functional equation of Van der Vaart (1991),

$$\frac{\partial \beta(F_\theta)}{\partial \theta} = E[\psi(z_i)S(z_i)^T], \tag{3.5}$$

as the score $S(z)$ varies over those for regular parametric models. This is a functional equation that the influence function uniquely solves. In many cases $\partial\beta(F_\theta)/\partial\theta$ will be a linear functional of the score. Root-n consistent estimation will not be possible unless $\partial\beta(F_\theta)/\partial\theta$ is a mean square continuous functional of $S(z_i)$, in which case the Riesz representation theorem will imply existence of $\psi(z_i)$ satisfying equation (3.5). Thus $\psi(z_i)$ can be thought of as the unique function that gives a Riesz representation.

Equation (3.4) provides a way to calculate the solution to equation (3.5). Consider F_τ as a parametric submodel with parameter $\theta = \tau$. Suppose that F and G_z^j are both absolutely continuous with respect to a measure μ with pdf f_0 and derivative g_z^j respectively such that

$f_\tau = (1-\tau)f_0 + \tau g_z^j$ is nonnegative for all small enough τ. The score for F_τ will be $S_j(z_i) = \partial \ln[(1-\tau)f_0(z_i)+\tau g_z^j(z_i)]/\partial\tau = g_z^j(z_i)/f_0(z_i)-1$, ignoring regularity conditions for the moment. For this score equation (3.5) is

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

If $\psi(\tilde{z})$ is continuous at z then $\partial\beta(F_\tau)/\partial\tau$ will approach $\psi(z)$ as $g_z^j(\tilde{z})$ approaches a spike at z. We illustrate this calculation in examples below as well as give regularity conditions for its validity.

This calculation allows us to avoid figuring out the solution the solution to the functional equation (3.5). For instance, there is no need to find the Riesz representations in Proposition 4 or 5 of Newey (1994). Instead, the influence function $\psi(z)$ emerges from the Gateaux derivative limit calculation. The form of F_τ also turns out to be useful for guaranteeing that $\beta(F_\tau)$ exists for NP2SLS. These advantages of equation (3.4) are highlighted in the examples to follow.

Another way to calculate the influence function is as the limit of a projection. Consider a parametric submodel of the form $f_\theta(z) = f_0(z)\left[1 + p(z)^T\theta\right]$ where $p(z)$ is a vector of bounded approximating functions, such as splines or wavelets, normalized to have $E[p(z_i)] = 0$. For this likelihood the score vector is $S(z) = p(z)$. Then by equation (3.5)

$$\psi_p(z) = \frac{\partial\beta(F_\theta)}{\partial\theta}(E[p(z_i)p(z_i)^T])^{-1}p(z)$$

is the population least squares projection of $\psi(z)$ on $p(z)$. One could calculate $\psi(z)$ as the mean square limit of $\psi_p(z)$ as the dimension of $p(z)$ grows so that linear combinations of $p(z)$ can approximate any function. Variants of the related approach to finding the asymptotic variance by taking the limit of

$$E[\psi_p(z_i)\psi_p(z_i)^T] = \frac{\partial\beta(F_\theta)}{\partial\theta}(E[p(z_i)p(z_i)^T])^{-1}\frac{\partial\beta(F_\theta)^T}{\partial\theta}$$

have proven useful in several settings, see Ai and Chen (2003, 2007) and Chen and Liao (2015). We could use this series expansion to compute the influence function but it is not an explicit calculation and the limit may be difficult to compute in general. In comparison, the Gateaux derivative approach just requires taking the limit of a derivative. Also, the Gateaux derivative calculation does not rely on finding a natural choice of $p(z)$ or any other natural parameterization of the limit of the estimator.

We can use the Gateaux derivative limit to characterize the influence function for semiparametric M-estimators. Let $Q(\beta, F_\tau)$ denote the limit of the objective function $\hat{Q}(\beta)$ when the CDF of z_i is F_τ. Then under standard regularity conditions $\hat{\beta}$ will converge to

$$\beta(F_\tau) = \arg\max_{\beta\in B} Q(\beta, F_\tau).$$

For notational convenience let $\beta_\tau = \beta(F_\tau)$. Suppose that $Q(\beta, F_\tau)$ is twice continuously differentiable in β and τ for each j and β_τ is in the interior of the parameter set. Then β_τ satisfies the first order conditions $\partial Q(\beta_\tau, F_\tau)/\partial\beta = 0$. It then follows from the implicit function theorem that $d\beta_\tau/d\tau = -H^{-1}\partial^2 Q(\beta_0, F_\tau)/\partial\tau\partial\beta|_{\tau=0}$, so that

$$\lim_{j\longrightarrow\infty}\left[\frac{d}{d\tau}\beta_\tau\right] = -H^{-1}\xi(z), \xi(z) = \lim_{j\longrightarrow\infty}\left[\partial^2 Q(\beta_0, F_\tau)/\partial\tau\partial\beta|_{\tau=0}\right].$$

Here we see that the function $\xi(z)$ will be the limit of the cross derivative of $Q(\beta, F_\tau)$ with respect to β and τ.

We can apply this characterization to derive the influence function for GMM. Let $\gamma_\tau = \gamma(F_\tau)$ and W_τ denote the limits of the first step estimator $\hat{\gamma}$ and the weighting matrix, respectively, when z_i has CDF F_τ. Also, let $E_\tau[\cdot]$ denote the expectation with respect to F_τ. For GMM $Q(\beta, F_\tau) = -E_\tau[m(z_i, \beta, \gamma_\tau)]^T W_\tau E_\tau[m(z_i, \beta, \gamma_\tau)]/2$, so that

$$\partial Q(\beta, F_\tau)/\partial\beta = -\partial E_\tau[m(z_i, \beta, \gamma_\tau)^T]/\partial\beta W_\tau E_\tau[m(z_i, \beta, \gamma_\tau)].$$

Then under correct specification of the moments where $E[m(z_i, \beta_0, \gamma_0)] = 0$ the chain rule gives

$$\begin{aligned}
\lim_{j\longrightarrow\infty}\left[\frac{d}{d\tau}\beta_\tau\right] &= -(M^T W M)^{-1} M^T W \lim_{j\longrightarrow\infty}\left[\frac{d}{d\tau}E_\tau[m(z_i, \beta_0, \gamma_\tau)]\right] \\
&= -(M^T W M)^{-1} M^T W \lim_{j\longrightarrow\infty}\left\{\frac{d}{d\tau}E_\tau[m(z_i, \beta_0, \gamma_0)] + \frac{d}{d\tau}E[m(z_i, \beta_0, \gamma_\tau)]\right\} \\
&= -(M^T W M)^{-1} M^T W \lim_{j\longrightarrow\infty}\left\{\int m(\tilde{z}, \beta_0, \gamma_0) G_z^j(d\tilde{z}) + \frac{d}{d\tau}E[m(z_i, \beta_0, \gamma_\tau)]\right\} \\
&= -(M^T W M)^{-1} M^T W \left\{m(z, \beta_0, \gamma_0) + \phi(z)\right\},
\end{aligned}$$

where the second equality follows by the chain rule, the third by the definition of F_τ, and the fourth equality will hold under continuity of $m(z, \beta_0, \gamma_0)$ in z with

$$\phi(z) = \lim_{j\longrightarrow\infty}\left\{\frac{d}{d\tau}E[m(z_i, \beta_0, \gamma_\tau)]\right\}. \tag{3.6}$$

Here we see by $\gamma_\tau = \gamma(F_\tau)$ that $\phi(z)$ is the influence function of $\int m(z, \beta_0, \hat{\gamma})F_0(dz)$. This $\phi(z)$ is the adjustment term that accounts for the presence of the first step estimator $\hat{\gamma}$ in the moment conditions as discussed in Newey (1994). It can be calculated as the Gateaux derivative limit in equation (3.6).

For M-estimators, certain nonparametric components of $\hat{Q}(\beta)$ can be ignored in deriving the influence function. The ignorable components are those that have been "concentrated out," meaning they have a limit that maximizes the limit of $\hat{Q}(\beta)$. In such cases the dependence of these functions on β captures the whole asymptotic effect of their estimation. To show this

result, suppose that there is a function γ of β and possibly other variables and a function $\tilde{Q}(\beta, \gamma, F)$ such that $Q(\beta, F_\tau) = \tilde{Q}(\beta, \gamma_\tau, F_\tau)$ where

$$\gamma_\tau = \arg\max_\gamma \tilde{Q}(\beta, \gamma, F_\tau).$$

Here $\tilde{Q}(\beta, \gamma_\tau, F_\tau)$ is the limit of $\hat{Q}(\beta)$ and γ_τ the limit of a nonparametric estimator on which $\hat{Q}(\beta)$ depends, when z_i has CDF F_τ. Since γ_τ maximizes over all γ it must maximize over function $\gamma_{\tilde{\tau}}$ as $\tilde{\tau}$ varies. The first order condition for maximization over $\tilde{\tau}$ is

$$\left.\frac{\partial \tilde{Q}(\beta, \gamma_{\tilde{\tau}}, F_\tau)}{\partial \tilde{\tau}}\right|_{\tilde{\tau}=\tau} = 0.$$

This equation holds identically in β, so that we can differentiate both sides of the equality with respect to β, evaluate at $\beta = \beta_0$ and $\tau = 0$, and interchange the order of differentiation to obtain

$$\frac{\partial^2 \tilde{Q}(\beta_0, \gamma_\tau, F_0)}{\partial \tau \partial \beta} = 0.$$

Then it follows by the chain rule that

$$\xi(z) = \lim_{j \longrightarrow \infty} \left[\partial^2 \tilde{Q}(\beta_0, \gamma_\tau, F_\tau)/\partial\tau\partial\beta|_{\tau=0}\right] = \lim_{j \longrightarrow \infty} \left[\partial^2 \tilde{Q}(\beta_0, \gamma_0, F_\tau)/\partial\tau\partial\beta|_{\tau=0}\right]. \tag{3.7}$$

That is, the function $\xi(z)$ can be obtained setting $\gamma_\tau = \gamma_0$ and differentiating $\tilde{Q}(\beta_0, \gamma_0, F_\tau)/\partial\beta$ with respect τ. In this calculation we are treating the limit γ_τ as if it is equal the true value γ_0.

Equation (3.7) generalizes Proposition 2 of Newey (1994) and Theorem 3.4 of Ichimura and Lee (2010) to objective functions that are not necessarily a sample average of a function of β and γ. There are many important estimators included in this generalization. One of those is NP2SLS where the residual includes both parametric and nonparametric components. The result implies that estimation of the function of β that is the nonparametric component γ can be ignored in calculating the influence function of β. Another interesting estimator is partially linear regression with generated regressors. There the estimation of the nonparametric component can be ignored in deriving the the influence function, just as in Robinson (1988), though the presence of generated regressors will often affect the influence function, as in Hahn and Ridder (2013, 2016) and Mammen, Rothe, and Schienle (2012).

To use the Gateaux derivative formula to calculate the influence function we need to specify G_z^j. Various kinds of restrictions on G_z^j may be needed to insure that $F_\tau \in \mathcal{F}$. In the surplus bound example $E[y_i|x_i]$ must be well defined where $\omega(x_i) > 0$. In other examples an identification condition may need to be satisfied. We are free to choose G_z^j in whatever way is convenient for imposing these restrictions and ensuring that equation (3.4) holds. A particularly convenient form of G_z^j is

$$G_z^j(\tilde{z}) = E[1(z_i \leq \tilde{z})\delta(z_i)], \tag{3.8}$$

where $\delta(z_i)$ is a bounded function with $E[\delta(z_i)] = 1$. The variable $\tilde{z}$ represents a possible value of the random variable z_i, and we suppress a j superscript and z subscript on $\delta(z_i)$ for notational convenience. This $G_z^j(\tilde{z})$ will approach the CDF of the constant $\tilde{z}$ as $\delta(z)f_0(z)$ approaches a spike at $\tilde{z}$. The boundedness of $\delta(z)$ makes this distribution convenient, as further discussed below. Also F_τ will have other useful properties.

We will assume that $z \in \mathbb{R}^r$ and that F_0 is absolutely continuous with respect to a product measure μ on $\mathbb{R}^r$ with pdf $f_0(z)$. This assumption allows for individual components of z_i to be continuously or discretely distributed, or some mixture of the two. By $\delta(\tilde{z})$ bounded $F_\tau = (1-\tau)F_0 + \tau G_z^j$ will be a CDF for small enough τ with pdf with respect to μ given by

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

Note that by $S(\tilde{z})$ bounded there is C such that for small enough τ,

$$(1-\tau)f_0/C \leq f_\tau \leq Cf_0, \tag{3.9}$$

so that f_τ and f_0 will be absolutely continuous with respect to each other. Also, for any measurable function y_i and components x_i of z_i the marginal pdf $f_\tau(\tilde{x})$ of x_i, conditional expectation $E_\tau[y_i|x_i]$ of y_i given x_i, and its derivative with respect to τ are

$$\begin{aligned} f_\tau(\tilde{x}) &= f_0(\tilde{x})\{1 + \tau E[S(z_i)|x_i = \tilde{x}]\}, \\ E_\tau[y_i|x_i] &= \frac{E[y_i|x_i] + \tau E[y_i S(z_i)|x_i]}{1 + \tau E[S(z_i)|x_i]}, \frac{\partial E_\tau[y_i|x_i]}{\partial \tau} = E[\{y_i - E[y_i|x_i]\}\delta(z_i)|x_i], \end{aligned} \tag{3.10}$$

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

We are free to choose $\delta(z)$ in whatever way is convenient for the problem at hand. A particular choice of $\delta(z)$ that is useful in several cases is a ratio of a sharply peaked pdf to the true density. To specify such a $\delta(z)$ let $K(u)$ be a pdf that is symmetric around zero, has bounded support, and is continuously differentiable of all orders with bounded derivatives. Also let $\bar{\mu}_\ell^j = j \int K((z_\ell - \tilde{z}_\ell)j) d\mu_\ell(\tilde{z}_\ell)$. We will assume that $\mu(\mathcal{N}) > 0$ for any open set $\mathcal{N}$ containing z, so that $\bar{\mu}_\ell^j > 0$ for every ℓ and j. Let

$$g(\tilde{z}) = \Pi_{\ell=1}^r \kappa_\ell^j(\tilde{z}_\ell), \kappa_\ell^j(\tilde{z}_\ell) = \left(\bar{\mu}_\ell^j\right)^{-1} jK((z_\ell - \tilde{z}_\ell)j). \tag{3.11}$$

A corresponding $\delta(\tilde{z})$ is

$$\delta(\tilde{z}) = g(\tilde{z})1(f_0(\tilde{z}) \geq 1/j)f_0(\tilde{z})^{-1}. \tag{3.12}$$

For this choice of $\delta(\tilde{z})$ equation, we will have $E[\delta(z_i)] = 1$ for large enough j and equation (3.4) will hold when $\psi(\tilde{z})$ is continuous at z and $f_0(\tilde{z})$ is bounded away from zero on a set of $\tilde{z}$ that has full μ measure locally to z, as shown and further discussed below.

