Locally Robust Semiparametric Estimation

Victor Chernozhukov* Juan Carlos Escanciano† Hidehiko Ichimura‡
MIT *Indiana University* *University of Tokyo*

Whitney K. Newey§
MIT

July 27, 2016

Abstract

This paper shows how to construct locally robust semiparametric GMM estimators, meaning equivalently moment conditions have zero derivative with respect to the first step and the first step does not affect the asymptotic variance. They are constructed by adding to the moment functions the adjustment term for first step estimation. Locally robust estimators have several advantages. They are vital for valid inference with machine learning in the first step, see Belloni et. al. (2012, 2014), and are less sensitive to the specification of the first step. They are doubly robust for affine moment functions, where moment conditions continue to hold when one first step component is incorrect. Locally robust moment conditions also have smaller bias that is flatter as a function of first step smoothing leading to improved small sample properties. Series first step estimators confer local robustness on any moment conditions and are doubly robust for affine moments, in the direction of the series approximation. Many new locally and doubly robust estimators are given here, including for economic structural models. We give simple asymptotic theory for estimators that use cross-fitting in the first step, including machine learning.

Keywords: Local robustness, double robustness, semiparametric estimation, bias, GMM.
JEL classification: C13; C14; C21; D24.

*Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: vchern@mit.edu.
†Department of Economics, Indiana University, Bloomington, IN 47405–7104, U.S.A E-mail: jescanci@indiana.edu.
‡Faculty of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033. E-mail: ichimura@e.u-tokyo.ac.jp.
§Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: wnewey@mit.edu.

1 Introduction

There are many economic parameters that depend on nonparametric or large dimensional first steps. Examples include games, dynamic discrete choice, average consumer surplus, and treatment effects. This paper shows how to construct GMM estimators that are locally robust to the first step, meaning equivalently that moment conditions have a zero derivative with respect to the first step and that estimation of the first step does not affect their influence function.

Locally robust moment functions have several advantages. Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2014) showed that local robustness, also referred to as orthogonality, is important for correct inference about parameters of interest when machine learning is used in the first step. Locally robust moment conditions are also nearly correct when the nonparametric part is approximately correct. This robustness property is appealing in many settings where it may be difficult to get the first step completely correct. Furthermore, local robustness implies the small bias property analyzed in Newey, Hsieh, and Robins (1998, 2004; NHR henceforth). As a result asymptotic confidence intervals based on locally robust moments have actual coverage probability closer to nominal than for other moments. Also, bias is flatter as a function of first step bias for locally robust estimators than for other estimators. This tends to make their mean-square error (MSE) flatter as a function of smoothing also, so their performance is less sensitive to smoothing. In addition, by virtue of their smaller bias, locally robust estimators have asymptotic MSE that is smaller than other estimators in important cases and undersmoothing is not required for root-n consistency. Finally, asymptotic variance estimation is straightforward with locally robust moment functions, because the first step is already accounted for.

Locally robust moment functions are constructed by adding to the moment functions the terms that adjust for (or account for) first step estimation. This construction gives moment functions that are locally robust. It leads to new estimators for games, dynamic discrete choice, average surplus, and other important economic parameters. Also locally robust moments that are affine in a first step component are globally robust in that component, meaning the moments continue to hold when that component varies away from the truth. This result allows construction of doubly robust moments in the sense of Scharfstein, Rotnitzky, and Robins (1999) and Robins, Rotnitzky, and van der Laan (2000) by adding to affine moment conditions an affine adjustment term. Here we construct many new doubly robust estimators, e.g. where the first step solves a conditional moment restriction or is a density.

Certain first step estimators confer the small bias property on moment functions that are not locally robust, including series estimators of mean-square projections (Newey, 1994), sieve

maximum likelihood estimators (Shen, 1996, Chen and Shen, 1997), bootstrap bias corrected first steps (NHR), and higher-order kernels (Bickel and Ritov, 2003). Consequently the inference advantages of locally robust estimators may be achieved by using one of these first steps. These first steps only make moment conditions locally robust in certain directions. Locally robust moments have the small bias property in a wider sense, that the moments are nearly zero as the first step varies in a general way. This property is important when the first step is chosen by machine learning or in other very flexible, data based ways; see Belloni et. al. (2014).

First step series estimators have some special robustness properties. Moments without the adjustment term can be interpreted as locally robust because there is an estimated adjustment term with average that is identically zero. This property corresponds to first step series estimators conferring local robustness in the direction of the series approximation. Also, first step series estimators are doubly robust in those directions when the moment functions and first step estimating equations are affine.

The theoretical and Monte Carlo results of NHR show the bias and MSE advantages of locally robust estimators for linear functionals of a density. The favorable properties of a twicing kernel first step versus a standard first step found there correspond to favorable properties of locally robust moments versus original moments, because a twicing kernel estimator is numerically equivalent to adding an estimated adjustment term. The theoretical results show that using locally robust moment conditions increases the rate at which bias goes to zero but only raises the variance constant, and so leads to improved asymptotic MSE. The Monte Carlo results show that the MSE of the locally robust estimator is much flatter as a function of bandwidth and has a smaller minimum than the original moment functions, even with quite small samples. Advantages have also been found in the literature on doubly robust estimation of treatment effects, as in Bang and Robins (2005) and Firpo and Rothe (2016). These results from earlier work suggest that locally robust moments provide a promising approach to improving the properties of semiparametric estimators.

This paper builds on other earlier work. Locally robust moment conditions are semiparametric versions of Neyman (1959) $C(\alpha)$ test moments for parametric models, with parametric extensions to nonlikelihood settings given by Wooldridge (1991), Lee (2005), Bera et. al. (2010), and Chernozhukov, Hansen, and Spindler (2015). Hasminskii and Ibragimov (1978) suggested an estimator of a functional of a nonparametric density estimator that can be interpreted as adding the first step adjustment term. Newey (1990) derived the form of the adjustment term in some cases. Newey (1994) showed local robustness of moment functions that are derivatives of an objective function where the first step has been "concentrated out," derived the form of the adjustment term for many important cases, and showed that moment functions based on series nonparametric regression have small bias. General semiparametric model results on doubly robust estimators were given in Robins and Rotnitzky (2001).

NHR showed that adding the adjustment term gives locally robust moments for functionals of a density integral and showed the important bias and MSE advantages for locally robust estimators mentioned above. Robins et. al. (2008) showed that adding the adjustment term gives local robustness of explicit functionals of nonparametric objects, characterized some doubly robust moment conditions, and considered higher order adjustments that could further reduce bias. The form of the adjustment term for first step estimation has been derived for a variety of first step estimators by Pakes and Olley (1995), Ai and Chen (2003), Bajari, Chernozhukov, Hong, and Nekipelov (2009), Bajari, Hong, Krainer, and Nekipelov (2010), Ackerberg, Chen, and Hahn (2012), Ackerberg, Chen, Hahn, and Liao (2014), and Ichimura and Newey (2016), among others. Locally and doubly robust moments have been constructed for a variety of estimation problems by Robins, Rotnitzky, and Zhao (1994, 1995), Robins and Rotnitzky (1995), Scharfstein, Rotznitzky, and Robins (1999), Robins, Rotnitzky, and van der Laan (2000), Robins and Rotnitzky (2001), Belloni, Chernozhukov, and Wei (2013), Belloni, Chernozhukov, and Hansen (2014), Ackerberg, Chen, Hahn, and Liao (2014), Firpo and Rothe (2016), and Belloni, Chernozhukov, Fernandez-Val, and Hansen (2016).

Contributions of this paper are a general construction of locally robust estimators in a GMM setting, a general nonparametric construction of doubly robust moments, and deriving bias and other large sample properties. The special robustness properties of first step series estimators are also shown here. We use these results to obtain many new locally and doubly robust estimators, such as those where the first step allows for endogeneity or is a conditional choice probability in an economic structural model. We expect these estimators to have the advantages mentioned above, that machine learning can be used in the first step, the estimators have appealing robustness properties, smaller bias and MSE, are less sensitive to bandwidth, have closer to nominal coverage for confidence intervals, and standard errors that can be easily computed.

Section 2 describes the general construction of locally robust moment functions for semiparametric GMM. Section 3 shows how the first step adjustment term can be derived and shows the local robustness of the adjusted moments. Section 4 introduces local double robustness, shows that affine, locally robust moment functions are doubly robust, and gives new classes of doubly robust estimators. Section 5 describes how locally robust moment functions have the small bias property and a smaller remainder term. Section 6 considers first step series estimation. Section 7 characterizes locally robust moments based on conditional moment restrictions. Section 8 gives locally robust moment conditions for conditional choice probability estimation of discrete game and dynamic discrete choice models. Section 9 gives asymptotic theory based on cross fitting with easily verifiable regularity conditions for the first step, including machine learning.

2 Constructing Locally Robust Moment Functions

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

To describe the type of estimator we consider let $m(z, \beta, \gamma)$ denote an $r \times 1$ vector of functions of the data observation z, parameters of interest β, and a function γ that may be vector valued. The function γ can depend on β and z through those arguments of m. Here the function γ represents some possible first step, such as an estimator, its limit, or a true function. A GMM estimator can be based on a moment condition where β_0 is the unique parameter vector satisfying

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

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

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

where A^T denotes the transpose of a matrix A and B is the parameter space for β.

As usual a choice of $\hat{W}$ that minimizes the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$ will be a consistent estimator of the inverse of the asymptotic variance of $\sqrt{n}\hat{m}(\beta_0)$. Of course that efficient $\hat{W}$ may include adjustment terms for the first step estimator $\hat{\gamma}$. This optimal $\hat{W}$ also gives an efficient estimator in the wider sense shown in Ackerberg, Chen, Hahn, and Liao (2014). The optimal $\hat{W}$ makes $\hat{\beta}$ efficient in a semiparametric model where the only restrictions imposed are equation (2.1).

To explain and analyze local robustness we consider limits when the true distribution of a single observation z_i is F, and how those limits vary with F over a general class of distributions. This kind of analysis can be used to derive the asymptotic variance of semiparametric estimators, as in Newey (1994), and is also useful here. Let $\gamma(F)$ denote the limit of $\hat{\gamma}$ when F is the true distribution of z_i. Here $\gamma(F)$ is understood to be the limit of $\hat{\gamma}$ under general misspecification where F need not satisfy the conditions used to construct $\hat{\gamma}$. We also consider parametric models F_τ where τ denotes a vector of parameters, with F_τ equal to the true distribution F_0 at $\tau = 0$.

We will restrict each parametric model to be regular in the sense used in the semiparametric efficiency bounds literature, so that F_τ has a score $S(z)$ (derivative of the log-likelihood in many cases, e.g. see Van der Vaart, 1998, p. 362) at $\tau = 0$ and possibly other conditions are satisfied. We also require that the set of scores over all regular parametric family has mean square closure that includes all functions with mean zero and finite variance. Here we are assuming that the set of scores for regular parametric models is unrestricted, the precise meaning of the domain of $\gamma(F)$ being a general class of distributions. We define local robustness in terms of such families of regular parametric models.

DEFINITION 1: *The moment functions $m(z, \beta, \gamma)$ are locally robust if and only if for all regular parametric models*

$$\left.\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau}\right|_{\tau=0} = 0.$$

This zero pathwise derivative condition means that moment conditions are nearly zero as the first step limit $\gamma(F)$ departs from the truth γ_0 along any path $\gamma(F_\tau)$. Below we use a functional derivative condition, but for now this pathwise derivative definition is convenient. Throughout the remainder of the paper we evaluate derivatives with respect to τ at $\tau = 0$ unless otherwise specified.

In general, locally robust moment functions can be constructed by adding to moment functions the term that adjusts for (or accounts for) first step estimation. Under conditions discussed below there is a unique vector of functions $\phi(z, \beta, \gamma)$ such that $E[\phi(z_i, \beta_0, \gamma_0)] = 0$ and

$$\sqrt{n}\hat{m}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n m(z_i, \beta_0, \hat{\gamma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{m(z_i, \beta_0, \gamma_0) + \phi(z_i, \beta_0, \gamma_0)\} + o_p(1). \tag{2.2}$$

Here $\phi(z, \beta_0, \gamma_0)$ adjusts for the presence of $\hat{\gamma}$ in $m(z, \beta_0, \hat{\gamma})$. Locally robust moment functions can be constructed by adding $\phi(z, \beta, \gamma)$ to $m(z, \beta, \gamma)$ to obtain new moment functions

$$g(z, \beta, \gamma) = m(z, \beta, \gamma) + \phi(z, \beta, \gamma). \tag{2.3}$$

For $\hat{g}(\beta) = \sum_{i=1}^n g(z_i, \beta, \hat{\gamma})/n$, a locally robust semiparametric GMM estimator is obtained as

$$\hat{\beta} = \arg\min_\beta \hat{g}(\beta)'\hat{W}\hat{g}(\beta).$$

In a parametric setting it is easy to see how adding the adjustment term for first step estimation gives locally robust moment conditions. Suppose that the first step estimator $\hat{\gamma}$ is a function of a finite dimensional vector of parameters λ where there is a vector of functions $h(z, \lambda)$ satisfying $E[h(z_i, \lambda_0)] = 0$ and the first step parameter estimator $\hat{\lambda}$ satisfies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n h(z_i, \hat{\lambda}) = o_p(1). \tag{2.4}$$

For $H = \partial E[h(z_i, \lambda)]/\partial\lambda|_{\lambda=\lambda_0}$ the usual expansion gives

$$\sqrt{n}(\hat{\lambda} - \lambda_0) = -H^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(z_i, \lambda_0) + o_p(1).$$

For notational simplicity let the moment functions depend directly on λ (rather than $\gamma(\lambda)$) and so take the form $m(z, \beta, \lambda)$. Let $M_\lambda = \partial E[m(z_i, \beta_0, \lambda_0)]/\partial\lambda$. Another expansion gives

$$\begin{aligned}\sqrt{n}\hat{m}(\beta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i, \beta_0, \lambda_0) + M_\lambda\sqrt{n}(\hat{\lambda} - \lambda_0) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(z_i, \beta_0, \lambda_0) - M_\lambda H^{-1}h(z_i, \lambda_0)\} + o_p(1).\end{aligned}$$

Here we see that the adjustment term is

$$\phi(z, \beta, \lambda) = -M_\lambda H^{-1}h(z, \lambda). \tag{2.5}$$

We can add this term to the original moment functions to produce new moment functions of the form

$$g(z, \beta, \lambda) = m(z, \beta, \lambda) + \phi(z, \beta, \lambda) = m(z, \beta, \lambda) - M_\lambda H^{-1}h(z, \lambda).$$

Local robustness of these moment functions follows by the chain rule and

$$\left.\frac{\partial E[g(z_i, \beta_0, \lambda)]}{\partial\lambda}\right|_{\lambda=\lambda_0} = \left.\frac{\partial E[m(z_i, \beta_0, \lambda) - M_\lambda H^{-1}h(z_i, \lambda)]}{\partial\lambda}\right|_{\lambda=\lambda_0} = M_\lambda - M_\lambda H^{-1}H = 0.$$

Neyman (1959) used scores and the information matrix to form such $g(z, \beta, \lambda)$ in a parametric likelihood setting, where $g(z, \beta_0, \lambda_0)$ has an orthogonal projection interpretation. There the purpose was to construct tests, based on $g(z, \beta, \lambda)$, where estimation of the nuisance parameters λ did not affect the distribution of the tests. The form here was given in Wooldridge (1991) for nonlinear least squares and Lee (2005), Bera et. al (2010), and Chernozhukov et. al. (2015) for GMM. What appears to be new here is construction of locally robust moment functions by adding the adjustment term to original moment functions.

In general the adjustment term $\phi(z, \beta, \gamma)$ may depend on unknown components that are not present in the original moment functions $m(z, \beta, \gamma)$. In the above parametric example the matrix $M_\lambda H^{-1}$ is unknown so its elements should really be included in γ, along with λ. If we do this local robustness will continue to hold with these additional components of γ because $E[h(z_i, \lambda_0)] = 0$. For notational simplicity we will take γ to be the first step for the locally robust moment functions $g(z, \beta, \gamma) = m(z, \beta, \gamma) + \phi(z, \beta, \gamma)$, with the understanding $\phi(z, \beta, \gamma)$ will generally depend on first step functions that are not included in $m(z, \beta, \gamma)$.

In general semiparametric settings the form of the adjustment term $\phi(z, \beta, \gamma)$ and local robustness of $g(z, \beta, \gamma)$ can be explained in terms of influence functions. We will do so in the

next Section. In many interesting cases the form of the adjustment term $\phi(z,\beta,\gamma)$ is already known, allowing construction of locally robust estimators. We conclude this section with an important class of examples.

The class of examples we consider is one where the first step $\hat{\gamma}_1$ is based on a conditional moment restriction $E[\rho(z_i,\gamma_{10})|x_i]=0$ for a residual $\rho(z,\gamma_1)$ and instrumental variables x. The conditional mean or median of y_i given x_i are included as special cases where $\rho(z,\gamma_1)=y-\gamma_1(x)$ and $\rho(z,\gamma_1)=2\cdot 1(y<\gamma_1(x))-1$ respectively, as are versions that allow for endogeneity where γ_1 depends on variables other than x. We take $\hat{\gamma}_1$ to have the same limit as the nonparametric two-stage least squares (NP2SLS) estimator of Newey and Powell (1989, 2003) and Newey (1991). Thus, $\hat{\gamma}_1$ has limit $\gamma_1(F)$ satisfying

$$\gamma_1(F)=\arg\min_{\gamma_1\in\Gamma} E_F[\{E_F[\rho(z_i,\gamma_1)|x_i]\}^2],$$

and E_F denotes the expectation under the distribution F. Suppose that there is $\gamma_{20}(x)$ in the mean square closure of the set of derivatives $\partial E[\rho(z_i,\gamma_1(F_\tau))|x_i]/\partial\tau$ as F_τ varies over regular parametric models such that

$$\frac{\partial E[m(z_i,\beta_0,\gamma_1(F_\tau))]}{\partial\tau}=-E[\gamma_{20}(x_i)\frac{\partial E[\rho(z_i,\gamma_1(F_\tau))|x_i]}{\partial\tau}]. \tag{2.6}$$

Then from Ichimura and Newey (2016) the adjustment term is

$$\phi(z,\beta,\gamma)=\gamma_2(x,\beta)\rho(z,\gamma_1). \tag{2.7}$$

A function $\gamma_{20}(x)$ satisfying equation (2.6) exists when the set of derivatives $\partial E[\rho(z_i,\gamma_1(F_\tau))|x_i]/\partial\tau$ is linear as F_τ varies over parametric models, $\partial E[m(z_i,\beta_0,\gamma_1(F_\tau))]/\partial\tau$ is a linear functional of $\partial E[\rho(z_i,\gamma_1(F_\tau))|x_i]/\partial\tau$, and that functional is continuous in mean square. Existence of $\gamma_{20}(x)$ then follows from the Riesz representation theorem. Special cases of this characterization of $\gamma_{20}(x)$ are in Newey (1994), Ai and Chen (2007), and Ackerberg, Chen, Hahn, and Liao (2014). When $\partial E[m(z_i,\beta_0,\gamma_1(F_\tau))]/\partial\tau$ is not a mean square continuous functional of $\partial E[\rho(z_i,\gamma_1(F_\tau))|x_i]/\partial\tau$ then first step estimation should make the moments converge slower than $1/\sqrt{n}$, as shown by Newey and McFadden (1994) and Severini and Tripathi (2012) for special cases. The adjustment term given here includes Santos (2011) as a special case with $m(z,\beta,\gamma_1)=\int v(x)\gamma_1(x)dx-\beta$, though Santos (2011) is more general in allowing for nonidentification of γ_{10}.

There are a variety of ways to construct an estimator $\phi(z_i,\beta,\hat{\gamma})$ of the adjustment term to be used in forming locally robust moment functions, see NHR and Ichimura and Newey (2016). A relatively simple and general one when the first step is a series or sieve estimator is to treat the first step as if it were parametric and use the parametric formula in equation (2.5). This approach to estimating the adjustment term is known to be asymptotically valid in a variety

of settings, see Newey (1994), Ackerberg, Chen, and Hahn (2012), and Ichimura and Newey (2016). For completeness we give a brief description here.

We parameterize an approximation to γ_1 as $\gamma_1 = \gamma(\lambda)$ where λ is a finite dimensional vector of parameters as before. Let $m(z_i, \beta, \lambda) = m(z_i, \beta, \gamma_1(\lambda))$ and $\hat{\lambda}_i$ denote an estimator of λ_0 solving $E[h(z_i, \lambda_0)] = 0$, that is allowed to depend on observation i. Being a series or sieve estimator the dimension of λ and hence of $h(z, \lambda)$ will increase with sample size. Also, let $\mathcal{I}(i)$ be a set of observation indices that can also depend on i. An estimator of the adjustment term is give by

$$\phi(z_i, \beta, \hat{\gamma}) = -\hat{M}_{i\lambda}(\beta)\hat{H}_i^{-1}h(z_i, \hat{\lambda}_i), \hat{M}_{i\lambda}(\beta) = \sum_{j \in \mathcal{I}(i)} \frac{\partial m(z_j, \beta, \hat{\lambda}_i)}{\partial \lambda}, \hat{H}_i = \sum_{j \in \mathcal{I}(i)} \frac{\partial h(z_j, \hat{\lambda}_i)}{\partial \lambda}.$$

This estimator allows for cross fitting where $\hat{\lambda}_i$, $\hat{M}_{i\lambda}$, and $\hat{H}_i$ depend only on observations other than the i^{th}. cross fitting is known to improve performance in some settings, such as in "leave one out" kernel estimators of averages, e.g. see NHR. This adjustment term will lead to locally robust moments in a variety of settings, as further discussed below.

3 Influence Functions, Adjustment Terms, and Local Robustness

Influence function calculations can be used to derive the form of the adjustment term $\phi(z, \beta, \gamma)$ and show local robustness of the adjusted moment functions $g(z, \beta, \gamma) = m(z, \beta, \gamma) + \phi(z, \beta, \gamma)$. To explain influence functions note that many estimators are asymptotically equivalent to a sample average. The object being averaged is the unique influence function. For example, in equation (2.2) we are assuming that the influence function of $\hat{m}(\beta_0)$ is $g(z, \beta_0, \gamma_0) = m(z, \beta_0, \gamma_0) + \phi(z, \beta_0, \gamma_0)$. This terminology is widely used in the semiparametric estimation literature.

In general an estimator $\hat{\mu}$ of a true value μ_0 and its influence function $\psi(z)$ satisfy

$$\sqrt{n}(\hat{\mu} - \mu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + o_p(1), E[\psi(z_i)] = 0, E[\psi(z_i)\psi(z_i)'] \text{ exists.}$$

The function $\psi(z)$ can be characterized in terms of the functional $\mu(F)$ that is the limit of $\hat{\mu}$ under general misspecification where F need not satisfy the conditions used to construct $\hat{\mu}$. As before, we allow F to vary over a family of regular parametric models where the set of scores for the family has mean square closure that includes all mean zero functions with finite variance. As shown by Newey (1994) the influence function $\psi(z)$ is then the unique solution to a derivative equation of Van der Vaart (1991),

$$\frac{\partial \mu(F_\tau)}{\partial \tau} = E[\psi(z_i)S(z_i)], E[\psi(z_i)] = 0, \tag{3.1}$$

as F_τ (and hence $S(z)$) varies over the general family of regularly parametric models. Ichimura and Newey (2016) also showed that when $\psi(z)$ has certain continuity properties it can be computed as

$$\psi(z) = \lim_{h \longrightarrow 0} \frac{\partial \mu(F_\tau^h)}{\partial \tau}, F_\tau^h = (1-\tau)F_0 + \tau G_z^h, \tag{3.2}$$

where G_z^h is constructed so that F_τ^h is in the domain of $\mu(F)$ and G_z^h approaches the point mass at z as $h \longrightarrow 0$.

These results can be used to derive the adjustment term $\phi(z, \beta, \gamma)$ and to explain local robustness. Let $\gamma(F)$ denote the limit of the first step estimator $\hat{\gamma}$ under general misspecification when a single observation has CDF F, as discussed above. From Newey (1994, pp. 1356-1357) we know that the adjustment term $\phi(z, \beta_0, \gamma_0)$ is the influence function of $\mu(F) = E[m(z_i, \beta_0, \gamma(F))]$ where $E[\cdot]$ denotes the expectation at the truth. Thus $\phi(z, \beta, \gamma)$ can be calculated as in equation (3.2) for that $\mu(F)$. Also $\phi(z, \beta, \gamma)$ satisfies equation (3.1) for $\mu(F) = E[m(z_i, \beta_0, \gamma(F))]$, i.e. for the score $S(z)$ at $\tau = 0$ for any regular parametric model $\{F_\tau\}$,

$$\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau} = E[\phi(z_i, \beta_0, \gamma_0)S(z_i)], E[\phi(z_i, \beta_0, \gamma_0)] = 0. \tag{3.3}$$

Also, $\phi(z, \beta_0, \gamma_0)$ can be computed as

$$\phi(z, \beta_0, \gamma_0) = \lim_{h \longrightarrow 0} \frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau^h))]}{\partial \tau}, F_\tau^h = (1-\tau)F_0 + \tau G_z^h,$$

The characterization of $\phi(z, \beta_0, \gamma_0)$ in equation (3.3) can be used to specify another local robustness property that is equivalent to Definition 1. We have defined local robustness as the derivative on the left of the first equality in equation (3.3) being zero for all regular parametric models. If that derivative is zero for all parametric models then $\phi(z, \beta_0, \gamma_0) = 0$ is the unique solution to this equation, by the set of scores being mean square dense in the set of mean zero random variables with finite variance. Also, if $\phi(z, \beta_0, \gamma_0) = 0$ then the derivative on the left is always zero. Therefore we have

PROPOSITION 1: $\phi(z, \beta_0, \gamma_0) = 0$ *if and only if* $m(z, \beta, \gamma)$ *is locally robust.*

Note that $\phi(z, \beta, \gamma)$ is the term in the influence function of $\hat{m}(\beta_0)$ that accounts for the first step estimator $\hat{\gamma}$. Thus Proposition 1 gives an alternative characterization of local robustness, that first step estimation does not affect the influence function of $\hat{m}(\beta_0)$. This result is a semiparametric version of Theorem 6.2 of Newey and McFadden (1994). It also formalizes the discussion in Newey (1994, pp. 1356-1357).

Local robustness of the adjusted moment function $g(z, \beta, \gamma) = m(z, \beta, \gamma) + \phi(z, \beta, \gamma)$ follows from Proposition 1 and $\phi(z, \beta, \gamma)$ being a nonparametric influence function. Because $\phi(z, \beta, \gamma)$ is an influence function it has mean zero at all true distributions, i.e. $\mu(F) \stackrel{def}{=}$

$\int \phi(z,\beta_0,\gamma(F))F(dz) \equiv 0$ identically in F. Consequently the derivative in equation (3.1) is zero, so that (like Proposition 1) the influence function of $\mu(F)$ is zero. Consequently, under appropriate regularity conditions $\bar{\phi} = \sum_{i=1}^n \phi(z_i,\beta_0,\hat{\gamma})/n$ has a zero influence function and so $\sqrt{n}\bar{\phi} = o_p(1)$. It then follows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n g(z_i,\beta_0,\hat{\gamma}) = \sqrt{n}\hat{m}(\beta_0) + \sqrt{n}\bar{\phi} = \sqrt{n}\hat{m}(\beta_0) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(z_i,\beta_0,\gamma_0) + o_p(1), \quad (3.4)$$

where the last equality follows by equation (2.2). Here we see that the adjustment term is zero for the moment functions $g(z,\beta,\gamma)$. From Proposition 1 with $g(z,\beta,\gamma)$ replacing $m(z,\beta,\gamma)$ it then follows that $g(z,\beta,\gamma)$ is locally robust.