We can calculate the influence function for the surplus bound using this $\delta(z)$. We assume that the joint pdf $f_0(\tilde{z})$ of $z_i = (y_i, x_i)$ is bounded away from zero on a neighborhood of z, so for large enough j, $f_0(\tilde{z}) \geq 1/j$ on the set where $g(\tilde{z})$ is nonzero, and hence $\delta(z) = g(z)/f_0(\tilde{z})$, for large enough j. In this case the formula for the derivative of the conditional expectation in equation (3.10) gives

$$\begin{aligned}\frac{\partial \beta(F_\tau)}{\partial \tau} &= \frac{\partial}{\partial \tau}\int \omega(x) E_\tau[y_i|x_i = x]dx = \int \omega(x)\frac{\partial}{\partial \tau}E_\tau[y_i|x_i = x]dx \\ &= \int \omega(x)E[\{y_i - \gamma_0(x)\}\frac{g(y_i,x)}{f_0(y_i,x)}|x_i = x]dx = \int \frac{\omega(x)}{f_0(x)}[\int \{y - \gamma_0(x)\}g(y,x)dy]dx \\ &= \int \alpha(x)(y - \gamma_0(x))g(y,x)dydx,\ \alpha(x) = \omega(x)/f_0(x).\end{aligned}$$

Assume that $\gamma_0(\tilde{x})$, $\omega_1(\tilde{x}_1)$, $\omega_2(\tilde{x}_2)$, and $1/f_0(\tilde{x})$ are continuous at x, so that $\alpha(\tilde{x})[\tilde{y} - \gamma_0(\tilde{x})]$ is continuous at z. Then as $j \longrightarrow \infty$ we have

$$\frac{d}{d\tau}\beta(F_\tau) \longrightarrow \alpha(x)[y - \gamma_0(x)]. \tag{3.13}$$

We can also characterize the influence function in this example using Proposition 4 of Newey (1994). To do so we need to find the solution to the Riesz representation in equation (4.4) of Newey (1994). Multiplying and dividing by the marginal density $f_0(x)$ gives

$$\int \omega(x)\gamma(x)dx = E[\alpha(x_i)\gamma(x_i)], \alpha(x) = \omega(x)/f_0(x).$$

Here $\alpha(x) = f_0(x)^{-1}\omega(x)$ gives the Riesz representation in Proposition 4 of Newey (1994), so the influence function of the surplus bound is in equation (3.13).

This example shows how the Gateaux derivative formula in equation (3.4) is a direct calculation of the influence function. At no point in the Gateaux derivative calculation did we need to solve for $\alpha(x)$. Instead the expression for $\alpha(x)$ emerged from the derivative calculation. In contrast, to apply Proposition 4 of Newey (1994) we had to find the solution to a Riesz representation. We will show that the Gateaux derivative is similarly useful in the even more challenging and novel example of NP2SLS.

We give a precise theoretical justification for the formula in equation (3.4) by assuming that an estimator is asymptotically linear and then showing that equation (3.4) is satisfied under a few mild regularity conditions. One of the regularity conditions we use is local regularity of $\hat{\beta}$ along the path F_τ.

DEFINITION 1: *$\hat{\beta}$ is locally regular for F_τ if there is a fixed random variable Y such that for any $\tau_n = O(1/\sqrt{n})$ and $z_1, ..., z_n$ i.i.d. with distribution F_{τ_n},*

$$\sqrt{n}[\hat{\beta} - \beta(F_{\tau_n})] \xrightarrow{d} Y.$$

Local regularity means that $\sqrt{n}[\hat{\beta} - \beta(F_{\tau_n})]$ has the same limiting distribution under a sequence of local alternatives as it has when $\tau_n = 0$ for all n, i.e. at F_0. Recall that the pdf of F_τ will be $f_\tau(z) = f_0(z)[1 + \tau S(z)]$ for bounded $S(z)$. We expect that such a deviation is so well behaved that $\hat{\beta}$ will be locally regular in most settings under sufficient, more primitive regularity conditions. For these reasons we view local regularity as a mild condition for the influence function calculation.

The next result shows that the influence function formula (3.4) is valid for G_z^j as specified in equation (3.8). It would be straightforward to extend this validity result to more general classes of G_z^j but the result we give should suffice for most cases.

THEOREM 1: *Suppose that i) $\hat{\beta}$ is asymptotically linear with influence function $\psi(z_i)$ and for each j is locally regular for the parametric model $(1-\tau)F_0+\tau G_z^j$; ii) μ is a product measure; iii) $\mu(\mathcal{N}) > 0$ for any open set containing z; iv) there is an open set $\mathcal{N}$ containing z and a subset $\bar{\mathcal{N}}$ of $\mathcal{N}$ such that a) $\mu(\mathcal{N}) = \mu(\bar{\mathcal{N}})$, b) $\psi(\tilde{z})$ is continuous at z for $\tilde{z} \in \bar{\mathcal{N}}$; c) there is $\varepsilon > 0$ such that $\mu(\bar{\mathcal{N}} \cap \{z : f_0(z) \geq \varepsilon\}) = \mu(\bar{\mathcal{N}})$. Then $\partial\beta(F_\tau)/\partial\tau$ exists for j large enough and satisfies equation (3.4).*

The proof of Theorem 1 is given in the Appendix.

We want to emphasize that the purpose of Theorem 1 is quite different than Van der Vaart (1991, 1998) and other important contributions to the semiparametric efficiency literature. Here $\beta(F)$ is not a parameter of some semiparametric model. Instead $\beta(F)$ is associated with an *estimator* $\hat{\beta}$, being the limit of that estimator when F is a distribution that is unrestricted except for regularity conditions, as formulated in Newey (1994). Our goal is to use $\beta(F)$ to calculate the influence function of $\hat{\beta}$ under the assumption that $\hat{\beta}$ is asymptotically linear. The purpose of Theorem 1 is to justify this calculation via equation (3.4). In contrast, the goal of the semiparametric efficiency literature is to find the efficient influence function for a parameter of interest when F belongs to a family of distributions.

To highlight this contrast, note that the Gateaux derivative limit calculation can be applied to obtain the influence function under misspecification while efficient influence function calculations generally impose correct specification. Indeed, the definition of $\beta(F)$ *requires* that misspecification be allowed for, because $\beta(F)$ is limit of the estimator $\hat{\beta}$ under all distributions F that are unrestricted except for regularity condition. Of course correct specification may lead to simplifications in the form of the influence function. Such simplifications will be incorporated automatically when the Gateaux derivative limit is taken at an F_0 that satisfies model restrictions.

Theorem 1 shows that if an estimator is asymptotically linear and locally regular then the

influence function satisfies equation (*3.4*), justifying that calculation. This result is like Van der Vaart (1991) in having differentiability of $\beta(F_\tau)$ as a conclusion. It differs in restricting the paths to have the form $(1-\tau)F_0 + \tau G_z^j$. Such a restriction on the paths actually weakens the local regularity hypothesis because $\hat{\beta}$ only has to be locally regular for a particular kind of path rather than the general class of paths in Van der Vaart (1991). The conditions of Theorem 1 are stronger than Van der Vaart (1991) in assuming that the influence function is continuous at z and that the pdf of z_i is bounded away from zero on a neighborhood of z. We view these as weak restrictions that will be satisfied almost everywhere with respect to the dominating measure μ in many cases. We also note that this result allows for distributions to have a discrete component because the dominating measure μ may have atoms.

The weak nature of the local regularity condition highlights the strength of the asymptotic linearity hypothesis. Primitive conditions for asymptotic linearity can be quite strong and complicated. For example, it is known that asymptotic linearity of estimators with a nonparametric first step generally requires some degree of smoothness in the functions being estimated, see Ritov and Bickel (1990). Our purpose here is to bypass those conditions in order to justify the Gateaux derivative formula for the influence function. The formula for the influence function can then be used in all the important ways outlined in the introduction, including as a starting point for formulating more primitive conditions for asymptotic linearity, which we do below.

4 Nonparametric Two Stage Least Squares

In this Section we derive the influence function for a semiparametric GMM estimator where the first step γ_0 is the nonparametric two stage least squares estimator (NP2SLS) of Newey and Powell (1989, 2003) and Newey (1991), abbreviated NP henceforth. Ai and Chen (2003) considered a semiparametric version of this estimator, giving conditions for root-n consistency and asymptotic normality of the finite dimensional component. As with consumer surplus, the form of the influence function emerges from the calculation of the derivative. Also, we show that the limit $\gamma(F_\tau)$ of the NP2SLS estimator exists and is unique for $\tau \neq 0$ as is essential for calculation of the influence function. The uniqueness and existence result depends on the specification $F_\tau = (1-\tau)F_0 + \tau G_z^j$. In this way the approach of this paper is important for a key hypothesis that the limit of the NP2SLS exists and is unique.

We begin with a first step that is based on a linear, nonparametric, instrumental variables model in NP where

$$y_i = \gamma_0(w_i) + \varepsilon_i, E[\varepsilon_i | x_i] = 0, \tag{4.14}$$

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

are instrumental variables. We also first consider the case where the conditional expectations $E[\Delta(w_i)|x_i]$ and $E[\tilde{\Delta}(x_i)|w_i]$ are both complete as a function of Δ and $\tilde{\Delta}$ respectively. The identification condition for $\gamma_0(w_i)$ in this model is completeness of $E[\Delta(w_i)|x_i]$ and the other completeness condition is important for the influence function calculations. If x_i and w_i are continuously distributed with the same dimension and support equal to a rectangle then these completeness conditions hold generically, as shown by Andrews (2011) and Chen, Chernozhukov, Lee, and Newey (2014). Genericity justifies our assumption of completeness, although as with other important generic conditions (e.g. existence of moments), completeness cannot be tested (see Canay, Santos, Shaikh, 2013).

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

$$\gamma_\tau = \arg\min_{\gamma \in \Gamma} Q_\tau(\gamma) = \arg\min_{\gamma \in \Gamma} E_\tau[\{y_i - E_\tau[\gamma(w_i)|x_i]\}^2].$$

A problem with this calculation is that we do not know if γ_τ exists or is unique when Γ is unrestricted. This problem occurs because γ appears inside a conditional expectation. Our framework helps. The use of the F_τ we are working with allows us to specify G_z^j in such a way that γ_τ exists and is unique when τ is small enough. For this purpose we modify our choice of $\delta(z)$. Let $g(y, w)$ be as specified in equation (3.11) except that the product is only taken over components of (y, w). Let $\Delta_j(w_i)$ be a bounded function with $E[\Delta_j(w_i)] \neq 0$ and other properties discussed below. We then choose $\delta(z)$ to be

$$\delta(\tilde{z}) = 1(f_0(\tilde{y}, \tilde{w}|x) \geq 1/j) \left[f_0(\tilde{y}, \tilde{w}|\tilde{x})\right]^{-1} g(\tilde{y}, \tilde{w}) \delta_x(\tilde{x}), \delta_x(\tilde{x}) = E[\Delta_j(w_i)|x_i = \tilde{x}]/E[\Delta_j(w_i)].$$

We impose the following condition.

ASSUMPTION 1: *a)* $E[\Delta(w_i)|x_i]$ *and* $E[\tilde{\Delta}(x_i)|w_i]$ *are complete as functions of* $\Delta(w_i)$ *and* $\tilde{\Delta}(x_i)$ *respectively and b) the conditional pdf* $f_0(\tilde{y}, \tilde{w}|\tilde{x})$ *of* (y_i, w_i) *conditional on* x_i *is bounded away from zero on a neighborhood of* (y, w) *uniformly in* x.

The following result shows that for the function $\delta(\tilde{z})$ above the minimum γ_τ exists for small enough τ and derives the form of γ_τ.

LEMMA 2: *If Assumption 1 is satisfied, $\Delta_j(w_i)$ is bounded, and $E[\Delta_j(w_i)] \neq 0$, for all τ small enough then $\gamma_\tau = \arg\min_\gamma Q_\tau(\gamma)$ exists and is unique and there is $c_j(\tau)$ with $c_j(0) = 0$ and*

$$\gamma_\tau(w_i) = \gamma_0(w_i) + c_j(\tau)\Delta_j(w_i), \frac{\partial c_j(\tau)}{\partial \tau} = \int [\tilde{y} - \gamma_0(\tilde{w})] g(\tilde{y}, \tilde{w}) d\mu / E[\Delta_j(w_i)].$$

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

$$\hat{\beta} = \sum_{i=1}^{n} v(w_i)\hat{\gamma}(w_i)/n, \tag{4.15}$$

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

$$\beta_\tau = E_\tau[v(w_i)\gamma_\tau(w_i)].$$

As shown by Severini and Tripathi (2012), the following condition is necessary for root-n consistent estimability of this β_τ.

ASSUMPTION 2: *There exists $\alpha(x_i)$ such that $v(w_i) = E[\alpha(x_i)|w_i]$ and $E[\alpha(x_i)^2] < \infty$.*

This assumption and generalizations to follow will be key conditions for the form of the adjustment term $\phi(z)$ for first step NP2SLS. To calculate the influence function of $\hat{\beta}$ note that by Assumptions 1 and 2 there is a unique $\alpha(x_i)$ such that

$$\begin{aligned} E[v(w_i)\Delta_j(w_i)] &= E[E[\alpha(x_i)|w_i]\Delta_j(w_i)] = E[\alpha(x_i)\Delta_j(w_i)] = E[\alpha(x_i)E[\Delta_j(w_i)|x_i]] \\ &= E[\alpha(x_i)\delta_x(x_i)]E[\Delta_j(w_i)]. \end{aligned}$$

By the chain rule and Lemma 2

$$\begin{aligned} \frac{\partial \beta(F_\tau)}{\partial \tau} &= \frac{\partial E_\tau[v(w_i)\gamma_\tau(w_i)]}{\partial \tau} \\ &= E[v(w_i)\gamma_0(w_i)S(z_i)] + \frac{\partial c_j(\tau)}{\partial \tau} E[v(w_i)\Delta_j(w_i)] \\ &= \int \{v(\tilde{w})\gamma_0(\tilde{w}) - \beta_0\} g(\tilde{y}, \tilde{w}) d\mu + \left\{ \int [\tilde{y} - \gamma_0(\tilde{w})] g(\tilde{y}, \tilde{w}) d\mu \right\} E[\alpha(x_i)\delta_x(x_i)]. \end{aligned} \tag{4.16}$$

As $j \longrightarrow \infty$ we will have $\int \{v(\tilde{w})\gamma_0(\tilde{w}) - \beta_0\} g(\tilde{y}, \tilde{w}) d\mu \longrightarrow v(w)\gamma_0(w) - \beta_0$ and $\int [\tilde{y} - \gamma_0(\tilde{w})] g(\tilde{y}, \tilde{w}) d\mu \longrightarrow y - \gamma_0(w)$ for $v(\tilde{w})$ and $\gamma_0(\tilde{w})$ continuous at w by the construction of $g(\tilde{y}, \tilde{w})$. Also, as shown

in the proof of Theorem 3 below, there will exist bounded $\Delta_j(w)$ with $E[\Delta_j(w_i)] \neq 0$ such that $E[\alpha(x_i)\delta_x(x_i)] \longrightarrow \alpha(x)$ for $\alpha(\tilde{x})$ and $f_0(\tilde{x})$ continuous at x, so we have

THEOREM 3: *If Assumptions 1 and 2 are satisfied, $f_0(x) > 0$, and each of $\alpha(\tilde{x})$, $f_0(\tilde{x}), v(\tilde{w})$, and $\gamma_0(\tilde{w})$ are continuous at (w, x) then there is $\Delta_j(w_i)$ such that for NP2SLS*

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

Here we find that the influence function of $\hat{\beta}$ of equation (4.15) is $\psi(z)$ of Theorem 3. Like the consumer surplus example a nonparametric residual $y - \gamma_0(w)$ emerges in the calculation of $\psi(z)$. Unlike the surplus example the residual is from the structural equation (4.14) rather than a nonparametric regression. The function $\alpha(x)$ of the instrumental variables is a key component of the influence function. Here $\alpha(x)$ is defined implicitly rather than having an explicit form. This implicit form seems inherent to the NP2SLS first step, where existence of $\alpha(x)$ satisfying Assumption 2 is required for root-n consistency of $\hat{\beta}$, as shown by Severini and Tripathi (2012).

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

$$b_0 = E[v(w_i)(w_i^T d_0)].$$

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

$$v(w)(w^T d_0) - b_0 + a(x)[y - w^T d_0], a(x) = x^T \left(E[w_i x_i^T]\right)^{-1} E[w_i v(w_i)].$$

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

Assumption 2 will only be satisfied when $v(w)$ satisfies certain conditions. When $E[v(w_i)^2] < \infty$ Assumption 2 requires that $v(w_i)$ have Fourier coefficients, with respect to the singular value basis corresponding to the operator $E[\cdot|w_i]$ (for compact $E[\cdot|w_i]$) that decline fast enough relative to the inverse of the singular values; see Section 15.4 of Kress (1989). This condition requires some "smoothness" of $v(w_i)$ and will rule out some functions, such as indicator functions of intervals.

Although asymptotic variances and efficiency bounds have previously been derived by Ai and Chen (2007, 2012), Santos (2011), and Severini and Tripathi (2012), the influence function of Theorem 3 appears to be novel. As discussed above, the form of the influence function is useful for many purposes, such as constructing estimators with improved properties. In this way the NP2SLS influence function formulae of this paper may be useful even for the well studied model of conditional moment restrictions that are linear in the first step.

There is a different way of estimating β_0 that is analogous to Santos (2011). By Assumption 2 and iterated expectations

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

Based on the last equality an estimator of β_0 could be constructed as $\tilde{\beta} = \sum_{i=1}^n \hat{\alpha}(x_i)y_i/n$ where $\hat{\alpha}$ is an estimator $\alpha(x)$. The influence function for this estimator is the same as in Theorem 3. This equality of influence functions occurs because equation (4.17) is satisfied for any F_τ where Assumption 2 holds, i.e. $v(w_i) = E_\tau[\alpha_\tau(x_i)|w_i]$ for some $\alpha_\tau(x)$. Therefore equation (4.17) will hold with E_τ replacing E and α replacing α_τ, so that $\tilde{\beta}$ will have the same limit as $\hat{\beta}$ when the distribution of a single observation is F_τ. Because the influence function is calculated from the limit of the estimator, equality of the limits will mean that $\tilde{\beta}$ and $\hat{\beta}$ have the same influence function.

This influence function calculation can be extended beyond the estimator of (4.15) to other semiparametric GMM-estimators. This extension requires that we specify how $E[m(z_i, \beta_0, \gamma_\tau)]$ depends on γ_τ, as we do in the following condition:

ASSUMPTION 3: *There exists $v(w_i)$ such that for all F_τ, $\partial E[m(z_i, \beta_0, \gamma_\tau)]/\partial\tau = E[v(w_i)\partial\gamma_\tau(w_i)/\partial\tau]$.*

For $m(z, \beta, \gamma) = v(w)\gamma(w) - \beta$ Assumption 3 holds with the $v(w)$ in $m(z, \beta, \gamma)$. More generally the Riesz representation theorem implies that Assumption 3 is equivalent to $\partial E[m(z_i, \beta_0, \gamma_\tau)]/\partial\tau$ being a mean square continuous functional of $\partial\gamma_\tau(w_i)/\partial\tau$. In addition the Riesz representation theorem implies equivalence of Assumption 2 with mean square continuity of $\partial E[m(z_i, \beta_0, \gamma_\tau)]/\partial\tau$ in $E[\partial\gamma_\tau(w_i)/\partial\tau|x_i]$. If Assumption 2 holds then substituting $E[\alpha(x_i)|w_i]$ for $v(w_i)$ and applying iterated expectations gives

$$E[v(w_i)\frac{\partial\gamma_\tau(w_i)}{\partial\tau}] = E[E[\alpha(x_i)|w_i]\frac{\partial\gamma_\tau(w_i)}{\partial\tau}] = E[\alpha\left(x_i\right)E[\frac{\partial\gamma_\tau(w_i)}{\partial\tau}|x_i]],$$

which is a mean square continuous, linear functional of $E[\partial\gamma_\tau(w_i)/\partial\tau|x_i]$. Also, mean square continuity of $E[v(w_i)\partial\gamma_\tau(w_i)/\partial\tau]$ in $E[\partial\gamma_\tau(w_i)/\partial\tau|x_i]$ and the Riesz representation theorem imply that there exists $\alpha(x_i)$ such that

$$E[v(w_i)\frac{\partial\gamma_\tau(w_i)}{\partial\tau}] = E[\alpha\left(x_i\right)E[\frac{\partial\gamma_\tau(w_i)}{\partial\tau}|x_i]] = E[E[\alpha(x_i)|w_i]\frac{\partial\gamma_\tau(w_i)}{\partial\tau}],$$

where the second equality holds by iterated expectations. As $\partial\gamma_\tau(w_i)/\partial w$ varies over a mean square dense set of functions then this equation implies Assumption 2. Similar uses of the Riesz representation theorem are given in Newey (1994), Ai and Chen (2007), and Ackerberg, Chen, Hahn, and Liao (2014).

We can use Assumptions 2 and 3 and Lemma 2 to calculate the influence function of a semiparametric GMM estimator with a NP2SLS first step. Recall from the discussion of equation (2.2) that the influence function of semiparametric GMM is determined by the correction term $\phi(z)$ for the first step and that $\phi(z)$ is the influence function of $\int m(z,\beta,\hat{\gamma})F_0(dz)$. When Assumptions 2 and 3 are satisfied it follows exactly as in equation (4.16) that

$$\begin{aligned}\frac{\partial E[m(z_i,\beta_0,\gamma(F_\tau))]}{\partial\tau} &= \frac{\partial E[m(z_i,\beta_0,\gamma_\tau)]}{\partial\tau} = \frac{\partial c_j(\tau)}{\partial\tau}E[v(w_i)\Delta_j(w_i)] \\ &= \left\{\int[\tilde{y}-\gamma_0(\tilde{w})]g(\tilde{y},\tilde{w})d\mu\right\}E[\alpha(x_i)\delta_x(x_i)].\end{aligned}$$

Then as for Theorem 3 we have:

THEOREM 4: *If Assumption 3 and the other hypotheses of Theorem 3 are satisfied then for the NP2SLS first step*

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

An interesting example is the average derivative estimator of Ai and Chen (2007), where $m(z,\beta,\gamma) = \bar{v}(w)\partial\gamma(w)/\partial w_k - \beta$ for some known $\bar{v}(w)$. Let $v(w) = -f_0(w)^{-1}\partial[\bar{v}(w)f_0(w)]/\partial w_k$ and suppose that Assumption 2 is satisfied for this $v(w)$. Integration by parts and interchanging the order of differentiation and integration gives

$$\frac{\partial E[m(z_i,\beta_0,\gamma_\tau)]}{\partial\tau} = \frac{\partial E[\bar{v}(w_i)\partial\gamma_\tau(w_i)/\partial w]}{\partial\tau} = \frac{\partial E[v(w_i)\gamma_\tau(w_i)]}{\partial\tau} = E[v(w_i)\frac{\partial\gamma_\tau(w_i)}{\partial\tau}],$$

so that Assumption 3 is satisfied. Then by Theorem 4 the adjustment term is $\phi(z) = \alpha(x)[y-\gamma_0(w)]$. It follows by $m(z,\beta,\gamma) = \bar{v}(w)\partial\gamma(w)/\partial w_k - \beta$ that the influence function of $\hat{\beta} = n^{-1}\sum_{i=1}^n \bar{v}(w_i)\partial\hat{\gamma}(w_i)/\partial w_k$ is

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

The asymptotic variance associated with this influence function is

$$E[\psi(z_i)^2] = E[\left\{\bar{v}(w_i)\frac{\partial\gamma_0(w_i)}{\partial w_k} - \beta_0 + \alpha(x_i)[y-\gamma_0(w_i)]\right\}^2]. \tag{4.18}$$

In Appendix B we show that this asymptotic variance formula is identical to that of Ai and Chen (2007) under their conditions, which includes the requirement that there exists $\Delta^*(w_i)$

such that $\alpha(x_i) = E[\Delta^*(w_i)|x_i]$. We do not need this condition here because we have imposed correct specification in calculation of the influence function. When misspecification is allowed it becomes important that such a $\Delta^*(w_i)$ exists, as in Proposition 6 below and in Ai and Chen (2007). In this way the results here contribute by showing that one condition Ai and Chen (2007) impose is not necessary for calculating the influence function. The results also contribute by providing the form of the influence function, and not just the asymptotic variance.

5 Orthogonality Conditions

The above results can be generalized to a setting where the first step estimation is based on orthogonality of a possibly nonlinear residual $\rho(z, \gamma)$ with a set of instrumental variables, where γ denotes a function of the endogenous variables w. In this Section we give this generalization. We continue to use the structure of F_τ in the derivation but depart from the previous Section in just assuming that the limit γ_τ of the NP2SLS estimator exists and is unique when F_τ is the true distribution of a single observation. The previous demonstration of existence and uniqueness motivates our proceeding under those conditions. We expect existence and uniqueness continue to hold in this more general setting but leave that demonstration to future work.

Let $p^K(x_i) = (p_{1K}(x_i), ..., p_{KK}(x_i))^T$ be a vector of instrumental variables that are functions of x_i. We will analyze estimators based on the population orthogonality conditions

$$E[p^K(x_i)\rho(z_i, \gamma_0)] = 0 \text{ for all } K. \tag{5.19}$$

The idea is that as K grows linear combinations of $p^K(x)$ can approximate any element of some nonparametric, infinite dimensional set of instrumental variables. The set of instrumental variables could be all functions of x with finite variance, in which case the orthogonality conditions will be equivalent to the conditional mean zero restriction $E[\rho(z_i, \gamma_0)|x_i] = 0$. The set of instrumental variables could also be smaller, such as the set of functions that are additive in functions of x_i or depend on only some function of x_i, such as a subvector of x_i. We could define the orthogonality conditions in terms of the set of instrumental variables, but it actually saves on notation and detail to formulate orthogonality conditions and define the set of instrumental variables in terms of $p^K(x)$.

These orthogonality conditions motivate estimating γ_0 by minimizing a quadratic form in sample cross products of $p^K(x_i)$ and $\rho(z_i, \gamma)$. We consider the estimator that uses the inverse second moment matrix of the instrumental variables as the middle matrix in this quadratic form. The corresponding objective function is

$$\hat{Q}(\gamma) = n^{-1} \sum_{i=1}^n p^K(x_i)\rho(z_i, \gamma)^T \left(\sum_{i=1}^n p^K(x_i)p^K(x_i)^T \right)^- \sum_{i=1}^n p^K(x_i)\rho(z_i, \gamma),$$

where A^- denotes a generalized inverse of a matrix A. This function can be interpreted as the average of the squared predictions from regressing $\rho(z_i, \gamma)$ on $p^K(x_i)$. The estimator $\hat{\gamma}$ we consider is

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma_n} \hat{Q}(\gamma),$$

where Γ_n is some set of functions that may impose restrictions on $\hat{\gamma}$. We continue to refer to this as NP2SLS because it has the same form as the NP2SLS estimator in NP. We differ here from NP in allowing $p^K(x_i)$ to be restricted in some way and in not necessarily imposing compactness on Γ_n.

To describe the limit of $\hat{\gamma}$ let $F_\tau = (1-\tau)F_0 + \tau G_z^j$ as above and $E_\tau[\cdot]$ denote the corresponding expectation. For each τ close enough to zero define the set of instrumental variables corresponding to $\{p^K(x)\}_{K=1}^\infty$ and F_τ to be

$$\mathcal{A}_\tau = \{a(x) : E_\tau[a(x_i)^2] < \infty \text{ and } \exists b^K \text{ s.t. } \lim_{K \longrightarrow \infty} E_\tau[(a(x_i) - p^K(x_i)^T b^K)^2] = 0\}.$$

By construction $\mathcal{A}_\tau$ is linear and closed in mean square. Also, $\mathcal{A}_\tau = \mathcal{A}_0$ by equation (3.9). Consider the Hilbert space $\mathcal{H}_\tau$ of functions of z_i that have finite second moment at F_τ and inner product $E_\tau[a(z_i)b(z_i)]$. For $b(z_i) \in \mathcal{H}_\tau$ let $\pi_\tau(\rho(\gamma), x_i)$ and $\pi_\tau(b, x_i)$ denote the orthogonal projection of $\rho(z_i, \gamma)$ and $b(z_i)$ on $\mathcal{A}_\tau$ when the true distribution is F_τ. It follows exactly as in Newey (1991) that the limit of the NP2SLS objective function will be

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

Suppose that there is a set Γ that is linear and closed in $\mathcal{H}_0$, contains each Γ_n, and that for all $\Delta \in \Gamma$ there exists $\Delta_n \in \Gamma_n$ with $\Delta_n \longrightarrow \Delta$ in $\mathcal{H}_0$. For example, Γ might be the closure in $\mathcal{H}_0$ of additive functions of some subvectors of w_i. Also, it follows by $\delta(z)$ bounded that closed sets in $\mathcal{H}_0$ are also closed in $\mathcal{H}_\tau$. It then follows as in Newey (1991) that the limit of the NP2SLS estimator will be

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

We will use first order conditions for γ_τ to characterize the adjustment term $\phi(z)$ for NP2SLS in this setting. To do so it is helpful to introduce some additional objects and be more specific about the structure. Let ζ denote a scalar and $\Delta(w)$ some function of the endogenous variables w with finite variance such that $\Delta(w_i) \in \Gamma$. Then we have $\gamma_\tau(w_i) + \zeta\Delta(w_i) \in \Gamma$ for any ζ by Γ linear. We will assume that there is $d_\tau(w_i, x_i)$ such that for $\Delta_\tau(w_i) = \partial\gamma_\tau(w_i)/\partial\tau$

$$\left.\frac{\partial \pi_\tau(\rho(\gamma_\tau + \zeta\Delta), x_i)}{\partial \zeta}\right|_{\zeta=0} = \pi_\tau(d_\tau\Delta, x_i), \frac{\partial \pi_0(\rho(\gamma_\tau), x_i)}{\partial \tau} = \pi_0(d_0\Delta_\tau, x_i). \tag{5.20}$$

Let

$$\begin{aligned}\mathbf{A} &= \{\pi_0(d_0\Delta, x_i) : \Delta \in \Gamma\},\ \bar{\mathbf{A}} \text{ equal the closure of } \mathbf{A} \text{ in } \mathcal{H}_0,\\ &\quad \pi^*(b, w_i) \text{ equal the projection of } b(z_i) \text{ on } \Gamma \text{ in } \mathcal{H}_0.\end{aligned}$$

The following condition will be used to characterize the adjustment term $\phi(z)$ for $\hat{\gamma}$.