PROPOSITION 2: *For the influence function $\phi(z,\beta_0,\gamma_0)$ of $\mu(F) = E[m(z_i,\beta_0,\gamma(F))]$ the adjusted moment function $g(z,\beta,\gamma) = m(z,\beta,\gamma) + \phi(z,\beta,\gamma)$ is locally robust.*

Local robustness of $g(z,\beta,\gamma)$ also follows directly from the identity $\int \phi(z,\beta_0,\gamma(F))F(dz) \equiv 0$ as discussed in the Appendix. Also, the adjusted moments $\hat{g}(\beta_0)$ have the same asymptotic variance as the original moments, as in the second equality of equation (3.4). That is, adding $\phi(z,\beta,\gamma)$ to $m(z,\beta,\gamma)$ does not affect the asymptotic variance. Thus the asymptotic benefits of the locally robust moments are in their higher order properties. Other modifications of the moments may also improve higher-order properties of estimators, such as the cross fitting described above (like "leave on out" in NHR) and the higher order bias corrections in Robins et. al. (2008) and Cattaneo and Jansson (2014).

4 Local and Double Robustness

The zero derivative condition in Definition 1 is an appealing robustness property in and of itself. Mathematically a zero derivative is equivalent to the moments remaining closer to zero than τ as τ varies away from zero. This property can be interpreted as local robustness of the moments to the value of γ being plugged in, with the moments remaining close to zero as γ varies away from its true value. Because it is difficult to get nonparametric functions exactly right, especially in high dimensional settings, this property is an appealing one.

Such robustness considerations, well explained in Robins and Rotnitzky (2001), have motivated the development of doubly robust estimators. For our purposes doubly robust moments have expectation zero if just one first stage component is incorrect. When there are only two first stage components this means that the moment conditions hold when only one of the first stage components is correct. Doubly robust moment conditions allow two chances for the moment conditions to hold.

It turns out that locally robust moment functions are automatically doubly robust in a local sense that the derivative with respect to each individual, distinct first stage component is zero. In that way the moment conditions nearly hold as each distinct component varies in a neighborhood of the truth. Furthermore, when locally robust moment functions are affine functions of a distinct first step component they are automatically globally robust in that component. Thus, locally robust moment functions that are affine in each distinct first step are doubly robust.

These observations suggest a way to construct doubly robust moment functions. Starting with any two step semiparametric moment function we can add the adjustment term to get a locally robust moment function. When we can choose a first step of that moment function so that it enters in an affine way the new moment function will be doubly robust in that component.

To give these results we need to define distinct components of γ. A distinct component is one where there are parametric models F_τ with that component varying in an unrestricted way but the other components of γ not varying. For a precise definition we will focus on the first component γ_1 of $\gamma = (\gamma_1, ..., \gamma_J)$.

DEFINITION 2: *A component γ_1 of γ is distinct if and only if there is F_τ such that*

$$\gamma(F_\tau) = (\gamma_1(F_\tau), \gamma_{20}, ..., \gamma_{J0}),$$

and $\gamma_1(F_\tau)$ is unrestricted as F_τ varies across parametric models.

An example is the moment function $g(z, \beta, \gamma_1, \gamma_2) = m(z, \beta, \gamma_1) + \gamma_2(x, \beta)\rho(z, \gamma_1)$, where $E[\rho(z_i, \gamma_{10})|x_i] = 0$. In that example the two components γ_1 and γ_2 are often distinct because γ_1 depends only on the conditional distribution of z_i given x_i and $\gamma_{20}(x, \beta)$ depends on the marginal distribution of x_i in an unrestricted way.

Local robustness means that the derivative must be zero for any model, so in particular it must be zero for any model where only the distinct component is varying. Thus we have

PROPOSITION 3: *If γ_1 is distinct then for $g(z, \beta, \gamma) = m(z, \beta, \gamma) + \phi(z, \beta, \gamma)$ and regular parametric models F_τ as in Definition 2,*

$$\frac{\partial E[g(z_i, \beta_0, \gamma_1(F_\tau), \gamma_{20}, ..., \gamma_{J0})]}{\partial \tau} = 0.$$

This result is an application of the simple fact that when a multivariable derivative is zero the partial derivative must be zero when the variables are allowed to vary in an unrestricted way. Although this fact is simple, it is helpful in understanding when local robustness holds for individual components. This means that locally robust moment functions automatically have a local double robustness property, that the expectation of the moment function remains nearly zero as each distinct first stage component varies away from the truth. For example, for

a first step conditional moment restriction where $g(z,\beta,\gamma) = m(z,\beta,\gamma_1) + \gamma_2(x,\beta)\rho(z,\gamma_1)$, the conclusion of Proposition 3 is

$$\frac{\partial E[m(z_i,\beta_0,\gamma_1(F_\tau)) + \gamma_{20}(x_i)\rho(z_i,\gamma_1(F_\tau))]}{\partial \tau} = 0.$$

In fact, this result is implied by equation (2.6), so by construction $g(z,\beta,\gamma)$ is already locally robust in γ_1 alone. Local robustness in γ_2 follows by the conditional moment restriction $E[\rho(z_i,\gamma_{10})|x_i] = 0$.

Moments that are locally robust in a distinct component γ_1 will be globally robust in γ_1 if γ_1 enters the moments in an affine way, meaning that for any γ_1 and $\gamma = (\gamma_1,\gamma_{20},...,\gamma_{J0})'$ and any z,

$$g(z,\beta,(1-\tau)\gamma_0 + \tau\gamma) = (1-\tau)\cdot g(z,\beta,\gamma_0) + \tau \cdot g(z,\beta,\gamma). \tag{4.1}$$

Global robustness holds because an affine function with zero derivative is constant. For simplicity we state a result when F_τ can be chosen so that $\gamma(F_\tau) = (1-\tau)\gamma_0 + \tau\gamma$ though it will hold more generally. Note that here

$$E[g(z_i,\beta_0,\gamma(F_\tau))] = (1-\tau)E[g(z_i,\beta_0,\gamma_0)] + \tau \cdot E[g(z_i,\beta_0,\gamma)] = \tau \cdot E[g(z_i,\beta_0,\gamma)].$$

Here the derivative of the moment condition with respect to τ is just $E[g(z_i,\beta_0,\gamma)]$ so Proposition 3 gives the following result:

PROPOSITION 4: *If equation (4.1) is satisfied and there is F_τ with $\gamma(F_\tau) = ((1-\tau)\gamma_{10} + \tau\gamma_1,\gamma_{20},...,\gamma_{J0})'$ then $E[g(z_i,\beta_0,\gamma_1,\gamma_{20},...,\gamma_{J0})] = 0$.*

Thus we see that locally robust moment functions that are affine in a distinct first step component are globally robust in that component. This result includes many existing examples of doubly robust moment functions and can be used to construct new ones.

A general class of doubly robust moment functions that appears to be new and includes many new and previous examples has first step satisfying a conditional moment restriction $E[\rho(z_i,\gamma_{10})|x_i] = 0$ where $\rho(z,\gamma_1)$ and $m(z,\beta_0,\gamma_1)$ are affine in γ_1. Suppose that $E[m(z_i,\beta_0,\gamma_1)]$ is a mean-square continuous linear functional of $E[\rho(z_i,\gamma_1)|x_i]$ for γ_1 in a linear set Γ. Then by the Riesz representation theorem there is $\gamma^*(x)$ in the mean square closure Π of the image of $E[\rho(z_i,\gamma_1)|x_i]$ such that

$$E[m(z_i,\beta_0,\gamma_1)] = -E[\gamma^*(x_i)E[\rho(z_i,\gamma_1)|x_i]] = -E[\gamma^*(x_i)\rho(z_i,\gamma_1)], \gamma_1 \in \Gamma. \tag{4.2}$$

Let $\gamma_{20}(x)$ be any function such that $\gamma_{20}(x_i) - \gamma^*(x_i)$ is orthogonal to Π and $g(z,\beta,\gamma) = m(z,\beta,\gamma_1) + \gamma_2(x,\beta)\rho(z,\gamma_1)$. Then $E[g(z_i,\beta_0,\gamma_1,\gamma_{20})] = 0$ by the previous equation. It also follows that $E[g(z_i,\beta_0,\gamma_{10},\gamma_2)] = 0$ by $E[\rho(z_i,\gamma_{10})|x_i] = 0$. Therefore $g(z,\beta,\gamma_1,\gamma_2)$ is doubly robust, showing the following result:

PROPOSITION 5: *If* $m(z_i, \beta_0, \gamma_1)$ *and* $\rho(z_i, \gamma_1)$ *are affine in* $\gamma_1 \in \Gamma$ *with* Γ *linear and* $E[m(z_i, \beta_0, \gamma_1)]$ *is a linear, mean square continuous functional of* $E[\rho(z_i, \gamma_1)|x_i]$ *then there is* $\gamma_{20}(x)$ *such that* $g(z, \beta, \gamma_1, \gamma_2) = m(z, \beta, \gamma_1) + \gamma_2(x, \beta)\rho(z, \gamma_1)$ *is doubly robust.*

Section 3 of Robins et. al. (2008) gives necessary and sufficient conditions for a moment function to be doubly robust when γ_1 and γ_2 enter the moment functions as functions evaluated at observed x. Proposition 5 is complementary to that work in deriving the form of doubly robust moment functions when the first step satisfies a conditional moment restriction and $m(z, \beta, \gamma_1)$ can depend on the entire function γ_1.

It is interesting to note that γ_{20} such that $E[g(z, \beta_0, \gamma_1, \gamma_{20})] = 0$ for all $\gamma_1 \in \Gamma$ is not unique when Π does not include all functions of x, the overidentified case of Chen and Santos (2015). This nonuniqueness can occur when there are multiple ways to estimate the first step γ_{10} using the conditional moment restrictions $E[\rho(z_i, \gamma_{10})|x_i] = 0$. As discussed in Ichimura and Newey (2016), the different $\gamma_{20}(x_i)$ correspond to different first step estimators, with $\gamma_{20}(x_i) = \gamma^*(x_i)$ corresponding to the NP2SLS estimator.

An important case is a linear conditional moment restrictions setup up like Newey and Powell (1989, 2003) and Newey (1991) where

$$\rho(z, \gamma_1) = y - \gamma_1(w), E[y_i - \gamma_{10}(w_i)|x_i] = E[\rho(z_i, \gamma_{10})|x_i] = 0. \tag{4.3}$$

Consider a moment function equal to $m(z, \beta, \gamma_1) = v(w)\gamma_1(w) - \beta$ for some known function $v(w)$, where the parameter of interest is $\beta_0 = E[v(w_i)\gamma_{10}(w_i)]$. If there is $\bar{\gamma}(x)$ such that $v(w_i) = E[\bar{\gamma}(x_i)|w_i]$ then we have

$$\begin{aligned} E[m(z_i, \beta_0, \gamma_1)] &= E[v(w_i)\{\gamma_1(w_i) - \gamma_{10}(w_i)\}] = E[E[\bar{\gamma}(x_i)|w_i]\{\gamma_1(w_i) - \gamma_{10}(w_i)\}] \\ &= E[\bar{\gamma}(x_i)\{\gamma_1(w_i) - \gamma_{10}(w_i)\}] = -E[\bar{\gamma}(x_i)\rho(z_i, \gamma_1)]. \end{aligned}$$

It follows that $g(z, \beta, \gamma) = m(z, \beta, \gamma_1) + \gamma_2(x)\rho(z, \gamma_1)$ is doubly robust for $\gamma_{20}(x) = \bar{\gamma}(x_i)$. Interestingly, the existence of $\bar{\gamma}$ with $v(w_i) = E[\bar{\gamma}(x_i)|w_i]$ is a necessary condition for root-n consistent estimability of β_0 as in Severini and Tripathi's (2012, Lemma 4.1). We see here that a doubly robust moment condition can always be constructed when this necessary condition is satisfied. Also, similarly to the above, the $\gamma_{20}(x)$ may not be unique.

COROLLARY 6: *If* $m(z, \beta, \gamma_1) = v(w)\gamma_1(w) - \beta$, *equation (4.3) is satisfied, and there is* $\bar{\gamma}(x)$ *such that* $v(w) = E[\bar{\gamma}(x)|w]$ *then* $g(z, \beta, \gamma_1, \gamma_2) = v(w)\gamma_1(w) - \beta + \gamma_2(x)[y - \gamma_1(w)]$ *is doubly robust for* $\gamma_{20}(x) - \bar{\gamma}(x)$ *orthogonal to* Π.

A new example of a doubly robust moment condition corresponds to the weighted average derivative of $\gamma_{10}(w)$ of Ai and Chen (2007). Here $m(z, \beta, \gamma_1) = \bar{v}(w)\partial\gamma_1(w)/\partial w - \beta$ for some

function $\bar{v}(w)$. Let $f_0(w)$ be the pdf of w_i. Assuming that $\bar{v}(w)\gamma_1(w)f_0(w)$ is zero on the boundary of the support of w_i, integration by parts gives

$$E[m(z_i,\beta_0,\gamma_1)] = E[v(w_i)\{\gamma_1(w_i) - \gamma_{10}(w_i)\}], v(w) = f_0(w)^{-1}\partial[\bar{v}(w)f_0(w)]/\partial w.$$

Assume that there exists $\bar{\gamma}(x)$ such that $v(w_i) = E[\bar{\gamma}(x_i)|w_i]$. Then as in Proposition 5 a doubly robust moment function is

$$g(z,\beta,\gamma) = \bar{v}(w)\frac{\partial\gamma_1(w)}{\partial w} - \beta + \gamma_2(x)[y - \gamma_1(w)].$$

A special case of this example is the doubly robust moment condition for the weighted average derivative in the exogenous case where $w_i = x_i$ given in Firpo and Rothe (2016).