ASSUMPTION 4: *a)* $p_{kK}(x) \in A_\tau$ *for all* K, $k \leq K$ *and each* τ; *b) there is* $d_\tau(w_i, x_i)$ *such that for all* τ *small enough and any* $\Delta \in \Gamma$, $\pi_\tau(\rho(\gamma_\tau + \zeta\Delta), x_i)$ *and* $\pi_0(\rho(\gamma_\tau), x_i)$ *satisfy equation (5.20); c) Assumption 3 is satisfied for some* $v(w_i)$ *and for the projection* $v^*(w_i)$ *of* $v(w_i)$ *on* Γ *in* $\mathcal{H}_0$ *there exists* $\alpha(x_i) \in \mathcal{A}_0$ *such that* $v^*(w_i) = -\pi^*(d_0\alpha, w_i)$.

Assumption 4 a) requires that the functions $p_{kK}(x)$ are nested in the sense that for every fixed $\bar{k}$ and $\bar{K}$ the function $p_{\bar{k}\bar{K}}(x)$ can be approximated in mean square by $p^K(x)^T b^K$ as K grows for some b^K. It will automatically hold when $p_{kK}(x)$ does not depend on K, e.g. as for power series, and will also hold for regression splines. Assumption 4 b) adds structure that is useful when $\rho(z, \gamma)$ is nonlinear in γ.

As an example consider $\beta_0 = E[v(w_i)\gamma_0(w_i)]$ with a linear residual $\rho(z, \gamma) = y - \gamma(w)$, $\gamma(w) = \gamma_1(w_1) + \gamma_2(w_2)$ restricted to be additive in distinct components of $w = (w_1, w_2)$, and the set of instrumental variables is all functions of x_i with finite mean-square. Here $d_0 = -1$, Γ will be the mean-square closure of the set of additive functions of the form $\gamma_1(w_1) + \gamma_2(w_2)$, and the instrument orthogonality condition is $E[y_i - \gamma_0(w_i)|x_i] = 0$. Then $\bar{\mathbf{A}} = \{\pi_0(\Delta, x_i) : \Delta \in \Gamma\}$ is the closure of the set of conditional expectations of additive functions. For Assumption 4 c) to hold in this example it is sufficient that there is $\alpha(x_i)$ such that $v(w_i) = E[\alpha(x_i)|w_i]$, because by iterated projections, it will follow that $v^*(w_i) = \pi^*(E[\alpha|w], w_i) = \pi^*(\alpha, w_i)$.

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

$$\begin{aligned}\frac{\partial E_\tau[\rho(z_i, \gamma_\tau + \zeta\Delta)|w_i, x_i]}{\partial\zeta} &= d_\tau(w_i, x_i)\Delta(w_i), \qquad (5.21)\\ d_\tau(w_i, x_i) &= f_\tau(\gamma_\tau(w_i)|w_i, x_i).\end{aligned}$$

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

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

giving the first equation of Assumption 4 b). The second equation of Assumption 4 b) follows similarly. More generally Assumption 4 b) will hold if $\rho(z,\gamma) = \rho(z,\gamma(w))$ with $d_\tau(w,x) = \partial E[\rho(z_i,\gamma_\tau(w)+\zeta)|w_i,x_i]/\partial\zeta|_{\zeta=0}$.

Assumption 4 c) generalizes Assumption 2 to nonlinear residuals. Similarly to Assumptions 2 and 3, the existence of $\alpha(x_i)$ can be shown to be equivalent to other conditions using the Riesz representation theorem. If the function $E[v(w_i)\Delta(w_i)]$ is a mean square continuous functional of $\pi_0(d_0\Delta,x_i)$ over all $\Delta\in\Gamma$ then Assumption 4 c) follows by the Riesz representation theorem.

We use Assumption 4 to obtain first order conditions for γ_τ. By calculus of variations and Assumption 4 the first order conditions for γ_τ are that

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

for all $\Delta\in\Gamma$. This equation is an identity in τ. Differentiating in τ at $\tau=0$ and applying the chain rule it follows that for all $\Delta\in\Gamma$

$$\begin{aligned} 0 &= E[\pi_0(d_0\Delta,x_i)\pi_0(d_0\Delta',x_i)] + E[\pi_0(\rho(\gamma_0),x_i)\pi_0(d_0\Delta,x_i)S(z_i)] \\ &\quad + E[\pi_0(d_0\Delta,x_i)\frac{\partial\pi_\tau(\rho(\gamma_0),x_i)}{\partial\tau}] + E[\pi_0(\rho(\gamma_0),x_i)\frac{\partial\pi_\tau(d_0\Delta,x_i)}{\partial\tau}] \\ &\quad + E[\pi_0(\rho(\gamma_0),x_i)\frac{\partial\pi_0(d_\tau\Delta,x_i)}{\partial\tau}]. \end{aligned} \tag{5.22}$$

The following result gives alternative expressions for the third and fourth terms in this equation.

LEMMA 5: *$\mathcal{A}_0 = \mathcal{A}_\tau$ and for any $a(x_i)\in\mathcal{A}_0$ and $b(z_i)$ with $E[b(z_i)^2]<\infty$,*

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

By applying Lemma 5 to the third and fourth terms of equation (5.22) and solving for the first term we see that

$$E[\pi_0(d_0\Delta,x_i)\pi_0(d_0\Delta',x_i)] = E[\phi_\Delta(z_i)S(z_i)] - E[\pi_0(\rho(\gamma_0),x_i)\frac{\partial\pi_0(d_\tau\Delta,x_i)}{\partial\tau}], \tag{5.23}$$

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

This formula can be combined with Assumption 4 to obtain the adjustment term when the first step is the NP2SLS estimator. We state the result as a Proposition, similarly to the Propositions of Newey (1994), where derivations use formal calculations without specifying a complete set of regularity conditions.

PROPOSITION 6: *Suppose that Assumption 4 is satisfied and the pdf of z_i is bounded away from zero in a neighborhood of z. Let $\alpha^*(x_i)$ be the projection of $\alpha(x_i)$ on $\bar{\mathbf{A}}$. If $\pi_0(\rho(\gamma_0), x_i) = 0$ and $\alpha^*(\tilde{x})$ and $\rho(\tilde{z}, \gamma_0)$ are continuous at z then the adjustment term is*

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

If $\pi_0(\rho(\gamma_0), x_i) \neq 0$, $\partial \pi_0(d_\tau \Delta, x_i)/\partial \tau = 0$ for all $\Delta \in \Gamma$, there exists $\Delta^* \in \Gamma$ such that $\alpha^*(x_i) = -\pi_0(d_0 \Delta^*, x_i)$, and $\phi_{\Delta^*}(\tilde{z})$ from equation (8.31) is continuous at z then the adjustment term is $\phi_{\Delta^*}(z)$.

We prove Proposition 6 in the Appendix. One important case where $\pi_0(\rho(\gamma_0), x_i) = 0$ holds is where the orthogonality condition (5.19) is satisfied. This is the correctly specified case. Another important case where $\pi_0(\rho(\gamma_0), x_i) = 0$ is where a solution to $\pi_0(\rho(\gamma), x_i) = 0$ exists even when the model is misspecified. We might think of this as an "exactly identified" case, similarly to Chen and Santos (2015).

In the misspecified case where $\pi_0(\rho(\gamma_0), x_i) \neq 0$ Proposition 6 assumes that $\partial \pi_0(d_\tau \Delta, x_i)/\partial \tau = 0$. We do not know if an influence function exists when $\pi_0(\rho(\gamma_0), x_i) \neq 0$ and $\partial \pi_0(d_\tau \Delta, x_i)/\partial \tau \neq 0$. The problem is that d_τ may be a nonparametric object evaluated at a point and hence $\partial \pi_0(d_\tau \Delta, x_i)/\partial \tau$ may not have a representation as an expected product with the score. In such cases $\sqrt{n}$ consistent estimation may not be possible. For example, for quantile IV, $d_\tau(w_i, x_i) = f_\tau(\gamma_\tau(w_i)|w_i, x_i)$. This is a nonparametric conditional density evaluated at a point, which is not root-n consistently estimable. In the misspecified case $m(z, \beta, \gamma)$ may depend on the value of this density at a point so that it may not be possible to root-n consistently estimate β. We leave analysis of this question to future work.

Under first step misspecification where $\pi_0(\rho(\gamma_0), x_i) \neq 0$ we also require that there exists $\Delta^*(w)$ with $\alpha^*(x_i) = -\pi_0(d_0 \Delta^*, x_i)$ This condition will impose additional smoothness conditions on $\alpha^*(x_i)$ and hence on $v^*(w_i)$. It requires $\alpha^*(x)$ be smooth like Assumption 3 requires that $v(w)$ be smooth. In addition, it may be that the existence of $\Delta^*(w)$ with $\alpha^*(x_i) = \pi_0(d_0 \Delta^*, x_i)$ is necessary for root-n consistency when $\pi_0(\rho(\gamma_0), x_i) \neq 0$. The existence of $\Delta^*(w)$ is like Assumption 3 for $v(w)$ that Severini and Tripathi (2012) showed is necessary condition for root-n consistency under correct specification. We leave analysis of this question to future work also.

We illustrate by giving the adjustment term when the residual is linear in an addditive function. As discussed earlier, $\alpha(x_i)$ is a function such that $v(w_i) = E[\alpha(x_i)|w_i]$. Then $\alpha^*(x_i)$ is the projection of $\alpha(x_i)$ on the closure of the set of conditional expectations of additive functions conditional on x_i. It follows that the adjustment term is

$$\phi(z_i) = \alpha^*(x_i)[y_i - \gamma_{10}(x_{1i}) - \gamma_{20}(x_{2i}).$$

Another illustration is the adjustment term for the average derivative of an endogenous quantile model where linear restrictions, such as additivity, may be imposed on the structural function. Here $m(z,\beta,\gamma) = \bar{v}(w)\partial\gamma(w)/\partial w_j - \beta$ as before, for some known weight function $\bar{v}(w)$. As before let $v(w) = -f_0(w)^{-1}\partial[\bar{v}(w)f_0(w)]/\partial w_j$. Let $\rho(z,\gamma) = 1(y < \gamma(w)) - \eta$ and $\mathcal{A}_0$ be the set of all functions of x_i with finite variance. In this case $\pi_0(b,x) = E[b(z_i)|x_i]$ and the orthogonality conditions of equation (5.19) is $E[\rho(z_i,\gamma_0)|x_i] = 0$, the endogenous quantile model. As noted above we here have $d_0(w_i,x_i) = f_0(\gamma_0(w_i)|w_i,x_i)$. Assumption 4 c) says that there must exist $\alpha(x)$ such that for the projection $v^*(w_i)$ of $v(w_i)$ on Γ,

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

Similar to Assumption 3, existence of such a $\alpha(x_i)$ places some "smoothness" restrictions on $v(w_i)$. Let $\alpha^*(x_i)$ be the projection of $\alpha(x_i)$ on $\bar{\mathbf{A}}$ in $\mathcal{H}_0$. Proposition 6 then gives the adjustment term

$$\phi(z_i) = \alpha^*(x_i)[1(y_i < \gamma_0(w_i)) - \eta].$$

The function $\alpha^*(x)$ is central to the form of the correction term in Proposition 6. This function quantifies how the instruments affects the influence function. Here $\alpha^*(x)$ is constrained to be an element of $\bar{\mathbf{A}}$ because NP2SLS projects functions of w on the instruments, just as parametric 2SLS does. By choosing the orthogonality conditions it is possible to also vary $\alpha^*(x_i)$ because the projection of functions of w on $\bar{\mathbf{A}}$ will change. In fact, by choosing the orthogonality conditions it appears possible to have $\alpha^*(x_i)$ to be equal to any function satisfying Assumption 4 c). Let $\alpha(x)$ be any function satisfying Assumption 4 c) and consider $a(x) = (\alpha(x), \tilde{a}(x))$ that is just identifying, in the sense of Chen and Santos (2015), meaning that

$$\bar{\mathbf{A}} = \{b(a(x)) : E[b(a(x_i))^2] < \infty\}.$$

In this case the projection of $\alpha(x_i)$ on $\bar{\mathbf{A}}$ is just $\alpha(x_i)$, so the correction term is $\alpha(x)\rho(z,\gamma_0)$. By choosing the orthogonality conditions in this way the correction term for the NP2SLS estimator can vary over all functions satisfying Assumption 4 c). This result is analogous to a parametric linear model, where 2SLS can be made to vary over all IV estimators by choosing the instruments to vary over all linear combinations of instrumental variables.

6 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 we know both $\psi(z)$ and $\hat{\beta}$ we know the remainder $\sqrt{n}(\hat{\beta}-\beta_0)-\sum_{i=1}^n \psi(z_i)/\sqrt{n}$ that must converge in probability to zero for $\hat{\beta}$ to be asymptotically

linear. This remainder term can then be analyzed to find conditions for it to be small. A decomposition of this remainder is often useful in the analysis. We consider here a remainder decomposition like Newey (1994), except that we impose stochastic equicontinuity on $\hat{m}(\beta_0)$ before linearizing and allow for weaker convergence conditions for other remainders. These conditions differ from the previous literature in requiring weaker hypotheses in one respect, as detailed below. We also provide primitive conditions for the linear remainder term for first step series regression. These conditions are sufficient for asymptotic linearity of a linear functional of a series regression. They improve upon previous results by combining conditions from Newey (1994) and Belloni, Chernozhukov, Chetverikov, and Kato (2015).