Doubly robust moment conditions can be used to identify parameters of interest. In general, if $g(z,\beta,\gamma_1,\gamma_2)$ is doubly robust and γ_{20} is identified then β_0 may be identified from

$$E[g(z_i,\beta_0,\bar{\gamma}_1,\gamma_{20})] = 0,$$

for any fixed $\bar{\gamma}_1$ when the solution β_0 to this equation is unique.

PROPOSITION 7: *If $g(z,\beta,\gamma_1,\gamma_2)$ is doubly robust, γ_{20} is identified, and for some $\bar{\gamma}_1$ the equation $E[g(z_i,\beta,\bar{\gamma}_1,\gamma_{20})] = 0$ has a unique solution then β_0 is identified as that solution.*

Applying this result to the NPIV setting gives an explicit formula for certain functionals of $\gamma_{10}(w)$ without requiring that the completeness identification condition of Newey and Powell (2003) be satisfied, similarly to Santos (2011). Suppose that $v(w)$ is identified, e.g. as for the weighted average derivative. Since both w and x are observed it follows that a solution $\gamma_{20}(x)$ to $v(w) = E[\gamma_{20}(x)|w]$ will be identified if such a solution exists. Plugging in $\bar{\gamma}_1 = 0$ in the equation $E[g(z_i,\beta_0,\bar{\gamma}_1,\gamma_{20})] = 0$ gives

COROLLARY 8: *If $v(w_i)$ is identified and there exists $\gamma_{20}(x_i)$ such that $v(w_i) = E[\gamma_{20}(x_i)|w_i]$ then $\beta_0 = E[v(w_i)\gamma_{10}(w_i)]$ is identified as $\beta_0 = E[\gamma_{20}(x_i)y_i]$.*

Note that this result holds without the completeness condition. Identification of $\beta_0 = E[v(w_i)\gamma_{10}(w_i)]$ for known $v(w_i)$ with $v(w_i) = E[\gamma_{20}(x_i)|w_i]$ follows from Severini and Tripathi (2006). Santos (2011) gives a related formula for a parameter $\beta_0 = \int \tilde{v}(w)\gamma_{20}(w)dw$. The formula here differs from Santos (2011) in being an expectation rather than a Lebesgue integral. Santos (2011) constructed an estimator. That is beyond the scope of this paper.

Another new example of a doubly robust estimator is a weighted average over income values of an average (across heterogenous individuals) of exact consumer surplus bounds, as in Hausman and Newey (2016). Here y is quantity consumed, $w = x = (x_1,x_2)'$, x_1 is price, x_2 is income, $\gamma_{10}(x_i) = E[y_i|x_i]$, price is changing between $\breve{x}_1$ and $\bar{x}_1$, and B is a bound on the income

effect. Let $v_2(x_2)$ be some weight function and $v_1(x_1) = 1(\breve{x}_1 \leq x_1 \leq \bar{x}_1)e^{-B(x_1-\breve{x}_1)}$. For the moment function $m(z,\beta,\gamma_1) = v_2(x_2)\int v_1(u)\gamma_1(u,x_2)du - \beta$ the true parameter β_0 is a bound on the average of equivalent variation over unobserved individual heterogeneity and income. Let $f_{10}(x_1|x_2)$ denote the conditional pdf of x_1 given x_2. Note that

$$\begin{aligned} E[m(z_i,\beta_0,\gamma_1)] &= E[v_2(x_{2i})\int v_1(u)[\gamma_1(u,x_{2i}) - \gamma_{10}(u,x_{2i})]du] \\ &= E[f_{10}(x_{1i}|x_{2i})^{-1}v_1(x_{1i})v_2(x_{2i})\{\gamma_1(x_i) - \gamma_{10}(x_i)\}] \\ &= -E[\gamma_{20}(x_i)E[y_i - \gamma_1(x_i)|x_i]], \gamma_{20}(x) = f_{10}(x_1|x_2)^{-1}v_1(x_1)v_2(x_2). \end{aligned}$$

Then it follows by Proposition 5 that a doubly robust moment function is

$$g(z,\beta,\gamma) = v_2(x_2)\int v_1(u)\gamma_1(u,x_2)du - \beta + \gamma_2(x)[y - \gamma_1(x)].$$

When the moment conditions are formulated so that they are affine in the first step Proposition 4 applies to many previously developed doubly robust moment conditions. Data missing at random is a leading example. Let β_0 be the mean of a variable of interest w where w is not always observed, $y \in \{0,1\}$ denote an indicator for w being observed, and x a vector of covariates. Assume w is mean independent of y conditional on covariates x. We consider estimating β_0 using the propensity score $P_0(x_i) = \Pr(y_i = 1|x_i)$. We specify an affine conditional moment restriction by letting $\gamma_1(x) = 1/P(x)$ and $\rho(z,\gamma_1) = \gamma_1(x)y - 1$. We have $\beta_0 = E[\gamma_{10}(x_i)y_iw_i]$, as is well known. An affine moment function is then $m(z,\beta,\gamma_1) = \gamma_1(x)yw - \beta$. Note that

$$\begin{aligned} E[m(z_i,\beta_0,\gamma_1)] &= E[E[y_iw_i|x_i]\{\gamma_1(x_i) - \gamma_{10}(x_i)\}] = -E[\gamma_{20}(x_i)\rho(z_i,\gamma_1)], \\ \gamma_{20}(x_i) &= -\gamma_{10}(x_i)E[y_iw_i|x_i]. \end{aligned}$$

Then Proposition 5 implies that a doubly robust moment function is given by

$$g(z,\beta,\gamma) = \gamma_1(x)yw - \beta - \gamma_2(x)[\gamma_1(x)y - 1].$$

This is the well known doubly robust moment function of Robins, Rotnitzky, and Zhao (1994).

This example illustrates how applying Propositions 4 and 5 require specifying the first step so that the moment functions are affine. These moment conditions were originally shown to be doubly robust when the first step is taken to be the propensity score $P(x)$. Propositions 4 and 5 only apply when the first step is taken to be $1/P(x)$. More generally we expect that particular formulations of the first step may be needed to make the moment functions affine in the first step and so use Propositions 4 and 5 to derive doubly robust moment functions.

Another general class of doubly robust moment functions depend on the pdf γ_1 of a subset of variables x_i and are affine in γ_1. An important example of such a moment function is the average where $\beta_0 = \int f_0(x)^2dx$ and $m(z,\beta,\gamma_1) = \gamma_1(x) - \beta$. Another is the density

weighted average derivative (WAD) of Powell, Stock, and Stoker (1989) where $m(z,\beta,\gamma_1) = -2y \cdot \partial\gamma_1(x)/\partial x - \beta$. Assume that $E[m(z_i,\beta_0,\gamma_1)]$ is a function of $\gamma_1 - \gamma_{10}$ that is continuous in the norm $\left[\int [\gamma_1(u) - \gamma_{10}(u)]^2 du\right]^{1/2}$. Then by the Riesz representation theorem there is $\gamma_{20}(x)$ with

$$E[m(z_i,\beta_0,\gamma_1)] = \int \gamma_{20}(u)[\gamma_1(u) - \gamma_{10}(u)]du. \tag{4.4}$$

The adjustment term for $m(z,\beta,\gamma)$, as in Proposition 3 of Newey (1994), is $\phi(z,\beta,\gamma) = \gamma_2(x) - \int \gamma_2(u)\gamma_1(u)du$. The corresponding locally robust moment function is

$$g(z,\beta,\gamma_1,\gamma_2) = m(z,\beta,\gamma_1) + \gamma_2(x) - \int \gamma_2(u)\gamma_1(u)du. \tag{4.5}$$

This function is affine in γ_1 and γ_2 separately so when they are distinct Proposition 4 implies double robustness. Double robustness also follows directly from

$$\begin{aligned} E[g(z_i,\beta_0,\gamma)] &= \int \gamma_{20}(u)[\gamma_1(u) - \gamma_{10}(u)]du + \int \gamma_2(u)\gamma_{10}(u)du - \int \gamma_2(u)\gamma_1(u)du \\ &= -\int [\gamma_2(u) - \gamma_{20}(u)][\gamma_1(u) - \gamma_{10}(u)]du. \end{aligned}$$

Thus we have the following result:

PROPOSITION 9: *If $m(z_i,\beta,\gamma_1)$ is affine in γ_1 and $E[m(z_i,\beta_0,\gamma_1)]$ is a linear function of $\gamma_1 - \gamma_{10}$ that is continuous in the norm $\left[\int \{\gamma_1(x) - \gamma_{10}(x)\}^2 dx\right]^{1/2}$, then for $\gamma_{20}(x)$ from equation (4.4), $g(z,\beta,\gamma) = m(z,\beta,\gamma_1) + \gamma_2(x) - \int \gamma_2(u)\gamma_1(u)du$ is doubly robust.*

We can use this result to derive doubly robust moment functions for the WAD. Let $\delta(x_i) = E[y_i|x_i]\gamma_{10}(x_i)$. Assuming that $\delta(u)\gamma_1(u)$ is zero on the boundary, integration by parts gives

$$E[m(z_i,\beta_0,\gamma_1)] = -2E[y_i\partial\gamma_1(x_i)/\partial x] - \beta_0 = 2\int [\partial\delta(u)/\partial u]\{\gamma_1(u) - \gamma_{10}(u)\}du,$$

so that $\gamma_{20}(x) = 2\partial\delta(x)/\partial x$. A doubly robust moment condition is then

$$g(z,\beta,\gamma) = -2y\frac{\partial\gamma_1(x)}{\partial x} - \beta + 2\frac{\partial\delta(x)}{\partial x} - \int 2\frac{\partial\delta(u)}{\partial u}\gamma_1(u)du.$$

The double robustness of this moment condition appears to be a new result. As shown in Newey, Hsieh, and Robins (1998), a "delete-one" symmetric kernel estimator based on this moment function gives the twicing kernel estimator of NHR. Consequently the MSE comparisons of NHR for twicing kernel estimators with the original kernel estimator correspond to comparison of a doubly (and locally) robust estimator with one based on unadjusted moment conditions, as discussed in the introduction.

It is interesting to note that Proposition 9 does not require that γ_1 and γ_2 are distinct first step components. For the average density $\gamma_1(x)$ and $\gamma_2(x)$ both represent the marginal density

of x and so are not distinct. Nevertheless the moment function $g(z, \beta_0, \gamma) = \gamma_1(x) - \beta_0 + \gamma_2(x) - \int \gamma_1(u)\gamma_2(u)du$ is doubly robust, having zero expectation if either γ_1 or γ_2 is correct. This example shows a moment function may be doubly robust even though γ_1 and γ_2 are not distinct. Thus, there are doubly robust moment functions that cannot be constructed using Proposition 4.

All of the results of this Section continue to hold with cross fitting. That is true because the results of this Section concern the moment and their expectation at various values of the first step, and not the particular way in which the first step is formed.

5 Small Bias of Locally Robust Moment Conditions

Adding the adjustment term improves the higher order properties of the estimated moments though it does not change their asymptotic variance. An advantage of locally robust moment functions is that the effect of first step smoothing bias is relatively small. To describe this advantage it is helpful to modify the definition of local robustness. In doing so we allow F to represent a more general object, an unsigned measure (charge). Let $\|\cdot\|$ denote a seminorm on F (a seminorm has all the properties of a norm but may be zero when F is not zero). Also, let $\mathcal{F}$ be a set of charges where $m(z, \beta_0, \gamma(F))$ is well defined.

DEFINITION 3: *$m(z, \beta, \gamma)$ is locally robust if and only if* $E[m(z_i, \beta_0, \gamma(F))] = o(\|F - F_0\|)$ for $F \in \mathcal{F}$.

Definition 1 requires that $\mu(F)$ have a zero pathwise derivative. Definition 3 requires a zero Frechet derivative for the semi-norm $\|\cdot\|$, generally a stronger condition than a zero pathwise derivative. The zero Frechet derivative condition is helpful in explaining the bias properties of locally robust moment functions.

Generally a first estimator will depend on some vector of smoothing parameters b. This b could be the bandwidth in a kernel estimator or the inverse of number of terms in a series estimator. Suppose that the limit of $\hat{\gamma}$ for fixed b is $\gamma(F_b)$ where F_b is a "smoothed" version of the true distribution that approaches the truth F_0 as $b \longrightarrow 0$. Then under regularity conditions $\hat{m}(\beta_0)$ will have limit $E[m(z_i, \beta_0, \gamma(F_b))]$. We can think of $\|F_b - F_0\|$ as a measure of the smoothing bias of F_b. Similarly $E[m(z_i, \beta_0, \gamma(F_b))]$ is a measure of the bias in the moment conditions caused by smoothing. The small bias property (SBP) analyzed in NHR is that the expectation of the moment functions vanishes faster than the nonparametric bias as $b \longrightarrow 0$.

DEFINITION 4: *$m(z, \beta, \gamma)$ and F_b have the small bias property if and only if* $E[m(z_i, \beta_0, \gamma(F_b))] = o(\|F_b - F_0\|)$ as $b \longrightarrow 0$.

As long as $F_b \in \mathcal{F}$, the set $\mathcal{F}$ in Definition 3, locally robust moments will have bias that vanishes faster than the nonparametric bias $\|F_b - F_0\|$ as $b \longrightarrow 0$. Thus locally robust moment functions have the small bias property.

PROPOSITION 10: *If $m(z, \beta, \gamma)$ is locally robust then $m(z, \beta, \gamma)$ has the small bias property for any $F_b \in \mathcal{F}$.*

Note that the bias of locally robust moment conditions will be flat as a function of the first step smoothing bias $\|F_b - F_0\|$ as that goes to zero. This flatter moment bias can also make the MSE flatter, meaning the MSE of the estimator does not depend as strongly on $\|F_b - F_0\|$ for locally robust moments as for other estimators.

By comparing Definitions 3 and 4 we see that the small bias property is a form of directional local robustness, with the moment being locally robust in the direction F_b. If the moments are not locally robust then there will be directions where the bias of the moments is not smaller than the smoothing bias F_b. Being locally robust in all directions can be important when the first step is allowed to be very flexible, such as when machine learning is used to construct the first step. There the first step can vary randomly across a large class of functions making the use of locally robust moments important for correct inference, e.g. see Belloni, Chernozhukov, and Hansen (2014).

This discussion of smoothing bias is based on sequential asymptotics where we consider limits for fixed b. This discussion provides useful intuition but it is also important to consider asymptotics where b could be changing with the sample size. We can analyze the precise effect of using locally robust moments by considering an expansion of the average moments. Let $\bar{m} = \sum_{i=1}^n m(z_i, \beta_0, \gamma_0)/n$, $\bar{g} = \sum_{i=1}^n g(z_i, \beta_0, \gamma_0)/n$, $\phi(z) = \phi(z, \beta_0, \gamma_0)$, and $\tilde{F}$ denote the empirical distribution. We suppose that $\hat{\gamma} = \gamma(\hat{F})$ for some estimator $\hat{F}$ of the true distribution F_0. Let $\mu(F) = E[m(z_i, \beta_0, \gamma(F))]$. By adding and subtracting terms we have

$$\hat{m}(\beta_0) = \bar{g} + \tilde{R}_1 + \hat{R}_2 + \hat{R}_3, \tilde{R}_1 = \int \phi(z)[\hat{F} - \tilde{F}](dz), \tag{5.1}$$

$$\hat{R}_2 = \mu(\hat{F}) - \int \phi(z)\hat{F}(dz), \hat{R}_3 = \hat{m}(\beta_0) - \bar{m} - \mu(\hat{F}).$$

The object $\int \phi(z)\hat{F}(dz) = \int \phi(z)[\hat{F} - F_0](dz)$ is a linearization of $\mu(\hat{F}) = \mu(\hat{F}) - \mu(F_0)$ so $\hat{R}_2$ is a nonlinearity remainder that is second order. Also $\hat{R}_3$ is a stochastic equicontinuity remainder of a type familiar from Andrews (1994) that is also second order.

The locally robust counterpart $\hat{g}(\beta_0)$ to $\hat{m}(\beta_0)$ has a corresponding remainder term that is asymptotically smaller than $\tilde{R}_1$. To see this let $\hat{\phi}(z) = \phi(z, \beta_0, \gamma(\hat{F}))$ and note that the mean zero property of an influence function will generally give $\int \hat{\phi}(z)\hat{F}(dz) = 0$. Then by $\hat{g}(\beta_0) = \hat{m}(\beta_0) + \int \hat{\phi}(z)\tilde{F}(dz)$ we have

PROPOSITION 11: *If* $\int \hat{\phi}(z)\hat{F}(dz) = 0$ *then* $\hat{g}(\beta_0) = \bar{g} + \hat{R}_1 + \hat{R}_2 + \hat{R}_3$,

$$\hat{R}_1 = -\int [\hat{\phi}(z) - \phi(z)][\hat{F} - \tilde{F}](dz).$$

Comparing this conclusion with equation (5.1) we see that locally robust moments have the same expansion as the original moments except that the remainder $\tilde{R}_1$ has been replaced by the remainder $\hat{R}_1$. The remainder $\hat{R}_1$ will be asymptotically smaller than $\tilde{R}_1$ under sufficient regularity conditions. Consequently, depending on cross correlations with other terms, the locally robust moments $\hat{g}(\beta_0)$ can be more accurate than $\hat{m}(\beta_0)$. For instance, as shown by NHR the locally robust moments for linear kernel averages have a higher order bias term that converges to zero at a faster rate than the original moments, while only the constant term in the higher order variance is larger. Consequently, the locally robust estimator will have smaller MSE asymptotically for appropriate choice of bandwidth. In nonlinear cases the use of locally robust moments may not lead to an improvement in MSE because nonlinear remainder terms may be important, see Robins et. al. (2008) and Cattaneo and Jansson (2014). Nevertheless, using locally robust moments does make smoothing bias small, which can be an important improvement.

In some settings it is possible to obtain a corresponding improvement by changing the first step estimator. For example, as mentioned earlier, for linear kernel averages the locally robust estimator is identical to the original estimator based on a twicing version of the original kernel (see NHR). The improvement from changing the first step can be explained in relation to the remainder $\tilde{R}_1$, that is the difference of the integral of $\phi(z)$ over the estimated distribution $\hat{F}$ and its sample average. Note that $\hat{F} - \tilde{F}$ will be shrinking to zero so that $\tilde{R}_1 - E[\tilde{R}_1]$ should be a second order (stochastic equicontinuity) term. $E[\tilde{R}_1]$ is the most interesting term. If $E[\hat{F}] = F_b$ and integrals can be interchanged then

$$E[\tilde{R}_1] = \int \phi(z)F_b(dz) = \int \phi(z)[F_b - F_0](dz).$$

When a twicing kernel or any other higher order kernel is used this remainder becomes second order, depending on the smoothness of both the true distribution F_0 and the influence function $\phi(z)$, see NHR and Bickel and Ritov (2003). Thus, by using a twicing or higher order kernel we obtain a second order bias, so all of the remainder terms are second order. Furthermore, series estimator automatically have a second order bias term, as pointed out in Newey (1994). Consequently, for all of these first steps the remainders are all second order even though the moment function is not locally robust.

The advantage of locally robust moments are that the improvement applies to any first step estimator. One does not have to depend on the particular structure of the estimator, such as having a kernel of sufficiently high order. This feature is important when the first step is

complicated so that it is hard to analyze the properties of terms that correspond to $E[\tilde{R}_1]$. Important examples are first steps that use machine learning. In that setting locally robust moments are very important for obtaining root-n consistency; see Belloni et. al. (2014). Locally robust moments have the advantages we have discussed even for very complicated first steps.

6 First Step Series Estimators

First step series estimators have certain automatic robustness properties. Moment conditions based on series estimators are automatically locally robust in the direction of the series approximation. We also find that affine moment functions are automatically doubly robust in these directions. In this Section we present these results.

It turns out that for certain first step series estimators there is a version of the adjustment term that has sample mean zero, so that $\hat{g}(\beta) = \hat{m}(\beta)$. That is, locally robust moments are numerically identical to the original moments. This version of the adjustment term is constructed by treating the first step as if it were parametric with parameters given by those of the series approximation, and calculating a sample version of the adjustment described in Section 2. Suppose that the coefficients $\hat{\lambda}$ of the first step estimator satisfy $\sum_{i=1}^n h(z_i, \hat{\lambda})/n = 0$. Let $\hat{M}_\lambda(\beta) = n^{-1} \sum_{i=1}^n \partial m(z_i, \beta, \hat{\lambda})/\partial\lambda$, $\hat{H} = n^{-1} \sum_{i=1}^n \partial h(z_i, \hat{\lambda})/\partial\lambda$, and

$$\phi(z, \beta, \hat{\gamma}) = -\hat{M}_\lambda(\beta)\hat{H}^{-1}h(z, \hat{\lambda}) \tag{6.1}$$

be the parametric adjustment term described in Section 2, where $\hat{\gamma}$ includes the elements of $\hat{M}_\gamma(\beta)$ and $\hat{H}$ and there is no cross fitting. Note that

$$\frac{1}{n}\sum_{i=1}^n \phi(z_i, \beta, \hat{\gamma}) = -\hat{M}_\lambda(\beta)\hat{H}^{-1}\frac{1}{n}\sum_{i=1}^n h(z_i, \hat{\lambda}) = 0.$$

It follows that $\hat{g}(\beta) = \hat{m}(\beta)$, i.e. the locally robust moments obtained by adding the adjustment term are identical to the original moments. Thus, if $\sum_{i=1}^n h(z_i, \hat{\lambda})/n = 0$, we treat the first step series estimator as parametric, and use the parametric adjustment term the locally robust moments are numerically identical to the original moments. This numerical equivalence results is an exact version of local robustness of the moments in the direction of the series approximation.

In some settings it is known that $\phi(z, \beta, \hat{\gamma})$ in equation (6.1) is an estimated approximation to $\phi(z_i, \beta, \gamma_0)$, justifying its use. Newey (1994, p. 1369) showed that this approximation property holds when the first step is a series regression. Ackerberg, Chen, and Hahn (2012) showed that this property holds when the first step satisfies certain conditional moment restrictions or is part of a sieve maximum likelihood estimator. It is also straightforward to show that this approximation holds when the first step is a series approximates to the solution to the conditional moment restriction $E[\rho(z_i, \gamma_{10})|x_i] = 0$. We expect that in general $\phi(z, \beta, \hat{\gamma})$ is an estimator of $\phi(z, \beta, \gamma_0)$.

We note that the result that $\hat{g}(\beta) = \hat{m}(\beta)$ is dependent on $\hat{\lambda}$ not varying with the observations and on being constructed from the whole sample. If we use cross fitting in any form then the numerical equivalence of the original moments with their locally robust counterpart will generally not hold. Also $\hat{g}(\beta) \neq \hat{m}(\beta)$ will generally occur when different models are used for different elements of γ. Such different models will often be present when machine learning is used for constructing the estimators of the different elements of γ. See for example Chernozhukov et. al. (2016).

There are interesting cases where the original moment functions $m(z, \beta, \gamma)$ with a series estimator for γ are doubly robust in certain directions, with $E[m(z_i, \beta_0, \gamma_1)] = 0$ when γ_1 is a series approximation to γ_{10}. Here we show this directional double robustness property for series estimators of solutions to conditional moment restrictions and orthogonal series density estimators. Consider first a conditional moment restriction where the residual $\rho(z, \gamma_1)$ is affine in γ_1, $m(z, \beta, \gamma_1)$ is also affine in γ_1, and the first step is a linear series estimator. Suppose that the series estimator approximates γ_{10} by a linear combination $p^{K\prime}\lambda$ of a vector of functions $p^K(w) = (p_{1K}(w), ..., p_{KK}(w))'$. Let $q^K(x_i)$ be a $K \times 1$ vector of instrumental variables and $\hat{\lambda}$ be the instrumental variables estimator solving a moment condition $\sum_{i=1}^n q^K(x_i)\rho(z_i, p^{K\prime}\hat{\lambda}) = 0$. Under standard regularity conditions the limit λ_* of $\hat{\lambda}$ will solve the corresponding population moment condition $E[q^K(x_i)\rho(z_i, p^{K\prime}\lambda_*)] = 0$. Let $\gamma_{20}(x)$ satisfy equation (4.2). Then if $\gamma_{20}(x_i) = \theta' q^K(x_i)$ for some θ it follows that

$$E[m(z_i, \beta_0, p^{K\prime}\lambda_*)] = -E[\gamma_{20}(x_i)\rho(z_i, p^{K\prime}\lambda_*)] = -\theta' E[q^K(x_i)\rho(z_i, p^{K\prime}\lambda_*)] = 0.$$

Thus we have the result

PROPOSITION 12: *If $m(z_i, \beta, \gamma_1)$ and $\rho(z_i, \gamma_1)$ are affine in $\gamma_1 \in \Gamma$ with Γ linear, $E[m(z_i, \beta_0, \gamma_1)]$ is a mean square continuous functional of $E[\rho(z_i, \gamma_1)|x_i]$, and $\gamma_{20}(x_i)$ satisfying $E[m(z_i, \beta_0, \gamma_1)] = -E[\gamma_{20}(x_i)\rho(z_i, \gamma_1)]$ also satisfies $\gamma_{20}(x_i) = \theta' q^K(x_i)$ for some θ then $E[m(z_i, \beta_0, p^{K\prime}\lambda_*)] = 0$.*

The property shown in the conclusion of Proposition 12 is a directional double robustness condition that depends on γ_1 being equal to a series approximation to γ_{10} and on $\gamma_{20}(x)$ being restricted. These restrictions are not required for double robustness of $g(z, \beta, \gamma) = m(z, \beta, \gamma_1) + \gamma_2(x)\rho(z, \gamma_1)$. We will have $E[g(z_i, \beta_0, \gamma_{20}, \gamma_1)] = 0$ for all γ_1, and not just for the γ_1 that are a series approximation to γ_{10}, and for any $\gamma_{20}(x)$ and not just one that is a linear combination of $q^K(x)$. For series first steps the the original moment functions will be doubly robust in certain directions just as they are locally robust in certain directions.

Previous examples of Proposition 12 are given in Newey (1990, p. 116), Newey (1999), and Robins et. al. (2007). Proposition 12 allows for endogeneity and $m(z, \beta, \gamma_1)$ to depend on the entire function γ_1. The condition that the instruments $q^K(x_i)$ have the same dimension as the approximating functions $p^K(w)$ allows for more than K instrumental variables. As is well known,

any IV estimator of λ can be viewed as having only K instruments $q^K(x)$, each one of which is equal to a linear combination of all the instrumental variables. Here the existence of θ such that $\gamma_{20}(x_i) = \theta' q^K(x_i)$ is restrictive. It is not sufficient that $\gamma_{20}(x_i)$ be any linear combination of all the instrumental variables. We must have $\gamma_{20}(x_i)$ equal to a linear combination of the instruments $q^K(x_i)$ used in estimating λ_*. This result also extends to the case where an infinite number of instrumental variables are used in the limit. In that case $q^K(x_i)$ can also be interpreted as an infinite dimensional linear combination of instrumental variables.