We focus on the key condition that $\sqrt{n}\hat{m}(\beta_0)$ is asymptotically linear. Let $\mu(\gamma) = \int m(z, \beta_0, \gamma) F_0(dz)$ and $D(\gamma)$ be a linear functional that will be a Frechet derivative of $\mu(\gamma)$ with respect to γ at γ_0. Consider the decomposition

$$\sqrt{n}\hat{m}(\beta_0) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [m(z_i, \beta_0, \gamma_0) + \phi(z_i)] = \hat{R}_1 + \hat{R}_2 + \hat{R}_3,$$

$$\begin{aligned} \hat{R}_1 &= \sqrt{n}[\hat{m}(\beta_0) - \frac{1}{n} \sum_{i=1}^{n} m(z_i, \beta_0, \gamma_0) - \mu(\hat{\gamma})], \hat{R}_2 = \sqrt{n}\left[\mu(\hat{\gamma}) - D(\hat{\gamma} - \gamma_0)\right], \\ \hat{R}_3 &= \sqrt{n} D(\hat{\gamma} - \gamma_0) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(z_i). \end{aligned}$$

The term $\hat{R}_1$ is a stochastic equicontinuity remainder of the type considered by Andrews (1994) and Van der Vaart and Wellner (1996). Extensive conditions for $\hat{R}_1 \xrightarrow{p} 0$ can be found there. These conditions can allow for $m(z, \beta_0, \gamma)$ to not be smooth in γ. Many of these conditions involve boundedness of derivatives of $\hat{\gamma}$ with respect to its arguments. When sample splitting is used, where $\hat{\gamma}$ is estimated from different observations than in the average $\hat{m}(\beta_0)$, $\hat{R}_1 \xrightarrow{p} 0$ will hold under much less stringent conditions on $\hat{\gamma}$; see Chernozhukov et. al. (2016a,b).

The term $\hat{R}_2$ is a linearization remainder from approximating $\mu(\hat{\gamma})$ by the linear functional $D(\hat{\gamma} - \gamma_0)$. It will converge to zero if there is a pseudo norm $\|\cdot\|$ such that:

ASSUMPTION 5: *For some* $1 < \zeta \leq 2$, $\mu(\gamma) = D(\gamma - \gamma_0) + O(\|\gamma - \gamma_0\|^{\zeta})$ *and* $\|\hat{\gamma} - \gamma_0\| = o_p(n^{-1/2\zeta})$.

This condition separates nicely into two parts, one about the properties of the functional $\mu(\gamma)$ and another about a convergence rate for $\hat{\gamma}$. By virtue of the fact that $\mu(\gamma)$ does not depend on the data the $\mu(\gamma)$ condition of Assumption 5 will often be satisfied when $\|\gamma - \gamma_0\|$ is an L_p norm with $p < \infty$. This $\mu(\gamma)$ condition is somewhat stronger than Frechet differentiability.

Frechet differentiability would require that $\mu(\gamma)-D(\gamma-\gamma_0) = o(\|\gamma-\gamma_0\|)$, whereas Assumption 5 requires $\mu(\gamma) - D(\gamma-\gamma_0) = O(\|\gamma-\gamma_0\|^\zeta)$. When $\mu(\gamma)$ is twice Frechet differentiable with continuous second derivative the $\mu(\gamma)$ condition will hold with $\zeta = 2$. The $\hat{\gamma}$ condition requires that $\hat{\gamma}$ converge to γ_0 at a rate that depends on ζ. When $\zeta = 2$ the $\hat{\gamma}$ hypothesis becomes $\|\hat{\gamma}-\gamma_0\| = o_p(n^{-1/4})$, a $n^{-1/4}$ consistency condition like that previously used by Ait-Sahalia (1991), Andrews (1994), Newey (1994), Newey and McFadden (1994), Chen and Shen (1997), Chen, Linton, and van Keilegom (2003), Ichimura and Lee (2010), and others. Assumption 5 allows for $\zeta < 2$ at the price of a faster convergence rate for $\hat{\gamma}$, similarly to Chen (2007) and Ichimura and Lee (2010).

The condition that $\hat{R}_3 \xrightarrow{p} 0$ is equivalent to the statement that $D(\hat{\gamma}-\gamma_0)$ is asymptotically linear with influence function $\phi(z)$. Here $\phi(z)$ is the correction term discussed earlier in the paper. This specification of $\phi(z)$ will coincide with the earlier definition of $\phi(z)$ as the influence function of $\int m(z,\beta_0,\hat{\gamma})F_0(dz)$ because it is the linear term in the expansion that determines the influence function. As is discussed in Newey (1994) and Newey and McFadden (1994), $D(\hat{\gamma}-\gamma_0)$ being asymptotically linear requires that $D(\gamma-\gamma_0)$ be a mean square continuous functional of γ. When γ is a conditional expectation this means that there is $\alpha(x)$ such that

$$D(\gamma-\gamma_0) = E[\alpha(x_i)\{\gamma(x_i)-\gamma_0(x_i)\}]. \tag{6.24}$$

When γ is a pdf of x this means that

$$D(\gamma-\gamma_0) = \int \alpha(x)[\gamma(x)-\gamma_0(x)]dx.$$

The remainder $\hat{R}_3$ often has an important expectation component that is related to the bias of $\hat{m}(\beta_0)$. Let χ_n be some conditioning set (sigma algebra) such that $E[\phi(z_i)|\chi_n] = 0$. Here χ_n may be constant, in which case the conditional expectation is the unconditional expectation and $E[\phi(z_i)|\chi_n] = 0$ is automatically satisfied because $\phi(z_i)$ has mean zero. For first stage series regression estimators a convenient choice of χ_n is the observations on the regressors. In general, χ_n can be chosen in a way that is convenient for a particular estimator. Assuming that the expectation of the linear functional $D(\hat{\gamma})$ is the functional of the expectation,

$$E[\hat{R}_3|\chi_n] = \sqrt{n}E[D(\hat{\gamma}-\gamma_0)|\chi_n] = \sqrt{n}D(E[\hat{\gamma}|\chi_n]-\gamma_0).$$

Often $\hat{\gamma}$ can be thought of as the result of some smoothing operation applied to the empirical distribution. A corresponding interpretation of $D(E[\hat{\gamma}|\chi_n]-\gamma_0)$ is smoothing bias in a linear approximation to the moment conditions as a function of γ. Consequently, requiring that $\hat{R}_3 \xrightarrow{p} 0$ will include a requirement that $\sqrt{n}$ times this smoothing bias goes to zero. Also $\hat{R}_3 - E[\hat{R}_3|\chi_n]$ will need to go zero in order for $\hat{R}_3 \xrightarrow{p} 0$. The term $\hat{R}_3 - E[\hat{R}_3|\chi_n]$ will

generally be a stochastic equicontinuity remainder, which is bounded in probability for fixed γ and converges to zero as γ approaches γ_0. In examples this part of the remainder goes to zero under quite weak conditions, while the bias term going to zero requires important smoothness conditions, as discussed below.

A condition that controls both the bias and stochastic equicontinuity terms in $\hat{R}_3$ is

ASSUMPTION 6: $E[\phi(z_i)] = 0$, $E[\|\phi(z_i)\|^2] < \infty$, and there is a sigma algebra χ_n such that $E[\hat{R}_3|\chi_n] \xrightarrow{p} 0$ and $\hat{R}_3 - E[\hat{R}_3|\chi_n] \xrightarrow{p} 0$.

Assumptions 5 and 6 can be combined with stochastic equicontinuity and a few uniform convergence conditions to give a precise result for semiparametric GMM.

THEOREM 7: *If $\hat{\beta} \xrightarrow{p} \beta_0$, $\hat{W} \xrightarrow{p} W$, β_0 is in the interior of the parameter set, $\hat{m}(\beta)$ is continuously differentiable in a neighborhood of β_0 with probability approaching* 1, *for any $\bar{\beta} \xrightarrow{p} \beta_0$ we have $\partial\hat{m}(\bar{\beta})/\partial\beta \xrightarrow{p} M$, M^TWM is nonsingular, $\hat{R}_1 \xrightarrow{p} 0$ and Assumptions 5 and 6 are satisfied then $\hat{\beta}$ is asymptotically linear with influence function $-(M^TWM)^{-1}M^TW[m(z,\beta_0,\gamma_0)+\phi(z)]$.*

This result differs from those of Newey (1994), Newey and McFadden (1994), Chen, Linton, and van Keilegom (2003), and Ichimura and Lee (2010) in the separation into bias and stochastic equicontinuity terms in Assumption 6. For simplicity we have assumed that $\hat{m}(\beta)$ is differentiable in β but it would be straightforward to extend the results to allow $\hat{m}(\beta)$ to not be smooth in β, using results as in Chen, Linton, and Van Keilegom (2003) and Ichimura and Lee (2010).

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 function 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 best to use that approach for semiparametric estimators, where the first step often involves nonparametric estimation of a conditional expectation or pdf. Even when the estimator is Hadamard differentiable in the first step, the smoothing bias induced by a nonparametric first step depends on the smoothness of the influence function, e.g. as in the series estimators of Newey (1994). This feature of semiparametric estimators makes it important to explicitly account for remainders like the $E[\hat{R}_3|\chi_n]$ of Assumption 6.

A key step in using the remainder decomposition here to obtain asymptotic linearity is showing that Assumption 6 is satisfied. Specifying primitive conditions for Assumption 6 requires the most work because there are well known conditions for Assumptions 4 and 5 to be satisfied. We illustrate how this can be done by giving conditions when γ is a conditional expectation and $\hat{\gamma}$ is a series estimator. For γ a conditional expectation Proposition 4 of Newey (1994) gives

$$\phi(z) = \alpha(x)[y - \gamma_0(x)],$$

where $\alpha(x)$ is from the Riesz representation in equation (6.24). The result we obtain below may be of independent interest the size of the remainder term turns out to be as small as known to be possible when the regression function is smooth enough.

To describe a series estimator of a conditional expectation let the true first step be $\gamma_0(x) = E[y|x]$. A series estimator $\hat{\gamma}(x)$ of $\gamma_0(x)$ can be formed as the predicted value at x of a series regression of y_i on $p^K(x_i)$ where $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))^T$ is a vector of approximating functions. Here

$$\hat{\gamma}(x) = p^K(x)^T \hat{\Sigma}^{-1} \hat{b}, \hat{b} = \frac{1}{n} \sum_{i=1}^{n} p^K(x_i) y_i, \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} p^K(x_i) p^K(x_i)^T.$$

Conditions for the stochastic equicontinuity condition $\hat{R}_1 \xrightarrow{p} 0$ from Andrews (1994) will require that $\hat{\gamma}(x)$ have bounded derivatives of sufficiently high order in large samples. Primitive conditions for such properties as well as for Assumption 5 could be obtained by applying the results of Belloni, Chernozhukov, Chetverikov, and Kato (2015). Alternatively, much weaker conditions for stochastic equicontinuity will suffice when sample splitting is used as in Chernozhukov et. al. (2016a,b). Also, conditions for Assumption 5 may be obtained using mean-square convergence rates of Newey (1997) or uniform rates of Belloni et. al. (2015). Specifics will depend on the form of $\mu(\gamma)$.

Turning next to Assumption 6, note that as discussed above, root-n consistency requires that there is $\alpha(x)$ with $D(\gamma - \gamma_0) = E[\alpha(x_i)\{\gamma(x_i) - \gamma_0(x_i)\}]$. Let $v = E[\alpha(x_i)p^K(x_i)]$, $\tilde{\alpha}(x) = v^T \hat{\Sigma}^{-1} p^K(x)$, $\delta = \left(E[p^K(x_i)p^K(x_i)^T]\right)^{-1} E[p^K(x_i)\gamma_0(x_i)]$, and $\gamma_K(x) = p^K(x)^T\delta$. Note that

$$\int \alpha(x)\hat{\gamma}(x)F_0(dx) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\alpha}(x_i)y_i, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\alpha}(x_i)\gamma_K(x_i) = \sqrt{n}v'\delta = \sqrt{n}E[\alpha(x_i)\gamma_K(x_i)].$$

Then adding and subtracting terms gives, observing that $\phi(z_i) = \alpha(x_i)[y_i - \gamma_0(x_i)]$,

$$\begin{aligned}
\hat{R}_3 &= \frac{1}{\sqrt{n}} \sum_i \tilde{\alpha}(x_i)y_i - \sqrt{n}E[\alpha(x_i)\gamma_0(x_i)] - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(z_i) = \hat{R}_{31} + \hat{R}_{32} + \hat{R}_{33}, \\
\hat{R}_{31} &= \sqrt{n}E[\alpha(x_i)\{\gamma_K(x_i) - \gamma_0(x_i)\}], \hat{R}_{32} = v^T \hat{\Sigma}^{-1} \frac{1}{\sqrt{n}} \sum_i p^K(x_i)[\gamma_0(x_i) - \gamma_K(x_i)] \\
\hat{R}_{33} &= \frac{1}{\sqrt{n}} \sum_i [\tilde{\alpha}(x_i) - \alpha(x_i)][y_i - \gamma_0(x_i)],
\end{aligned}$$

Let χ_n be the sigma algebra generated by $\{x_1, ..., x_n\}$. Then since $\tilde{\alpha}(x_i)$ depends only on $\{x_1, ..., x_n\}$ and $E[y_i - \gamma_0(x_i)|\chi_n] = 0$,

$$E[\hat{R}_3|\chi_n] = \hat{R}_{31} + \hat{R}_{32}, \quad \hat{R}_3 - E[\hat{R}_3|\chi_n] = \hat{R}_{33}.$$

We consider first $\hat{R}_{31}$. For $\Sigma = E[p(x_i)p(x_i)^T]$ let $\delta_\alpha = \Sigma^{-1}v$ be the coefficients of the population projection of $\alpha(x_i)$ on $p^K(x_i)$ and $\alpha_K(x) = p^K(x)^T\delta_\alpha$ Then by $\gamma_0(x_i) - \gamma_K(x_i)$ being orthogonal to $p^K(x_i)$ in the population,

$$\hat{R}_{31} = E[\alpha(x_i)\{\gamma_K(x_i) - \gamma_0(x_i)\}] = -E[\{\alpha(x_i) - \alpha_K(x_i)\}\{\gamma_0(x_i) - \gamma_K(x_i)\}].$$

As pointed out in Newey (1994), the size of this bias term is determined by the product of series approximation errors to $\alpha(x_i)$ and to $\gamma_0(x_i)$. Thus, the bias of a series semiparametric estimator will generally be smaller than the nonparametric bias for a series estimate of $\gamma_0(x)$. For example, for splines of order κ, if $\gamma_0(x)$ and $\alpha(x)$ are continuously differentiable of order s_γ and s_α respectively, x is r-dimensional, and the support of x is compact then by standard approximation theory, for $\bar{s}_\gamma = \min\{s_r, \kappa + 1\}$ and $\bar{s}_\alpha = \min\{s_\alpha, \kappa + 1\}$

$$|E[\{\alpha(x_i) - \alpha_K(x_i)\}\{\gamma_0(x_i) - \gamma_K(x_i)\}]| \leq CK^{-(\bar{s}_\gamma + \bar{s}_\alpha)/r}.$$

Turning now to $\hat{R}_{32}$, note that $E[p^K(x_i)\{\gamma_0(x_i) - \gamma_K(x_i)\}] = 0$. Also, $\gamma_K(x_i)$ is the minimum mean square error approximation to $\gamma_0(x_i)$ so that $\gamma_K(x_i)$ will get close to $\gamma_0(x_i)$ as K grows. Then $\hat{R}_{32}$ is a stochastic bias term that can be dealt with using the results of Belloni et. al. (2015).