To illustrate, consider again the weighted average derivative example discussed above where $m(z, \beta, \gamma_1) = v(w)\partial\gamma_1(w)/\partial w - \beta$, $\rho(z, \gamma_1) = y - \gamma_1(w)$, and there is $\gamma_{20}(x)$ such that

$$E[\gamma_{20}(x)|w] = -f_0(w)^{-1}\partial[f_0(w)v(w)]/\partial w. \tag{6.2}$$

Suppose that the first step is a linear instrumental variables (IV) estimator with right-hand side variables $p^K(w_i)$ and instruments $q^K(x_i)$ and let λ_* be the limit of the IV coefficients. From Proposition 12 it follows that if there is a θ such that $\theta' q^K(x) = \gamma_{20}(x)$ then

$$E[v(w_i)\left\{\partial p^K(w_i)'\lambda_*\right\}/\partial w] - \beta_0 = E[m(z_i, \beta_0, p^{K\prime}\lambda_*)] = 0.$$

Thus, the weighted average derivative of the linear IV estimator will be consistent when $\gamma_{20}(x)$ is a linear combination of $q^K(x)$.

The case where $v(w)$ is 1, w_i is Gaussian, and $E[x_i|w_i]$ is linear in w_i is interesting. Partial out constants and means so that $(w_i', x_i')'$ has mean zero. Let $p^K(w_i) = w_i$ and let $q_i = q^K(x_i)$ be any linear combination x_i such that $\zeta = E[q_i w_i']$ is nonsingular. Normalize w_i and q_i so that each have an identity variance matrix. Then $f_0(w)^{-1}\partial f_0(w)/\partial w = -w$. Note that $E[q_i|w_i] = \zeta w_i$, so that equation (6.2) is satisfied with $\gamma_{20}(x) = -\zeta^{-1}q^K(x)$.Thus the conditions of Proposition 12 are satisfied, giving the following result

COROLLARY 13: *If $y_i = \gamma_{10}(w_i) + \varepsilon_i$, $E[x_i\varepsilon_i] = 0$, $E[y_i^2] < \infty$, w_i is Gaussian, and $E[x_i|w_i]$ is linear in w_i then for instruments equal to any linear combination q_i of x_i with $cov(q_i, w_i)$ nonsingular*,

$$E[\partial\gamma_{10}(w_i)/\partial w] = cov(q_i, w_i)^{-1}cov(q_i, y_i).$$

We can give a simple, direct proof that only uses q_i and ε_i uncorrelated, as is assumed in Corollary 11, rather than the conditional moment restriction we have been focusing on. With means partialed out we have $E[w_i] = E[q_i] = 0$ so that

$$\begin{aligned} cov(q_i, w_i)^{-1}cov(q_i, y_i) &= (E[q_i w_i'])^{-1}E[q_i y_i] = (E[q_i w_i'])^{-1}E[q_i\gamma_{10}(w_i)] \\ &= (E[E[q_i|w_i]w_i'])^{-1}E[E[q_i|w_i]\gamma_{10}(w_i)] = (E[\zeta w_i w_i'])^{-1}E[\zeta w_i\gamma_{10}(w_i)] \\ &= (E[w_i w_i'])^{-1}E[w_i\gamma_{10}(w_i)] = E[\partial\gamma_{10}(w_i)/\partial w], \end{aligned}$$

where the fourth equality follows by $E[q_i|w_i] = \zeta w_i$ and the last equality holds by Stoker (1986) and w_i Gaussian. This result generalizes that of Stoker (1986) to NPIV models where the right hand variables are Gaussian. Further generalizations to non Gaussian cases can be obtained by letting $q^K(x)$ and $p^K(w)$ be nonlinear in x and w.

Orthogonal series density estimators have a property analogous to Proposition 12. Suppose now that $p^K(x)$ is orthonormal with respect to Lebesgue measure on $(-\infty, \infty)$ so that $\int p^K(u)p^K(u)'du = I$. An orthogonal series pdf estimator is $\hat{\gamma}_1(x) = p^K(x)'\hat{\lambda}$, where $\hat{\lambda} = \sum_{i=1}^n p^K(x_i)/n$ has limit $\lambda_* = \int p^K(u)\gamma_{10}(u)du$. Suppose that $E[m(z_i, \beta_0, \gamma_1)]$ is a continuous linear functional of $\gamma_1 - \gamma_{10}$ so that by the Riesz representation theorem there is $\gamma_{20}(x)$ with $E[m(z_i, \beta_0, \gamma_1)] = \int \gamma_{20}(u)[\gamma_1(u) - \gamma_{10}(u)]du$. If there is θ with $\theta' p^K(x) = \gamma_{20}(x)$ then by $p^K(u)$ orthonormal and equation (4.4) we have

$$E[m(z_i, \beta_0, p^{K\prime}\lambda_*)] = \int \gamma_{20}(u)[p^K(u)'\lambda_* - \gamma_{10}(u)]du = \int \theta' p^K(u)[p^K(u)'\lambda_* - \gamma_{10}(u)]du$$
$$= \theta'[\int p^K(u)p^K(u)'du]\lambda_* - \theta' \int p^K(u)\gamma_{10}(u)du = \theta'\lambda_* - \theta'\lambda_* = 0.$$

Thus we have the following result:

PROPOSITION 14: *If i)* $m(z, \beta_0, \gamma_1)$ *is affine in* γ_1*, ii)* $E[m(z_i, \beta_0, \gamma_1)] - \beta_0$ *is a functional of* $\gamma_1(x) - \gamma_{10}(x)$ *that is continuous in the norm* $\left(\int [\gamma_1(u) - \gamma_{10}(u)]^2 du\right)^{1/2}$*, and iii)* $\gamma_{20}(x_i) = \theta' p^K(x_i)$ *for some* θ *then* $E[m(z_i, \beta_0, p^{K\prime}\lambda_*)] = 0$.

The orthogonal series estimators of linear functionals of a pdf discussed in Bickel and Ritov (2003) are examples. Those estimators are special cases of the estimator above where $m(z, \beta, \gamma_1) = \int \gamma_{20}(u)\gamma_1(u)du - \beta$ for prespecified $\gamma_{20}(u)$. Proposition 14 implies that the orthogonal series estimator of β_0 will be consistent if $\gamma_{20}(u)$ is a linear combination of the approximating functions. For example, if $p^K(x)$ is vector of polynomials of order $K-1$ then the orthogonal series estimator of moments of up to order $K-1$ is consistent for fixed K.

7 Conditional Moment Restrictions

Conditional moment restrictions are widely used in econometrics to identify parameters of interest. In this Section we expand upon the cases already considered to construct a wide variety of locally robust moment conditions. In particular we extend above results to residuals that may depend on parameters of interest with instrumental variables that can differ across residuals. Here we depart from deriving locally robust moments from the adjustment term for first step estimation. Instead we extend the form of previously derived locally robust moments to the more general setting of this Section.

To describe these results let $j = 2, ..., J$ index conditional moment restrictions, $\rho_j(z, \beta, \gamma_1)$ denote a corresponding residual, and x_j be corresponding conditioning variables. We will consider construction of locally robust moment conditions when the true parameters of interest β_0 and a first step γ_{10} satisfy conditional moment restrictions

$$E[\rho_j(z_i, \beta_0, \gamma_{10})|x_{ji}] = 0, (j = 2, ..., J). \tag{7.1}$$

Here γ_1 is specified to include all functions that affect any of the residuals $\rho_j(z_i, \beta, \gamma_1)$. We continue to assume that the unconditional moment restriction in equation (2.1) holds, though $m(z, \beta, \gamma_1)$ could be zero, with identification of β_0 coming from the conditional moment restrictions of equation (7.1). We will discuss this case below.

In this setting we consider locally robust moment conditions having the form

$$g(z, \beta, \gamma) = m(z, \beta, \gamma_1) + \sum_{j=2}^{J} \gamma_j(x_j, \beta)\rho_j(z, \beta, \gamma_1), \tag{7.2}$$

where $\gamma_j(x, \beta), (j = 2, ..., J)$ are unknown functions satisfying properties discussed below. These moment functions depend on J first step components $\gamma = (\gamma_1, ..., \gamma_J)$. By virtue of the conditional moment restrictions these moment functions will be doubly robust in $(\gamma_2, ..., \gamma_J)$, meaning that $E[g(z_i, \beta_0, \gamma_{10}, \gamma_2, ..., \gamma_J) = 0$. They will be locally robust in γ_1 if for the limit $\gamma_1(F)$ of $\hat{\gamma}_1$ and all regular parametric models F_τ as discussed in Section 2,

$$\frac{\partial}{\partial \tau} E[m(z_i, \beta_0, \gamma_1(F_\tau))] + E[\sum_{j=2}^{J} \gamma_{j0}(x_{ji}) \frac{\partial}{\partial \tau} E[\rho_j(z_i, \beta_0, \gamma_1(F_\tau))|x_{ji}]] = 0. \tag{7.3}$$

If $\partial E[m(z_i, \beta_0, \gamma_1(F_\tau))]/\partial \tau|_{\tau=0}$ is a linear mean-square continuous function of

$$(\partial E[\rho_2(z_i, \beta_0, \gamma_1(F_\tau))|x_{ji}]/\partial \tau, ..., \partial E[\rho_J(z_i, \beta_0, \gamma_1(F_\tau))|x_{ji}]/\partial \tau)|_{\tau=0}$$

and the mean-square closure of the set of such vectors over all parametric submodels is linear then existence of $\gamma_{j0}(x_j)$, $j \geq 2$ satisfying equation (7.3) will follow by the Riesz representation theorem. In addition, if $m(z, \beta_0, \gamma_1)$ and $\rho_j(z, \beta_0, \gamma_1)$, $(j \geq 2)$, are affine in γ_1 then we will have double robustness in γ_1 similarly to Proposition 12. Summarizing we have

PROPOSITION 15: *If equation (7.3) is satisfied then $g(z, \beta, \gamma)$ from equation (7.2) is locally robust. Furthermore, if i) $m(z, \beta_0, \gamma_1)$ and $\rho_j(z, \beta_0, \gamma_1), (j \geq 2)$ are affine in $\gamma_1 \in \Gamma$ with Γ linear; ii) $E[m(z_i, \beta_0, \gamma_1)]$ is a mean square continuous functional of $E[\rho_j(z_i, \beta_0, \gamma_1)|x_{ij}], (j \geq 2)$ then there is $\gamma_{j0}(x), (j \geq 2)$, such that $g(z, \beta, \gamma)$ is doubly robust.*

For local identification of β we also require that

$$rank(\partial E[g(z_i, \beta, \gamma_0)]/\partial \beta|_{\beta=\beta_0}) = \dim(\beta). \tag{7.4}$$

A model where β_0 is identified from semiparametric conditional moment restrictions with common instrumental variables x is a special case where $m(z,\beta,\gamma)$ is zero and $x_j = x, (j \geq 2)$. In this case let $\rho(z,\beta,\gamma_1) = (\rho_2(z,\beta,\gamma_1),...,\rho_J(z,\beta,\gamma_1))'$. The conditional moment restrictions of equation (7.1) can be summarized as

$$E[\rho(z_i,\beta_0,\gamma_{10})|x_i] = 0.$$

This model is considered by Chamberlain (1992) and Ai and Chen (2003, 2007, 2012). We allow the residual vector $\rho(z,\beta,\gamma_1)$ to depend on the entire function γ_1 and not just its value at some function of the observed data z_i. Also let $\varphi(x) = [\gamma_2(x),...,\gamma_J(x)]$ denote an $r \times (J-1)$ matrix of functions of x. A locally robust moment function $g(z,\beta,\gamma) = \varphi(x)\rho(z,\beta,\gamma_1)$ will be one which satisfies Definition 1 with $g(z,\beta,\gamma)$ replacing $m(z,\beta,\gamma)$, i.e. where

$$\frac{\partial E[g(z_i,\beta_0,\gamma(F_\tau))]}{\partial \tau} = E\left[\varphi(x_i)\frac{\partial E[\rho(z_i,\beta_0,\gamma_1(F_\tau))|x_i]}{\partial \tau}\right] = 0,$$

for all regular parametric models. We also require that equation (7.4) is satisfied.

To characterize local robustness here it is helpful to assume that the set of pathwise derivatives of $E[\rho(z_i,\beta_0,\gamma)|x_i]$ varies over a linear set as the regular parametric model F_τ varies. To be precise we will assume that $\gamma_1 \in \Gamma$ for Γ linear and for $\Delta \in \Gamma$ we let

$$m_\gamma(x,\Delta) = \left.\frac{\partial E[\rho(z_i,\beta_0,\gamma_{10}+\tau\Delta)|x_i]}{\partial \tau}\right|_{\tau=0}$$

denote the $(J-1)\times 1$ random vector that is Gateaux derivative of the conditional expectation $E[\rho(z_i,\beta_0,\gamma_1)|x_i]$ with respect to the first step γ_1 in the direction Δ. We assume that $m_\gamma(x,\Delta)$ is linear in Δ and that the mean square closure $\mathcal{M}_\gamma$ of the set $\{m_\gamma(x,\Delta) : \Delta \in \Gamma\}$ equals the mean-square closure of the set $\{\partial E[\rho(z_i,\beta_0,\gamma_1(F_\tau))|x_i]/\partial\tau\}$ as F_τ varies over all regular parametric models. The local robustness condition can then be interpreted as orthogonality of each row $\varphi(x_i)'e_j$ of $\varphi(x)$ with $\mathcal{M}_\gamma$ in the Hilbert space of functions of x with inner product $\langle a,b\rangle = E[a(x_i)'b(x_i)]$, where e_k is the k^{th} unit vector. Thus the condition for locally robust $g(z,\beta,\gamma) = \varphi(x)\rho(z,\beta,\gamma_1)$ is that

$$E[\varphi(x_i)m_\gamma(x_i,\Delta)] = 0 \text{ for all } \Delta \in \Gamma.$$

We refer to such $\varphi(x_i)$ as being orthogonal. They can be interpreted as instrumental variables where the effect of estimation of γ_1 has been partialed out.