We see that the term $\hat{R}_{33}$ is a stochastic equicontinuity term that will be $o_p(1)$ as $\tilde{\alpha}(x)$ gets close to $\alpha(x)$. In particular, since $\tilde{\alpha}(x_i)$ depends only on $x_1, ..., x_n$,

$$E[\hat{R}_{33}^2|\chi_n] = \frac{1}{n}\sum_{i=1}^{n}[\tilde{\alpha}(x_i) - \alpha(x_i)]^2 Var(y_i|x_i)$$

which will be small when $Var(y_i|x_i)$ is bounded and $\tilde{\alpha}(x_i) - \alpha(x_i)$ is small.

Turning now to the regularity conditions for asymptotic linearity, we follow Belloni et. al. (2015) and impose the following assumption:

ASSUMPTION 7: *$Var(y_i|x_i)$ is bounded, $E[\alpha(x_i)^2] < \infty$, 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 $c_K \longrightarrow 0$ and ℓ_K such that $\sqrt{E[\{\gamma_0(x_i) - \gamma_K(x_i)\}^2]} \leq c_K$, $\sup_{x\in\chi}|\gamma_0(x) - \gamma_K(x)| \leq \ell_K c_K$, and for $\xi_K = \sup_{x\in\chi}\left\|p^K(x)\right\|$, $\xi_K^2(\ln K)/n \longrightarrow 0$.*

We also make use of a bound on the error from a series approximation of $\alpha(x)$.

ASSUMPTION 8: $\sqrt{E[\{\alpha(x_i) - p^K(x_i)^T\delta_\alpha\}^2]} \leq c_K^\alpha$.

Belloni et. al. (2015) give an extensive discussion of the size of c_K, c_K^α, ℓ_K, and ξ_K for various kinds of series approximations and distributions for x_i. For power series, x_i continuously distributed on a rectangular bounded support in $\Re^r$, $\gamma_0(x)$ and $\alpha(x)$ continuously differentiable of order s_γ and s_α respectively, Assumptions 7 and 8 are satisfied with $c_K = CK^{-s_\gamma/r}$, $c_K^\alpha = CK^{-s_\alpha/r}$, and $\ell_K = K$, $\xi_K = K$, and $K^2 (\ln K)/n \longrightarrow 0$. For tensor product splines of order κ, $\bar{s}_\gamma = \min\{s_\gamma, \kappa + 1\}$, and $\bar{s}_\gamma = \min\{s_\gamma, \kappa + 1\}$ Assumptions 7 and 8 are satisfied with $c_K = CK^{-\bar{s}_\gamma/r}$, $c_K^\alpha = CK^{-\bar{s}_\alpha/r}$, $\ell_K = C$, $\xi_K = \sqrt{K}$, and $K (\ln K)/n \longrightarrow 0$.

THEOREM 8: *If* $D(\gamma - \gamma_0) = \int \alpha(x)[\gamma(x) - \gamma_0(x)]F_0(dx)$ *and Assumptions 7 and 8 are satisfied then for* $\phi(z) = \alpha(x)[y - \gamma_0(x)]$,

$$\sqrt{n}D(\hat{\gamma} - \gamma_0) = \sum_{i=1}^{n} \phi(z_i)/\sqrt{n} + O_p(\sqrt{n}c_K c_K^\alpha + c_K\ell_K + c_K^\alpha + \sqrt{\frac{\xi_K^2 \ln(K)}{n}}(1 + \sqrt{K}c_K\ell_K)).$$

Applying the previous discussion about regression splines to the remainder in Theorem 8 gives the following result on asymptotic linearity of $D(\hat{\gamma} - \gamma_0)$ when $\hat{\gamma}$ is a spline nonparametric regression.

COROLLARY 9: *If the hypotheses of Theorem 8 are satisfied,* $p^K(x)$ *are regression splines, the minimum knot width is bounded below by* C/K *for a constant* C, *and the support of* x_i *is* $[0,1]^r$, *then for* $\phi(z) = \alpha(x)[y - \gamma_0(x)]$,

$$\sqrt{n}D(\hat{\gamma}-\gamma_0) = \sum_{hi=1}^{n} \phi(z_i)/\sqrt{n}+O_p(\sqrt{n}K^{-(\bar{s}_\gamma+\bar{s}_\alpha)/r}+K^{-\bar{s}_\gamma/r}+K^{-\bar{s}_a/r}+\sqrt{\frac{K\ln(K)}{n}}(1+K^{(1/2)-(\bar{s}_\gamma/r)}))$$

Here we see that $\sqrt{n}D(\hat{\gamma} - \gamma_0)$ will be asymptotically linear when

$$\sqrt{n}K^{-(\bar{s}_\gamma+\bar{s}_\alpha)/r} \longrightarrow 0, \sqrt{\frac{K\ln(K)}{n}}(1 + K^{(1/2)-(\bar{s}_\gamma/r)}) \longrightarrow 0. \tag{6.25}$$

We have motivated these series estimation results as conditions for an important remainder in an asymptotically linear representation to be small. Theorem 8 and Corollary 9 apply directly to show asymptotic linearity of linear functionals $\hat{\beta} = D(\hat{\gamma})$ of a series regression as estimators of $\beta_0 = D(\gamma_0)$, where $\sqrt{n}D(\hat{\gamma} - \gamma_0) = \sqrt{n}(\hat{\beta} - \beta_0)$. For example consider the average surplus bound estimator $\hat{\beta} = \int \omega(x)\hat{\gamma}(x)dx$. Here $\alpha(x) = \omega(x)/f_0(x_1|x_2)$ which will be discontinuous at the upper and lower prices and at a pair of income values when $\omega_2(x_2)$ is an indicator function

for an interval. If we assume that $\alpha(x)$ is Lipschitz where it is nonzero then $\bar{s}_\alpha = 1$. Since $r = 2$ in this example, the conditions for asymptotic linearity are

$$\sqrt{n}K^{-(\bar{s}_\gamma+1)/2} \longrightarrow 0, \sqrt{\frac{K\ln(K)}{n}}(1 + K^{(1-\bar{s}_\gamma)/2}).$$

There will exist $K = K_n$ such that this condition holds if and only if $\bar{s}_\gamma > 1/2$. This is a mild smoothness condition that will be satisfied if $\gamma(x)$ is also Lipschitz in between a finite number of discontinuity points. More generally there will exist K satisfying equation (6.25) for certain combinations of smoothness conditions on $\alpha(x)$ and $\gamma_0(x)$, as further discussed below.

The conditions of Theorem 8 and Corollary 9 improve significantly on those of Theorem 6.1 of Newey (1994) for linear functionals of series estimators. For example, for splines Assumption 6.6 of Newey (1994) requires that $K^3/n \longrightarrow 0$, which is stronger than the condition that $K^{-2\bar{s}_\gamma/r}K^2\ln(K)/n \longrightarrow 0$ in Corollary 9.

It is interesting to ask how small the remainder bounds in Corollary 9 are relative to any known remainder bounds for series estimators of a functional. We restrict attention to estimators where some unknown function must be consistently estimated in order for $\hat{\beta}$ to be consistent. For example we do not consider optimal instrumental variables estimation. The smallest known bounds are for a series estimator $\hat{\beta}$ of β_0 in the partially linear regression $E[y_i|w_i, x_i] = w_i^T\beta_0 + \gamma_0(x_i)$. Donald and Newey (1994) show that

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i) + O_p(\sqrt{n}K^{-(\bar{s}_\gamma+\bar{s}_\alpha)/r} + K^{-\bar{s}_\gamma/r} + K^{-\bar{s}_a/r} + \sqrt{\frac{K}{n}}), \qquad (6.26)$$

under weaker regularity conditions than those for Corollary 9 (that allow $E[p(x_i)p(x_i)^T]$ to be singular), where the form of $\hat{\beta}$ and $\psi(z_i)$ are given in Donald and Newey (1994), $\bar{s}_\alpha = \min\{s_\alpha, \kappa + 1\}$, and s_α is the Holder smoothness of $\alpha(x_i) = E[w_i|x_i]$. The last term in the remainder bound in Corollary 9 is larger for two reasons. One reason is the presence of $\ln(K)$. The presence of $\ln(K)$ will not increase the remainder much because it grows slowly with K. The second reason is the presence of $K^{(1/2)-(\bar{s}_\gamma/r)}$. When $\bar{s}_\gamma \geq r/2$ the $K^{(1/2)-(\bar{s}_\gamma/r)}$ term will be bounded, so that the conclusion of Corollary 9 becomes

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i) + O_p(\sqrt{n}K^{-(\bar{s}_\gamma+\bar{s}_\alpha)/r} + K^{-\bar{s}_\gamma/r} + K^{-\bar{s}_a/r} + \sqrt{\frac{K\ln(K)}{n}}).$$

Here the remainder term is the same size as Donald and Newey (1994), except for the (relatively small) $\ln(K)$ term. In this sense Corollary 9 has a remainder that is nearly as small as known possible when $\gamma_0(x)$ is smooth enough. It is true that when $\bar{s}_\gamma < r/2$ the remainder in Corollary 9, will be larger, but the small size of the remainder when when $\bar{s}_\gamma \geq r/2$ is an important improvement on existing results.

The additional term $K^{(1/2)-(\bar{s}_\gamma/r)}$ will be important for the minimal conditions for root-n consistency allowed by Corollary 9. A linear combination of $\bar{s}_\gamma$ and $\bar{s}_\alpha$ must be large enough in order that there exist K where the remainder term in Corollary 9 goes to zero. Ignoring the $\ln(K)$ term a choice of K_n with $nK_n^{-2(\bar{s}_\gamma+\bar{s}_\alpha)/r}$ proportional to $K_n^{-2\bar{s}_\gamma/r}K_n^2/n$ will asymptotically maximize the rate at which the remainder in Corollary 9 goes to zero when $\bar{s}_\gamma \leq r/2$. Such a K_n is given by $K_n = n^{r/(r+\bar{s}_\alpha)}$. Note that $n(K_n)^{-2(\bar{s}_\gamma+\bar{s}_\alpha)/r} = n^{1-2(\bar{s}_\gamma+\bar{s}_\alpha)/(r+\bar{s}_\alpha)}$, which goes to zero if and only if

$$2\bar{s}_\gamma + \bar{s}_\alpha > r. \tag{6.27}$$

Note also that this condition is automatically satisfied when $\bar{s}_\gamma > r/2$. Thus this is the minimal smoothness condition for existence of K such that the remainder term of Corollary 9 vanishes asymptotically. These necessary smoothness conditions are slightly stronger than the corresponding conditions from Donald and Newey (1994), which are $2\bar{s}_\gamma + 2\bar{s}_\alpha > r$. Conditions exactly analogous to $2s_\gamma + 2s_\alpha > r$ are known to be a minimal smoothness condition for existence of any root-n consistent estimators in some related settings, see Ritov and Bickel (1988) and Robins et. al. (2009).

The estimator $\hat{\beta} = D(\hat{\gamma})$ may be asymptotically linear when K is chosen to maximize the rate at which the mean square error of $\hat{\gamma}_0(x)$ goes to zero. Setting $K^{-2\bar{s}_\gamma/r}$ proportional to K/n is such a choice of K, giving $K_n = n^{r/(r+2\bar{s}_\gamma)}$. In this case Corollary 9 gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i) + O_p(n^{(1/2)-(\bar{s}_\gamma+\bar{s}_\alpha)/(r+2\bar{s}_\gamma)} + n^{(1-\bar{\gamma})r/(r+2\bar{s}_\gamma)-(1/2)}).$$

Here the remainder term goes to zero for $\bar{s}_\gamma > r/2(1+r)$ and $\bar{s}_\alpha > r/2$, a stronger condition for $\bar{s}_\gamma$ and the same condition for $\bar{s}_\alpha$ as in Donald and Newey (1994).

7 Conclusion

In this paper we have introduced a Gateaux derivative limit formula for the influence function and used it to characterize the influence function of M and GMM estimators. We found that this approach is very useful for NP2SLS where we have derived the adjustment term for first step estimation. We also used the formula and the remainder that it implies to specify regularity conditions for asymptotic linearity. These included conditions for first step series regression that are nearly as weak as known to be possible when the regression function is smooth enough.

In further work we have used the influence function to construct estimators with small bias, see Chernozhukov, Escanciano, Ichimura, and Newey (2016). For this and other purposes it

would be interesting to develop additional methods for estimating the influence function. We leave this topic to future work.

8 Appendix A: Proofs

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

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

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

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

$$\int 1(\tilde{a}\in\mathcal{A}) f_\tau(\tilde{a}) d\mu = E[1_i] + \tau E[1_i E[S(z_i)|a_i]] = E[1_i] + \tau E[1_i S(z_i)] = E_\tau[1_i],$$

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

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

$$= E[1_i b(z_i)\{1+\tau S(z_i)\}] = E_\tau[1_i b(z_i)]. \qquad Q.E.D.$$

Proof of Theorem 1: Note that by $S(\tilde{z})$ bounded there is an open set T containing zero such that for all $\tau \in T$, $1+\tau S(\tilde{z})$ is positive, bounded away from zero, and $f_\tau(\tilde{z})^{1/2} = f_0(\tilde{z})^{1/2}[1+\tau S(\tilde{z})]^{1/2}$ is continuously differentiable in τ with

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

By $S(\tilde{z})$ bounded, $\int \left[C f_0(\tilde{z})^{1/2} S(\tilde{z})\right]^2 d\mu < \infty$. Then by the dominated convergence theorem $f_0(\tilde{z})^{1/2}[1+\tau S(\tilde{z})]^{1/2}$ is mean-square differentiable and $I(\tau) = \int s_\tau(\tilde{z})^2 d\mu$ is continuous in τ on a neighborhood of zero. Next, consider $\delta(\tilde{z})$ as specified in eq. (3.12). Take j large enough so $g(z_i)$ is equal to zero with positive probability and $g(z_i)$ are both nonzero on a neighborhood of z, so that $\delta(z_i)$ is not constant. Then $S(z_i)$ is not zero so that $I(\tau) > 0$. Then by Theorem 7.2 and Example 6.5 of Van der Vaart (1998) it follows that for any $\tau_n = O(1/\sqrt{n})$ a vector of n observations $(z_1, ..., z_n)$ that is i.i.d. with pdf $f_{\tau_n}(\tilde{z})$ is contiguous to a vector of n observations with pdf $f_0(\tilde{z})$. Therefore,

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

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

Next define $\mu_z^j = E[\psi(z_i)S(z_i)] = E[\psi(z_i)\delta(z_i)]$. Then by $E[\psi(z_i)] = 0$,

$$E_\tau[\psi(z_i)] = \tau\mu_z^j.$$

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

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

Note that $E_{\tau_n}[\breve{\psi}_n(z_i)] = 0$. Also, by τ_n bounded,

$$\begin{aligned}E_\tau[1(\left\|\breve{\psi}_n(z_i)\right\| &\geq M)\left\|\breve{\psi}_n(z_i)\right\|^2] \leq CE[1(\left\|\breve{\psi}_n(z_i)\right\| \geq M)\left\|\breve{\psi}_n(z_i)\right\|^2] \\ &\leq CE[1(\left\|\breve{\psi}_n(z_i)\right\| \geq M)(\|\psi(z_i)\|^2 + C)] \\ &\leq CE[1(\|\psi(z_i)\| \geq M - C)(\|\psi(z_i)\|^2 + C)] \longrightarrow 0,\end{aligned}$$

as $M \longrightarrow \infty$, so the Lindbergh-Feller condition for a central limit theorem is satisfied. Furthermore, it follows by similar calculations that $E_{\tau_n}[\breve{\psi}_n(z_i)\breve{\psi}_n(z_i)^T] \longrightarrow V$. Therefore, by the Lindbergh-Feller central limit theorem, $\sum_{i=1}^n \breve{\psi}_n(z_i)/\sqrt{n} \xrightarrow{d} N(0, V)$. Then$\sqrt{n}(\hat{\beta} - \beta_n) \xrightarrow{d} N(0, V)$ implies that

$$\sqrt{n}\tau_n\mu_z^j - \sqrt{n}(\beta_n - \beta_0) \longrightarrow 0. \tag{8.28}$$