There are many ways to construct orthogonal instruments. For instance, given a $r \times (J-1)$ matrix of instrumental variables $A(x)$ one could construct corresponding orthogonal ones $\varphi(x_i)$ as the matrix where each row is the residual of the least squares projection of the corresponding row of $A(x)$ on $\mathcal{M}_\gamma$. We focus on another way of constructing orthogonal instruments that leads

to an efficient estimator of β_0. Let $\Sigma(x)$ denote some positive definite matrix with smallest eigenvalue bounded away from zero, so that $\Sigma(x_i)^{-1}$ is bounded. Let $\langle a,b\rangle_\Sigma = E[a(x_i)'\Sigma(x_i)^{-1}b(x_i)]$ denote an inner product and note that $\mathcal{M}_\gamma$ is closed in this inner product by $\Sigma(x_i)^{-1}$ bounded. Let $\tilde{A}_k(x_i,A,\Sigma)$ denote the residual from the least squares projection of the k^{th} row $A(x)'e_k$ of $A(x)$ on $\mathcal{M}_\gamma$ with the inner product $\langle a,b\rangle_\Sigma$. Also let $\varphi(x_i,A,\Sigma)$ be the matrix with k^{th} row $\tilde{A}_k(x_i,A,\Sigma)'\Sigma(x_i)^{-1}$, $(k=1,...,r)$. Then for all $\Delta\in\Gamma$,

$$A(x_i)'e_k - \tilde{A}_k(x_i,A,\Sigma)\in\mathcal{M}_\gamma, E[\varphi(x_i,A,\Sigma)m_\gamma(x_i,\Delta)] = E[\tilde{A}(x_i,A,\Sigma)\Sigma(x_i)^{-1}m_\gamma(x_i,\Delta)] = 0,$$

so that $\varphi(x_i,A,\Sigma)$ are orthogonal instruments. Also, $\tilde{A}(x,A,\Sigma)$ can be interpreted as the solution to

$$\min_{\{M(x):M(x)'e_k\in\mathcal{M}_\gamma,k=1,...,r\}} E[\{A(x_i)-M(x_i)\}\Sigma(x_i)^{-1}\{A(x_i)-M(x_i)\}']$$

where the minimization is in the positive semidefinite sense.

The orthogonal instruments that minimize the asymptotic variance of GMM in the class of GMM estimators with orthogonal instruments are given by

$$\varphi^*(x_i)=\varphi(x_i,A^*,\Sigma^*), A^*(x_i)=\left.\frac{\partial E[\rho(z_i,\beta,\gamma_{10})|x_i]}{\partial\beta}\right|'_{\beta=\beta_0}, \Sigma^*(x_i)=Var(\rho(z_i,\beta_0,\gamma_{10})|x_i).$$

To see that $\varphi^*(x_i)$ minimizes the asymptotic variance note that for any orthogonal instrumental variable matrix $\varphi(x)$

$$G=E[\varphi(x_i)A^*(x_i)'] = E[\varphi(x_i)\tilde{A}(x_i,A^*,\Sigma^*)'] = E[\varphi(x_i)\rho(z_i,\beta_0,\gamma_{10})\rho(z_i,\beta_0,\gamma_{10})'\varphi^*(x_i)'],$$

where the first equality defines G and the second equality holds by $\varphi(x_i)$ orthogonal. Since the instruments are orthogonal the asymptotic variance matrix of GMM estimator with $\hat{W}\xrightarrow{p}W$ is the same as if $\hat{\gamma}_1=\gamma_{10}$. Define $m_i=G'W\varphi(x_i)\rho(z_i,\beta_0,\gamma_{10})$ and $m_i^*=\varphi^*(x_i)\rho(z_i,\beta_0,\gamma_{10})$. The asymptotic variance of the GMM estimator for orthogonal instruments $\varphi(x)$ is

$$\begin{aligned}&(G'WG)^{-1}G'WE[\varphi(x_i)\rho(z_i,\beta_0,\gamma_{10})\rho(z_i,\beta_0,\gamma_{10})'\varphi(x_i)']WG(G'WG)^{-1}\\ &=(E[m_im_i^{*\prime}])^{-1}E[m_im_i'](E[m_im_i^{*}])^{-1\prime}.\end{aligned}$$

The fact that this matrix is minimized in the positive semidefinite sense for $\varphi(x)=\varphi^*(x)$ follows from Theorem 5.3 of Newey and McFadden (1994) and can also be shown using the argument in Chamberlain (1987).

PROPOSITION 16: *The instruments $\varphi^*(x_i)$ give an efficient estimator in the class of IV estimators with orthogonal instruments.*

The asymptotic variance of the GMM estimator with optimal orthogonal instruments is

$$(E[m_i^*m_i^{*\prime}])^{-1} = E[\tilde{A}(x_i,A^*,\Sigma^*)\Sigma^*(x_i)^{-1}\tilde{A}(x_i,A^*,\Sigma^*)'])^{-1}.$$

This matrix coincides with the semiparametric variance bound of Ai and Chen (2003). Estimation of the optimal orthogonal instruments is beyond the scope of this paper. The series estimator of Ai and Chen (2003) could be used for this.

8 Structural Economic Examples

Estimating structural models can be difficult when that requires computing equilibrium solutions. Motivated by this difficulty there is increasing interest in two step semiparametric methods based on first step estimation of conditional choice probabilities (CCP). This two step approach was pioneered by Hotz and Miller (1993). In this Section we show how locally robust moment conditions can be formed for two kinds of structural models, the dynamic discrete choice model of Rust (1987) and the static model of strategic interactions of Bajari, Hong, Krainer, and Nekipelov (2010, BHKN). It should be straightforward to extend the construction of locally robust moments to other more complicated structural economic models. The use of such moment conditions will allow for conditional choice probabilities that are estimated by modern, machine learning methods.

8.1 Static Models of Strategic Interactions

We begin with a static model of interactions where results are relatively simple. To save space we describe the estimator of BHKN while only describing a small part of the motivational economic structure. Let x denote a vector of state variables for a fixed set of individuals and let y denote a vector of binary variables, each one representing a choice of an alternative by an individual. Let the observations $z_i = (y_i, x_i)$ represent repeated plays of a static game of interaction and $\gamma_{10}(x) = E[y_i|x_i = x]$ the vector of conditional choice probabilities given a value x of the state. In the semiparametric estimation problem described in Section 4.2 of BHKN there is a known function $r(x, \beta, \gamma_1(x))$ of the state variable x, a vector of parameters β, and a possible value $\gamma(x)$ of the conditional choice probabilities such that the true parameter β_0 satisfies

$$E[y_i|x_i = x] = r\left(x, \beta_0, \gamma_{10}\left(x\right)\right),$$

This model can be used to form moment functions

$$m(z, \beta, \gamma_1) = A(x)[y - r\left(x, \beta, \gamma_1\left(x\right)\right)],$$

where $A(x)$ is a matrix of instrumental variables; see equation (17) of BHKN.

To describe locally robust moment functions in this example, let $r_\gamma(x, \beta, \gamma_1) = \partial r(x, \beta, \gamma_1)/\partial\gamma_1$ where γ_1 here denotes a real vector representing a possible value of $\gamma_{10}(x)$. Then, it follows from

Proposition 4 of Newey (1994), as discussed in BHKN, that the adjustment term for first step estimation of $\gamma_{10}(x) = E[y_i|x_i = x]$ is

$$\phi(z, \beta, \gamma_1) = -A(x) r_\gamma(x, \beta, \gamma_1)[y - \gamma_1(x)].$$

This expression differs from BHKN in the appearance of $\gamma_1(x)$ at the end of the expression rather than $r_\gamma(x, \beta, \gamma_1(x))$, which is essential for local robustness. The locally robust moment conditions are then

$$g(z, \beta, \gamma_1) = A(x)\{y - r(x, \beta, \gamma_1(x)) - r_\gamma(x, \beta, \gamma_1(x))[y - \gamma_1(x)]\}.$$

For a first step estimator $\hat{\gamma}(x)$ of the conditional choice probabilities, the locally robust sample moments will be

$$\hat{g}(\beta) = \frac{1}{n} \sum_{i=1}^{n} A(x_i)\{y_i - r(x_i, \beta, \hat{\gamma}_1(x_i)) - r_\gamma(x_i, \beta, \hat{\gamma}_1(x_i))[y_i - \hat{\gamma}_1(x_i)]\}$$

Here the locally robust moments are constructed by subtracting from the structural residuals r a linear combination of the first step residuals. Using these moment functions should result in an estimator of the structural parameters with less bias and the other improved properties of locally robust estimators mentioned above.

The optimal instruments here are the same as discussed in BHKN. Let I denote the identity matrix, set $H(x) = I - \partial r(x_i, \beta_0, \gamma_{10}(x_i))/\partial \gamma_1$, and let $\Omega(x_i) = H(x_i) Var(y_i|x_i) H(x_i)^T$ denote the conditional variance of $H(x_i)(y_i - \gamma_{10}(x_i))$. The optimal instruments are given by

$$A^*(x_i) = \frac{\partial r(x_i, \beta_0, \gamma_{10}(x_i))'}{\partial \beta} \Omega(x)^-$$

where A^- denotes a generalized inverse of a positive semi-definite matrix A.

This model can also be viewed as a special case of the conditional moment restrictions framework with residual vector $\rho(z, \beta, \gamma_1) = (y - \gamma_1(x), y - r(x, \beta, \gamma_1(x)))^T$. An orthogonal instrument that gives the above locally robust moment function is $A(x)[-r_\gamma(x_i, \beta, \gamma_1(x_i)), I]$.

Here the locally robust moment function only depend on one first step function $\gamma_1(x)$. This feature is shared by all setups where the second step residual $r(x, \beta, \gamma_1)$ depends only on regressors that are included in the first step $\gamma_1(x)$. The static model of strategic interactions leads to this structure. The situation is not so simple in other structural economic models, as we see next.

8.2 Dynamic Discrete Choice

Dynamic discrete choice estimation is important for modeling economic decisions, Rust (1987). In this setting we find it helpful to describe the underlying economic model in order to explain

the form of the moment conditions. Here we give locally robust moment conditions moment conditions that depend on first step estimation of the conditional choice probabilities. We do this for the infinite horizon, stationary, dynamic discrete choice model of Rust (1987). It is straightforward to derive locally robust moment conditions for other structural econometric models. We also focus here on the case of data on many homogenous individuals, but discuss how the approach extends to time series on one individual.

Suppose that the per-period utility function for an agent making choice j in period t is given by

$$U_{jt} = v_j(x_t, \beta_0) + \epsilon_{jt}, (j = 1, ..., J; t = 1, 2, ...)$$

where we suppress the individual subscript i for notational convenience. The vector x_t is the observed state variables of the problem (*e.g.* work experience, number of children, wealth) and the vector β is unknown parameters. The disturbances $\epsilon_t = \{\epsilon_{1t}, ..., \epsilon_{Jt}\}$ are not observed by the econometrician. As in the majority of the literature we assume that ϵ_t is i.i.d. over time with known CDF that has support R^J and is independent of the state process x_t and that x_t is Markov of order 1. Let δ denote a time discount parameter, $\bar{v}(x)$ the expected value function, $y_{jt} \in \{0, 1\}$ the indicator that choice j is made and $\bar{v}_j(x_t) = v_j(x_t, \beta_0) + \delta E[\bar{v}(x_{t+1})|x_t, y_{tj} = 1]$ the expected value function for choice j. Also let $\tilde{v}_j$ denote a possible realization of $\tilde{v}_j(x_t) = \bar{v}_j(x_t) - \bar{v}_1(x_t)$, so that $\tilde{v}_1 \equiv 0$. Let $\tilde{v} = (\tilde{v}_2, ..., \tilde{v}_J)$ and $P_j(\tilde{v}) = \Pr(\tilde{v}_j + \epsilon_{jt} \geq \tilde{v}_k + \epsilon_{kt}; \tilde{v}_1 = 0; k = 1, ..., J)$, $(j = 1, ..., J)$ denote the choice probabilities associated with the distribution of ϵ_t. Here we normalize to focus on the difference with $\bar{v}_1(x)$ throughout. Let $\tilde{v}(x) = (\tilde{v}_2(x), ..., \tilde{v}_J(x))'$. Let $\gamma_a(x) = (\gamma_{a1}(x), ..., \gamma_{aJ}(x))^T$ be a vector of first step functions with true values $\gamma_{aj0}(x_t) = \Pr(y_{tj} = 1|x_t)$. From Rust (1987) we know that for

$$\gamma_{aj0}(x_t) = P_j(\tilde{v}(x_t)), (j = 1, ..., J).$$
$$\bar{v}(x_t) = E[\max_j\{\bar{v}_j(x_t) + \epsilon_{jt}\}|x_t] = \bar{v}_1(x_t) + E[\max_j\{\tilde{v}_j(x_t) + \epsilon_{jt}\}|x_t].$$

From Hotz and Miller (1993) we know that $P(\tilde{v}) = (P_1(\tilde{v}), P_2(\tilde{v}), ..., P_J(\tilde{v}))'$ is a one-to-one function of $\tilde{v}$, so that inverting this relationship it follows that $E[\max_j\{\tilde{v}_j(x_t) + \epsilon_{jt}\}|x_t]$ is a function of $\gamma_{a0}(x_t)$, say $E[\max_j\{\tilde{v}_j(x_t) + \epsilon_{jt}\}|x_t] = H(\gamma_{a0}(x_t))$, for some function $H(\cdot)$ (e.g. for binary logit $H(\gamma_{a0}(x_t)) = .5772 - \ln(\gamma_{a10}(x_t))$). Then the expected value function is given by

$$\bar{v}(x_t) = \bar{v}_1(x_t) + H(\gamma_{a0}(x_t))$$

To use these relationships to construct semiparametric moment conditions we normalize $v_1(x, \beta) = 0$ and make an additional assumption for $\bar{v}_1(x)$. The additional assumption is that $E[\bar{v}(x_{t+1})|x_t, 1]$ does not depend on x_t. With this normalization and assumption we have a constant choice specific value function for $j = 1$, that is $\bar{v}_1(x) = \bar{v}_1$, with

$$\bar{v}_1(x_t) = 0 + \delta E[\bar{v}(x_{t+1})|x_t, y_{t1} = 1] = \delta E[\bar{v}(x_{t+1})|y_{t1} = 1] = \bar{v}_1.$$

A sufficient condition for constant $\bar{v}_1(x_t)$ is that $j = 1$ is a "renewal" choice where the distribution of the future state does not depend on the current state. In the Rust (1987) example this state is the one where the bus engine is replaced.