Next, we follow the proof of Theorem 2.1 of Van der Vaart (1991). The above argument shows that local regularity implies that eq. (8.28) holds for all $\tau_n = O(1/\sqrt{n})$. Consider any sequence $r_m \longrightarrow 0$. Let n_m be the subsequence such that

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

Let $\tau_n = r_m$ for $n = n_m$ and $\tau_n = n^{-1/2}$ for $n \notin \{n_1, n_2, ...\}$. By construction, $\tau_n = O(1/\sqrt{n})$, so that eq (8.28) holds. Therefore it also holds along the subsequence n_m, so that

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

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

Next note that by the construction of $g(\tilde{z})$ for large enough j the function $g(\tilde{z})$ will be zero outside the neighborhood $\mathcal{N}$ in the statement of Theorem 1 and $1/j < \varepsilon$. For such large j,

$$\mu_z^j = E[\psi(z_i)\delta(z_i)] = \int_{\mathcal{N}} \psi(\tilde{z})\delta(\tilde{z})f_0(\tilde{z})d\mu = \int_{\bar{\mathcal{N}}} \psi(\tilde{z})g(\tilde{z})d\mu.$$

Consider any $\zeta > 0$. By continuity of $\psi(\tilde{z})$ at z on $\bar{\mathcal{N}}$ and the construction of $g(\tilde{z})$, for large enough j we will have $\|\psi(\tilde{z}) - \psi(z)\| < \zeta$ for all $\tilde{z} \in \bar{\mathcal{N}}$ where $g(\tilde{z}) > 0$. Then by $\int_{\bar{\mathcal{N}}} g(\tilde{z})d\mu = 1$,

$$\left\|\mu_z^j - \psi(z)\right\| = \|E[\psi(z_i)\delta(z_i)] - \psi(z)\| = \left\|\int_{\bar{\mathcal{N}}} [\psi(\tilde{z}) - \psi(z)]g(\tilde{z})d\mu\right\| \leq \int_{\bar{\mathcal{N}}} \|\psi(\tilde{z}) - \psi(z)\| \, g(\tilde{z})d\mu \leq \zeta.$$

Therefore $\lim_{j\longrightarrow\infty} \mu_z^j = \psi(z)$. *Q.E.D.*

Proof of Lemma 2: Define $\bar{\Delta}_j = E[\Delta_j(w_i)]$, $\delta_x(\tilde{x}) = E[\Delta_j(w_i)|x_i = \tilde{x}]/\bar{\Delta}_j$, and $\breve{\Delta}_j = \int \Delta_j(w)g(y,w)d\mu$. By Assumption 1 b) we have $f(y,w|x) \geq 1/j$ for all (y,w,x) with $g(y,w) > 0$ when j is large enough. For all such j

$$\delta(\tilde{z}) = f(\tilde{y},\tilde{w}|\tilde{x})^{-1}g(\tilde{y},\tilde{w})\delta_x(\tilde{x}).$$

Note that $\delta(z_i)$ is bounded by condition b) and by $\Delta_j(w_i)$ bounded. Also

$$E[\delta(z_i)\Delta_j(w_i)|x_i] = \breve{\Delta}_j\delta_x(x_i).$$

We will first prove that for j large enough $E_\tau[\Delta(w_i)|x_i]$ is complete as a function of $\Delta(w_i)$ for all τ small enough. Consider $\tau < \bar{\tau}$ where $\bar{\tau}$ is chosen so that $\bar{\Delta}_j + \tau(\breve{\Delta}_j - \bar{\Delta}_j) \neq 0$ for any $\tau < \bar{\tau}$. Consider $\Delta(w_i)$ with $E[\Delta(w_i)^2] < \infty$ and $E_\tau\left[\Delta(w_i)|x_i\right] = 0$. Note that $E[\Delta(w_i)\delta(z_i)|x_i]$ is finite with probability one by $\delta(z_i)$ bounded and

$$E[\Delta(w_i)\delta(z_i)|x_i] = \breve{\Delta}\delta_x(x_i), \breve{\Delta} = \int \Delta(\tilde{w})g(\tilde{y},\tilde{w})d\mu,$$

so that $\breve{\Delta}$ exists. Then by eq. (3.10),

$$0 = (1-\tau)E\left[\Delta(w_i)|x_i\right] + \tau\breve{\Delta}\delta_x(x_i). \tag{8.29}$$

If $\breve{\Delta} = 0$ note that $(1-\tau)E[\Delta(w_i)|x_i] = 0$ implying $\Delta(w_i) = 0$ by Assumption 1 a). Suppose $\breve{\Delta} \neq 0$. Then dividing through eq. (8.29) by $(1-\tau)$ and choosing $C = \tau\breve{\Delta}/[(1-\tau)\bar{\Delta}_j]$ we have, by the definition of $\delta_x(x_i)$,

$$0 = E[\Delta(w_i)|x_i] + CE[\Delta_j(w_i)|x_i] = E[\Delta(w_i) + C\Delta_j(w_i)|x_i].$$

Assumption 1 a) then implies $\Delta(w_i) = -C\Delta_j(w_i)$. Then $E_\tau[\Delta(w_i)|x_i] = 0$ and $C \neq 0$ implies $E_\tau[\Delta_j(w_i)|x_i] = 0$, which together with eq. (3.10) implies

$$0 = (1-\tau)E[\Delta_j(w_i)|x_i] + \tau\breve{\Delta}_j\delta_x(x_i) = [(1-\tau)\bar{\Delta}_j + \tau\breve{\Delta}_j]\delta_x(x_i).$$

by the definition of $\delta_x(x_i)$. Since $\bar{\Delta}_j \neq 0$ we must have $\Delta_j(w_i) \neq 0$ and hence $\delta_x(x_i) \neq 0$ by Assumption 1 a). Then the previous equation implies

$$0 = (1-\tau)\bar{\Delta}_j + \tau\breve{\Delta}_j = \bar{\Delta}_j + \tau(\breve{\Delta}_j - \bar{\Delta}_j),$$

contradicting the choice of τ. Therefore we have contradicted $\breve{\Delta} \neq 0$, so that we must have $\breve{\Delta} = 0$, and hence $\Delta(w_i) = 0$.

Next consider $\gamma(w_i, c) = \gamma_0(w_i) + c\Delta_j(w_i)$ for another constant c. Note that for $\bar{y} = \int \tilde{y} g(\tilde{y}, \tilde{w}) d\mu$ and $\bar{\gamma} = \int \gamma_0(\tilde{w}) g(\tilde{y}, \tilde{w}) d\mu$,

$$E[\delta(z_i) y_i | x_i] = \bar{y}\delta_x(x_i),\ E[\delta(z_i)\gamma_0(w_i)|x_i] = \bar{\gamma}\delta_x(x_i).$$

By completeness of $E_\tau[\Delta(w_i)|x_i]$ as a function of Δ and eq. (3.10), $E_\tau[y_i|x_i] = E_\tau[\gamma(w_i, c)|x_i]$ if and only if

$$\begin{aligned}
(1-\tau)E[y_i|x_i] + \tau\bar{y}\delta_x(x_i) &= (1-\tau)E[y_i|x_i] + \tau E[\delta(z_i)y_i|x_i] \\
&= (1-\tau)E[\gamma(w_i,c)|x_i] + \tau E[\delta(z_i)\gamma(w_i,c)|x_i] \\
&= (1-\tau)E[\gamma_0(x_i)|x_i] + (1-\tau)c\bar{\Delta}_j\delta_x(x_i) + \tau\bar{\gamma}\delta_x(x_i) + \tau c\breve{\Delta}_j\delta_x(x_i).
\end{aligned}$$

Noting that $E[y_i|x_i] = E[\gamma_0(w_i)|x_i]$ and $\delta_x(x_i) \neq 0$, this equation holds if any only if

$$\tau\bar{y} = (1-\tau)\bar{\Delta}_j c + \tau\bar{\gamma} + \tau c\breve{\Delta}_j = 0.$$

We can solve this equation for $c = c_j(\tau)$ to obtain

$$c_j(\tau) = \tau(\bar{y} - \bar{\gamma})/[(1-\tau)\bar{\Delta}_j + \tau\breve{\Delta}_j], \frac{\partial c_j(\tau)}{\partial \tau} = (\bar{y} - \bar{\gamma})/\bar{\Delta}_j.$$

Finally, let $\gamma_\tau(\tilde{w}) = \gamma_0(\tilde{w}) + c_j(\tau)\Delta_j(\tilde{w})$. We have $E_\tau[y_i - \gamma_\tau(w_i)|x_i] = 0$ by construction and $E_\tau[y_i - \gamma(w_i)|x_i] \neq 0$ for $\gamma(w_i) \neq \gamma_\tau(w_i)$ by completeness of $E_\tau[\Delta(w_i)|x_i]$. It therefore follows as in NP that $Q_\tau(\gamma)$ has a unique minimum at γ_τ. *Q.E.D.*

Proof of Theorem 3: By $\alpha(\tilde{x})$ continuous at x and $f_0(\tilde{x}) > 0$ positive in a neighborhood of x, there are bounded functions $\delta_j(x)$ such that $E[\delta_j(x)] = 1$ and $E[\alpha(x_i)\delta_j(x_i)] \longrightarrow \alpha(x)$. By completeness of $E[\tilde{\Delta}(x_i)|w_i]$ the mean square closure of the range of $E[\Delta(w_i)|x_i]$ is all functions of x_i with finite mean square. Because the set of bounded functions is mean square dense, for each j there exists bounded $\Delta_j(w)$ such that $E[\{\delta_j(x_i) - \delta_x(x_i)\}^2] \longrightarrow 0$, so that by the triangle and Cauchy-Schwartz inequalities

$$|\alpha(x) - E[\alpha(x_i)\delta_x(x_i)]| \leq |\alpha(x) - E[\alpha(x_i)\delta_j(x_i)]| + |E[\alpha(x_i)\{\delta_j(x_i) - \delta_x(x_i)\}]| \longrightarrow 0.$$

The proof then follows as in the text. *Q.E.D.*

Proof of Lemma 5: Note that for any $b(z) \geq 0$ we have $E_\tau [b(z_i)] = (1-\tau)E [b(z_i)] + \tau E [b(z_i)\delta(z_i)]$ so that

$$(1-\tau) E [b(z_i)] \leq E_\tau [b(z_i)] \leq [(1-\tau) + C] E [b(z_i)]$$

Apply this inequality to $b(z_i) = \left[a(z_i) - c^T p^K(x_i)\right]^2$ for each K to get $\mathcal{A}_0 = \mathcal{A}_\tau$. Next, note that for any $a(x) \in \mathcal{A}_\tau = \mathcal{A}_0$,

$$E_\tau [a(x_i) \{b(z_i) - \pi_\tau(b, x_i)\}] = 0 = E [a(x_i) \{b(z_i) - \pi_\tau(b, x_i)\}] + \tau E [a(x_i) \{b(z_i) - \pi_\tau(b, x_i)\} S(z_i)] .$$

It follows that

$$E [a(x_i) \{\pi_0(b, x_i) - \pi_\tau (b, x_i)\}] = E [a(x_i) \{b(z_i) - \pi_\tau (b, x_i)\}] = -\tau E [a(x_i) \{b(z_i) - \pi_\tau (b, x_i)\} S(z_i)]$$

Plugging in $a(x_i) = \pi_0(b, x_i) - \pi_\tau (b, x_i) = \Delta_\tau(x_i) \in \mathcal{A}_0$, it follows from the Cauchy Schwartz inequality and $S(z_i)$ bounded that

$$E \left[\Delta_\tau(x_i)^2\right] = -\tau E [\Delta_\tau(x_i) \{b(z_i) - \pi_\tau(b, x_i)\} S(z_i)] \leq \tau \left\{E \left[\Delta_\tau(x_i)^2\right]\right\}^{1/2} E \left[\{b(z_i) - \pi_\tau (b, x_i)\}^2\right] .$$

Dividing through by $\left\{E \left[\Delta_\tau (x_i)^2\right]\right\}^{1/2}$ it follows that $E \left[\Delta_\tau (x_i)^2\right] \longrightarrow 0$ as $\tau \longrightarrow 0$. Therefore, as $\tau \longrightarrow 0$

$$\begin{aligned} \frac{E [a(x_i)\pi_\tau(b, x_i)] - E [a(x_i)\pi_0(b, x_i)]}{\tau} &= E [a(x_i) \{b(z_i) - \pi_\tau (b, x_i)\} S(z_i)] \\ &= E [a(x_i) \{b(z_i) - \pi_0 (b, x_i)\} S(z_i)] + o(1). Q.E.D. \end{aligned}$$

Proof of Proposition 6: Note first that by Γ linear we have $\Delta' \in \Gamma$. Then by Assumption 4 c)

$$\begin{aligned} \frac{\partial E[m(z_i, \beta_0, \gamma_\tau)]}{\partial \tau} &= E[v(w_i)\Delta'(w_i)] = E[v^*(w_i)\Delta'(w_i)] = -E[\pi^*(d_0\alpha, w_i)\Delta'(w_i)] \quad (8.30) \\ &= -E[d_0(w_i, x_i)\alpha(x_i)\Delta'(w_i)] = -E[\alpha(x_i)\pi_0(d_0\Delta', x_i)] \\ &= -E[\alpha^*(x_i)\pi_0(d_0\Delta', x_i)], \end{aligned}$$

where the first equality follows by Assumption 3, the second by $\Delta' \in \Gamma$ and v^* being the projection of v on Γ, the third by Assumption 4 c), the fourth by $\Delta' \in \Gamma$ and π^* being the projection on Γ, the fifth by $\alpha \in \mathcal{A}_0$ and π_0 being the projection on $\mathcal{A}_0$, and the sixth by $\pi_0(d_0\Delta', x_i) \in \bar{\mathbf{A}}$ and $\alpha^*(x_i)$ being the projection of $\alpha(x_i)$ on $\bar{\mathbf{A}}$. By $\alpha^*(x_i) \in \bar{\mathbf{A}}$ there is Δ^k such that $\pi_0(d_0\Delta^k, x_i) \longrightarrow \alpha^*(x_i)$ in $\mathcal{H}_0$. Then by $S(z_i)$ bounded

$$\begin{aligned} E[\pi_0(d_0\Delta^k, x_i)\rho(z_i, \gamma_0)S(z_i)] &\longrightarrow E[\alpha^*(x_i)\rho(z_i, \gamma_0)S(z_i)], \\ E[\pi_0(d_0\Delta^k, x_i)\pi_0(d_0\Delta', x_i)] &\longrightarrow E[\alpha^*(x_i)\pi_0(d_0\Delta', x_i)]. \end{aligned}$$

Then for $\pi(\rho(\gamma_0), x_i) = 0$ equation (5.23) gives $E[\alpha^*(x_i)\pi_0(d_0\Delta', x_i)] = -E[\alpha^*(x_i)\rho(z_i, \gamma_0)S(z_i)]$. Combining this result with equation (8.30) we obtain

$$\frac{\partial E[m(z_i, \beta_0, \gamma_\tau)]}{\partial \tau} = E[\alpha^*(x_i)\rho(z_i, \gamma_0)S(z_i)].$$

Taking limits as $j \longrightarrow \infty$ we find that for for $\pi(\rho(\gamma_0), x_i) = 0$ the correction term is $\alpha^*(x)\rho(z, \gamma_0)$.

Next consider the case where $\pi(\rho(\gamma_0), x_i) \neq 0$ but $\partial\pi_0(d_\tau\Delta, x_i)/\partial\tau = 0$ and there exists $\Delta^*(w_i)$ with $\alpha^*(x_i) = -\pi_0(d_0\Delta^*, x_i)$. Then combining equations (5.23) and (8.30)

$$\begin{aligned} \frac{\partial E[m(z_i, \beta_0, \gamma_\tau)]}{\partial \tau} &= E[\phi(z_i)S(z_i)], \qquad (8.31) \\ \phi(z_i) &= \alpha^*(x)[\rho(z, \gamma_0) - \pi_0(\rho(\gamma_0), x)] - \pi_0(\rho(\gamma_0), x)[d_0(w, x)\Delta^*(w) + \alpha^*(x)] \\ &\quad + \alpha^*(x)\pi_0(\rho(\gamma_0), x) - E[\alpha^*(x_i)\pi_0(\rho(\gamma_0), x_i)].Q.E.D. \end{aligned}$$

Proof of Theorem 7: Consistency of $\hat{\beta}$ for β_0 in the interior of the parameter space implies that with probability approaching one (w.p.a.1) $\hat{\beta}$ will satisfy the first-order condition

$$0 = \hat{M}(\hat{\beta})^T\hat{W}\hat{m}(\hat{\beta}).$$

For any $\bar{\beta} \xrightarrow{p} \beta_0$ we will have $\hat{M}(\hat{\beta}) \xrightarrow{p} M = \partial E[m(z_i, \beta, \gamma_0]/\partial\beta|_{\beta=\beta_0}$, $\hat{M}(\bar{\beta}) \xrightarrow{p} M$, and $\hat{W} \xrightarrow{p} W$, so $\hat{M}(\hat{\beta})^T\hat{W}\hat{M}(\bar{\beta}) \xrightarrow{p} M^TWM$. Nonsingularity of M^TWM will then imply that $\hat{M}(\hat{\beta})^T\hat{W}\hat{M}(\bar{\beta})$ is nonsingular w.p.a.1 and $[\hat{M}(\hat{\beta})^T\hat{W}\hat{M}(\bar{\beta})]^{-1} \xrightarrow{p} (M^TWM)^{-1}$. By $\hat{R}_1 \xrightarrow{p} 0$ and Assumptions 5 and 6, the data i.i.d., and the Markov inequality we have

$$\sqrt{n}\hat{m}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(z_i, \beta_0, \gamma_0) + \phi(z_i)] + o_p(1) = O_p(1).$$

Expanding $\hat{m}(\hat{\beta})$ in β around β_0 and solving the first-order condition for $\sqrt{n}(\hat{\beta} - \beta_0)$ gives

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta_0) &= -[\hat{M}(\hat{\beta})^T\hat{W}\hat{M}(\bar{\beta})]^{-1}\hat{M}(\hat{\beta})^T\hat{W}\sqrt{n}\hat{m}(\beta_0) \qquad (8.32) \\ &= -(M^TWM)^{-1}M^TW\sqrt{n}\hat{m}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i) + o_p(1), \\ \psi(z) &= -(M^TWM)^{-1}M^TW[m(z, \beta_0, \gamma_0) + \phi(z)], \end{aligned}$$

where $\bar{\beta}$ is an intermediate value that lies on the line joining $\hat{\beta}$ and β_0 (and so is consistent for β_0) and actually differs from row-to-row of $\hat{M}(\beta)$. *Q.E.D.*

Proof of Theorem 8: It follows by Assumption 8 and Cauchy-Schwartz that

$$\left|\hat{R}_{31}\right| \leq \sqrt{n}c_K^\alpha c_K.$$

Next, by $\xi_K^2 \ln(K)/n \longrightarrow 0$ it follows as in Belloni et. al. (2015) that $\hat{\Sigma}$ is nonsingular with probability approaching one. Then since everything in $\hat{R}_{31}$ and $\hat{R}_{32}$ is invariant to nonsingular linear transformations of $p^K(x)$ it can be assumed without loss of generality that $\Sigma = E[p^K(x_i)p^K(x_i)^T] = I$. Note that $\hat{R}_{32} = \hat{R}_{321} + \hat{R}_{322}$ where

$$\begin{aligned}\hat{R}_{321} &= v^T \sum_{i=1}^n p^K(x_i)[\gamma_0(x_i) - \gamma_K(x_i)]/\sqrt{n} = \sum_{i=1}^n \alpha_K(x_i)[\gamma_0(x_i) - \gamma_K(x_i)]/\sqrt{n}, \\ \hat{R}_{322} &= v^T(\hat{\Sigma}^{-1} - I) \sum_{i=1}^n p^K(x_i)[\gamma_0(x_i) - \gamma_K(x_i)]/\sqrt{n}.\end{aligned}$$

Note that $E[\alpha_K(x_i)\{\gamma_0(x_i) - \gamma_K(x_i)\}] = 0$ and that

$$E[\alpha_K(x_i)^2\{\gamma_0(x_i) - \gamma_K(x_i)\}^2] \leq c_K^2 \ell_K^2 E[\alpha_K(x_i)^2] \leq c_K^2 \ell_K^2 C.$$

Therefore we have

$$\hat{R}_{321} = O_p(c_K \ell_K).$$

Also, note that $v^T v = E[\alpha_K(x_i)^2] \leq E[\alpha(x_i)^2]$ and that

$$E[\left\| \sum_{i=1}^n p^K(x_i)[\gamma_0(x_i) - \gamma_K(x_i)]/\sqrt{n} \right\|^2] \leq E[p^K(x_i)^T p^K(x_i)\{\gamma_0(x_i) - \gamma_K(x_i)\}^2] \leq K c_K^2 \ell_K^2.$$

Then it follows similarly to eqs. (4.12) and (4.14) of Lemma 4.1 of Belloni et. al. (2015) that

$$\hat{R}_{322} = O_p(\sqrt{\frac{\xi_K^2 \ln(K)}{n}} \sqrt{K} c_K \ell_K)$$

Next, by Assumption 8, $E[\{\alpha_K(x_i) - \alpha(x_i)\}^2] = O([c_K^\alpha]^2)$. Then by $Var(y_i|x_i)$ bounded and the Markov inequality,

$$\begin{aligned}\sum_{i=1}^n \{\tilde{\alpha}(x_i) - \alpha(x_i)\}^2 Var(y_i|x_i)/n &\leq C \sum_{i=1}^n \{\tilde{\alpha}(x_i) - \alpha(x_i)\}^2/n \\ &\leq C \sum_{i=1}^n \{\alpha_K(x_i) - \alpha(x_i)\}^2/n + C \sum_{i=1}^n \{v^T(\hat{\Sigma}^{-1} - I)p^K(x_i)\}^2/n \\ &= O_p([c_K^\alpha]^2) + v^T(\hat{\Sigma}^{-1} - I)\hat{\Sigma}(\hat{\Sigma}^{-1} - I)v = O_p([c_K^\alpha]^2 + \frac{\xi_K^2 \ln(K)}{n}),\end{aligned}$$

where the last equality follows as in Step 1 of the proof of Lemma 4.1 of Belloni et. al. (2015). The conclusion now follows by the triangle inequality which gives

$$\hat{R}_3 = O_p(\sqrt{n} c_K^\alpha c_K + c_K \ell_K + c_K^\alpha + \sqrt{\frac{\xi_K^2 \ln(K)}{n}}(1 + \sqrt{K} c_K \ell_K)).$$

9 Appendix B: Comparison of the Average Derivative Variance with Ai and Chen (2007).

In this Appendix we show that if $E[\tilde{\Delta}(x_i)|w_i]$ is complete as a function of $\tilde{\Delta}$, as holds generically in the setting of Section 4, then the asymptotic variance for the average derivative in equation (4.18) is the same as on p. 25 of Ai and Chen (2007). The asymptotic variance corresponding to the influence function of the average derivative in Section 4 is

$$E[\psi(z_i)^2] = E[\left\{ \bar{v}(w_i)\frac{\partial \gamma_0(w_i)}{\partial w_k} - \beta_0 + \alpha(x_i)[y_i - \gamma_0(w_i)] \right\}^2].$$

The Ai and Chen (2007) expression for the variance of the average derivative estimator is given on their p. 25 as

$$\begin{aligned} \Omega^*/C^2 &= E[\{\theta^* - a(Y_2)\nabla^s h_*(Y_2) + C^{-1}E[w^*(Y_2)|X_1](Y_1 - h_*(Y_2))\}^2] \\ C &= 1 + E[a(Y_2)\nabla^s w^*(Y_2)]. \end{aligned}$$

This expression is identical to $E[\psi(z_i)^2]$ if

$$\begin{aligned} y_i &= Y_1, w_i = Y_2, X_1 = x_i, \theta^* = \beta_0, a(Y_2) = \bar{v}(w_i), \nabla^s h_*(Y_2) = \frac{\partial \gamma_0(w_i)}{\partial w_j}, \\ \gamma_0(w_i) &= h_*(Y_2), \alpha(x_i) = -C^{-1}E[w^*(Y_2)|X_1]. \end{aligned}$$

All of these equalities except the last one hold by the respective definitions in the papers. To show the last equality let $\tilde{w}(Y_2) = -C^{-1}w^*(Y_2)$. Consider equation (20) in Ai and Chen (2007). Divide through by $-C$ and integrate the last expectation in equation (20) by parts with $\nabla^s = \partial/\partial w_j$ to get

$$\begin{aligned} 0 &= E[E[\tilde{w}(Y_2)|X_1]E[\delta(Y_2)|X_1]] - E[v(Y_2)\delta(Y_2)] = E[E[\tilde{w}(Y_2)|X_1]\delta(Y_2)] - E[v(Y_2)\delta(Y_2)] \\ &= E[\{E[E[\tilde{w}(Y_2)|X_1]|Y_2] - v(Y_2)\}\delta(Y_2)] \end{aligned}$$

for all $\delta(Y_2)$ with finite second moment. This equation can only hold if

$$E[E[\tilde{w}(Y_2)|X_1]|Y_2] = v(Y_2).$$

Since also $E[\alpha(X_1)|Y_2] = v(Y_2)$, it follows by completeness that $\alpha(X_1) = E[\tilde{w}(Y_2)|X_1]$, and hence the variance expressions are equal to each other.

As noted in Section 4 above the condition that $\alpha(X_1) = E[\tilde{w}(Y_2)|X_1]$ means that $\alpha(X_1)$ is restricted to be smooth in a way similar to $v(w)$ being restricted to be smooth.

10 References

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

Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica* 71, 1795–1843.

Ai, C. and X. Chen (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," *Journal of Econometrics* 141, 5–43.

Ai, C. and X. Chen (2012): "The Semiparametric Efficiency Bound for Models of Sequential Moment Restrictions Containing Unknown Functions," *Journal of Econometrics* 170, 442–457.

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

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

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

Bajari, P., V. Chernozhukov, H. Hong, and D. Nekipelov (2009): "Nonparametric and Semiparametric Analysis of a Dynamic Discrete Game," working paper, Stanford.

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

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

Bickel, P, C. Klaasen, Y. Ritov, and J. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models,* Washington, Johns Hopkins.

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

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

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

Carone, M., A.R. Luedtke, M.J. van der Laan (2016): "Toward computerized efficient estimation in infinite dimensional models," arXiv: 1608.08717v1 31Aug2016.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," Chapter 76, *Handbook of Econometrics*.

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

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

Chen, X., and Z. Liao (2015): "Sieve Semiparametric GMM Under Weak Dependence," *Journal of Econometrics* 189, 163-186.

Chen, X. and A. Santos (2015): "Overidentification in Regular Models," Cowles Foundation Discussion Paper No. 1999.

Chernozhukov, V. and C. Hansen (2004): "An IV Model of Quantile Treatment Effects," *Econometrica* 73, 245–261.

Chernozhukov, V., G. Imbens, and W. Newey (2007): "Instrumental Variables Estimation of Nonseparable Models," *Journal of Econometrics* 139, 4–14.

Chernozhukov, Escanciano, Ichimura, Newey (2016a): "Locally Robust Semiparametric Estimation," ArXIV, July 2016.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2016b): "Double Machine Learning for Treatment and Causal Parameters," MIT working paper.

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

Donald, S.G. and W.K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis* 50, 30-40.

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

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

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

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

Hahn, J. and G. Ridder (2016): "Three-stage Semi-Parametric Inference: Control Variables and Differentiability," working paper.

Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50, 1029-1054.

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

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

Hausman, J.A. and W.K. Newey (2016b): "Nonparametric Welfare Analysis," *Annual Review of Economics,* forthcoming.

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

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-index Models," *Journal of Econometrics* 58, 71-120.

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

Klein, R.W. and R.H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61, 387-421.

Kress, R. (1989): *Linear Integral Equations,* New York: Springer-Verlag.

Luedtke, A.R., M. Carone, M.J. van der Laan (2015): "Toward computerized efficient estimation in infinite dimensional models," arXiv: 1608.08717v1 31Aug2016.

Mammen, E., C. Rothe, M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics* 40, 1132–1170.

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

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

Newey, W.K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.

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., and J.L. Powell (1989): "Instrumental Variable Estimation of Nonparametric Models," presented at Econometric Society winter meetings, 1988.

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

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* 57, 1027-1057.

Powell, J.L., J.H. Stock, and T.M. Stoker (1989): "Semiparametric Estimation of Index

Coefficients," *Econometrica* 57, 1403-1430.

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

Ritov, Y. and P.J. Bickel (1990): "Achieving Information Bounds in Non and Semiparametric Models," *Annals of Statistics* 18, 925-938.

Robins, J.M., E.T. Tchetgen, L. Li, A. van der Vaart (2009): "Semiparametric Minimax Rates," *Electronic Journal of Statistics* 3, 1305–1321.

Robinson, P.M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica* 56, 931-954.

Santos, A. (2011): "Instrumental Variable Methods for Recovering Continuous Linear Functionals," *Journal of Econometrics* 161, 129-146.

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.

Sherman, R. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica* 61, 123-137.

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

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

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

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