With this normalization and assumption we now have

$$\begin{aligned}
\bar{v}_j(x) &= v_j(x, \beta_0) + \delta E[\bar{v}_1 + H(\gamma_{a0}(x_{t+1}))|x_t = x, y_{tj} = 1] \\
&= v_j(x, \beta_0) + \delta\bar{v}_1 + \delta E[H(\gamma_{a0}(x_{t+1}))|x_t = x, y_{tj} = 1], \\
\tilde{v}_j(x) &= v_j(x, \beta_0) + \delta E[H(\gamma_{a0}(x_{t+1}))|x_t = x, y_{tj} = 1] - \delta E[H(\gamma_{a0}(x_{t+1}))|y_{t1} = 1], (j = 2, ..., J).
\end{aligned}$$

The choice specific expected value differences $\tilde{v}_j(x)$ have a parametric part $v_j(x, \beta_0)$ and a nonparametric part that depends on $J - 1$ additional nonparametric regressions $\gamma_b(x, \gamma_a) = (\gamma_{b1}(x, \gamma_a), ..., \gamma_{b,J-1}(x, \gamma_a))^T$ and an unknown parameter γ_c where

$$\gamma_{bj0}(x, \gamma_a) = E[H(\gamma_a(x_{t+1}))|x_t = x, y_{t,j+1} = 1], j = 1, ..., J - 1; \gamma_{c0}(\gamma_a) = E[H(\gamma_a(x_{t+1}))|y_{t1} = 1].$$

Let $\gamma_1(x) = (\gamma_a(x)^T, \gamma_b(x, \gamma_a)^T, \gamma_c(\gamma_a))^T$ be a vector of first step objects and

$$\tilde{v}_j(x, \beta, \gamma_1) = v_j(x, \beta) + \delta[\gamma_{b,j}(x, \gamma_a) - \gamma_c(\gamma_a)], \tilde{v}(x, \beta, \gamma_1) = (\tilde{v}_2(x, \beta, \gamma_1), ..., \tilde{v}_J(x, \beta, \gamma_1))^T$$

denote the semiparametric choice specific expected value differences. Semiparametric moment conditions can then be formed by plugging in a nonparametric $\hat{\gamma}_1$ estimator of the first step into the expected differences and plugging those into the choice probability. Let $y_t = (y_{t1}, ..., y_{tJ})^T$ denote the vector of choice indicators for period t and $z = (y_1^T, x_1^T, ..., y_T^T, x_T^T)^T$ be the vector consisting of the observations on choice and state variables x_t for each time period t. Also let $A_t(x_t)$ be an $r \times 1$ vector of functions of x_t where r is the dimension of β. Then for each $t = 1, ..., T - 1$ we can form semiparametric moment conditions as

$$m_t(z, \beta, \gamma_1) = A_t(x_t)[y_t - P(\tilde{v}(x_t, \beta, \gamma_1))].$$

To derive locally robust moment functions we can derive the adjustment term for estimation of γ_1. The first step function γ_1 is more complicated than previously considered. It depends on two unknown conditional expectations, $E[y_t|x_t]$ and $E[\cdot|x_t, j]$, $(j = 2, ..., J)$. From Newey (1994, p. 1357) we know that the adjustment term will be the sum of two terms, each adjusting for one of the two conditional expectations while treating the other as if it was equal to the truth. In the Appendix we give a general form for each of these adjustment terms. Here we apply that general form to derive the corresponding locally robust moment functions for dynamic discrete choice.

We begin by deriving the adjustment terms for γ_b and γ_c because they are simpler than for γ_a. The adjustment term for γ_b and γ_c are obtained by applying Proposition A1 in the

Appendix. Let $\tilde{\gamma}_b(F), \tilde{\gamma}_c(F)$ have the same form as γ_b and γ_c except that $E[\cdot|x_t, y_t]$ is replaced by $E_F[\cdot|x_t, j]$. For $H_{t+1} = H\left(\gamma_{a0}(x_{t+1})\right)$ and $\pi_1 = E[y_{t1}]$ let

$$\begin{aligned}
\lambda_{bj}(z, \beta, \gamma_1) &= [y_{t,j+1}/\gamma_{a,j+1}(x_t)]\{H(\gamma_a(x_{t+1})) - \gamma_{b,j+1}(x_t, \gamma_a)\}, \\
\lambda_b(z, \beta, \gamma_1) &= (\lambda_{b1}(z, \beta, \gamma_1), ..., \lambda_{b,J-1}(z, \beta, \gamma_1))^T, \\
\lambda_c(z, \beta, \gamma_1, \pi_1) &= \pi_1^{-1} y_{t1} H(\gamma_a(x_{t+1})) - \gamma_c(\gamma_a).
\end{aligned}$$

Then for e a $(J-1) \times 1$ vector of $1's$ we have

$$\begin{aligned}
\frac{\partial E[m_t(z_i, \beta_0, \gamma_{a0}, \tilde{\gamma}_b(F_\tau), \tilde{\gamma}_c(F_\tau))]}{\partial \tau} &= E[\phi_b^t(z_i, \beta_0, \gamma_1, \gamma_2) S(z_i)], \\
\phi_b^t(z, \beta, \gamma_1, \pi) &= -\delta A(x_t) \frac{\partial P}{\partial \tilde{v}}(\tilde{v}(x_t, \beta, \gamma_1)) \lambda_b(z, \beta, \gamma_1) \\
&\quad + \delta E[A(x_t) \frac{\partial P}{\partial \tilde{v}}(\tilde{v}(x_t, \beta, \gamma_1))] e \cdot \lambda_c(z, \beta, \gamma_1, \pi_1)\}
\end{aligned}$$

The adjustment term for estimation of $\gamma_a(x)$ is obtained by applying Proposition A2. This term is somewhat complicated because $\gamma_a(x)$ is evaluated at $x = x_{t+1}$ rather than the conditioning argument x_t of its true value We assume that x_t is stationary over time so that x_{t+1} and x_t have the same pdf, eliminating the ratio of pdf's in the conclusion of Proposition A2. Let $\bar{\gamma}_a(F)$ have the same form as γ_{10} except that $\gamma_{10}(x)$ is replaced by $E_{F_t}[y_t|x_t = x]$. Also let

$$\begin{aligned}
\gamma_{2j0}(x, \beta, \gamma_1) &= E\left[\frac{y_{tj}}{\gamma_{aj}(x_t)} A(x_t) \frac{\partial P(\tilde{v}(x_t, \beta, x_t))}{\partial \tilde{v}} | x_{t+1} = x\right], (j = 1, ..., J-1), \\
\gamma_{30}(x, \pi_1) &= \pi_1^{-1} E[y_{t1}|x_{t+1} = x]|_{x=x_t}.
\end{aligned}$$

Then we have

$$\begin{aligned}
\frac{\partial E[m_t(z_i, \beta_0, \gamma_a(F_\tau), \gamma_{b0}, \gamma_{c0})]}{\partial \tau} &= E[\phi_a^t(z_i, \beta_0, \gamma_1, \gamma_2) S(z_i)], \\
\phi_a^t(z, \beta, \gamma_1, \gamma_2, \gamma_3, \pi_1) &= -\delta\{\gamma_2(x, \beta, \gamma_1) - E[A(x_t) P_{\tilde{v}}(\tilde{v}(x_t, \beta, \gamma_1))] \gamma_3(x_t, \pi_1)\} e \\
&\quad \times \frac{\partial H(\gamma_a(x_t))}{\partial \gamma_a} [y_t - \gamma_a(x_t)].
\end{aligned}$$

We can now form locally robust moment conditions as

$$g_t(z, \beta, \gamma_1, \gamma_2, \gamma_3, \pi_1) = m_t(z, \beta, \gamma_1) + \phi_a^t(z, \beta, \gamma_1, \gamma_2, \gamma_3, \pi_1) + \phi_b^t(z, \beta, \gamma_1, \pi_1).$$

With data z_i that is i.i.d. over individuals these moment functions can be used for any t to estimate the structural parameters β. Also, for data for a single individual we could use a time average $\sum_{t=1}^{T-1} g_t(z, \beta, \gamma_1, \gamma_2, \gamma_3, \pi_1)/(T-1)$ to estimate β, although the asymptotic theory we give does not apply to this estimator.

Bajari, Chernozhukov, Hong, and Nekipelov (2009) derived the adjustment term for the more complicated dynamic discrete game of imperfect information. Locally robust moment conditions for such games could be formed using their results. We leave that formulation to future work.

9 Asymptotic Theory for Locally Robust Moments

In this Section we give asymptotic theory for locally robust estimators. In keeping with the general applicability of locally robust moments to a variety of first steps we consider estimation and conditions that have the most general conditions we can find for the first step. In particular, the construction here only requires that the first step converge at rate slightly faster than $n^{-1/4}$ in norms specified below, a more generally applicable condition than in most of the literature. This formulation allows the results to be applied in settings where it is challenging to say much about the first step other than its convergence rate, such as when machine learning is used in the first step. The locally robust form of the moment conditions is essential for this formulation, as previously discussed.

We use cross fitting in the first step to obtain an estimator that is root-n consistent and asymptotically normal with under such generally applicable conditions. Chernozhukov et. al. (2016) gives results with cross fitting that allow for moment functions that are not smooth in parameters. Here we focus on the smooth in parameters case. Cross fitting has been previously used in the literature on semiparametric estimation. See Bickel, Klaasen, Ritov, and Wellner (1993) for discussion. This approach is different than that some previous work in semiparametric estimation, as in Andrews (1994), Newey (1994), Chen, Linton, and van Keilegom (2003), Ichimura and Lee (2010), where cross fitting was not used and the moment conditions need not be locally robust. The approach adopted here leads to general and simple conditions.

The estimator is formed by grouping observations into L distinct groups. Let $\mathcal{I}_\ell, (\ell = 1, ..., L)$ partition the set of observation indices $\{1, ..., n\}$. Let $\hat{\gamma}_{-\ell}$ be the first step constructed from all observations not in $\mathcal{I}_\ell$. Consider sample moment conditions of the form

$$\tilde{g}(\beta) = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in \mathcal{I}_\ell} g(z_i, \beta, \hat{\gamma}_{-\ell}).$$

We consider GMM estimators based on these moment functions. This is a special case of the cross fitting described earlier. Also, leave one out moment conditions are a further special case where each $\mathcal{I}_\ell$ consists of a single observation. We focus here on the case where the number of groups L is fixed to keep the conditions as simple as possible.

An important intermediate result is that the adjustment term for the first step is zero by virtue of $g(z, \beta, \gamma)$ being locally robust, that is

$$\sqrt{n}\tilde{g}(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(z_i, \beta_0, \gamma_0) + o_p(1). \tag{9.1}$$

With cross fitting this result holds under relatively weak and simple conditions:

ASSUMPTION 1: *For each $\ell = 1, ..., L$, i) $\int \|g(z, \beta_0, \hat{\gamma}_{-\ell}) - g(z, \beta_0, \gamma_0)\|^2 F_0(dz) \xrightarrow{p} 0$, ii) for $\zeta > 1, C > 0$ we have $\left\| \int g(z, \beta_0, \hat{\gamma}_{-\ell}) F_0(dz) \right\| \leq C \|\hat{\gamma}_{-\ell} - \gamma_0\|^\zeta$, and iii) $\sqrt{n} \|\hat{\gamma}_{-\ell} - \gamma_0\|^\zeta \xrightarrow{p} 0$.*

LEMMA 17: *If Assumption 1 is satisfied then equation (9.1) is satisfied.*

This Lemma is proved in the Appendix. There are two important components to this result. One component is a stochastic equicontinuity result

$$\sqrt{n}\tilde{g}(\beta_0) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(z_i, \beta_0, \gamma_0) - \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L} n_\ell \int g(z, \beta_0, \hat{\gamma}_{-\ell}) F_0(dz) \xrightarrow{p} 0,$$

where n_ℓ is the number of observations with $i \in \mathcal{I}_\ell$. Assumption 1 i) is sufficient for this result. Assumption 1 i) is a much weaker stochastic equicontinuity condition than appears in much of the literature, e.g. Andrews (1994). Those other conditions generally involve boundedness of some derivatives of $\hat{\gamma}$. In contrast Assumption 1 i) only requires that $\hat{\gamma}_{-\ell}$ have a mean square convergence property. The cross fitting is what makes this condition sufficient. Cattaneo and Jansson (2014) have also previously weakened the stochastic equicontinuity condition and established the validity of the bootstrap for kernel estimators under substantially weaker bandwidth conditions than usually imposed.

The second component of the result is that

$$\sqrt{n}\bar{g}(\hat{\gamma}_{-\ell}) \xrightarrow{p} 0, \bar{g}(\gamma) = \int g(z, \beta_0, \gamma) F_0(dz).$$

This component follows from Assumptions 1 ii) and iii). By comparing Assumption 1 ii) with Definition 2 we see that this condition implies local robustness in the sense that the Frechet derivative of $\bar{g}(\gamma)$ is zero at γ_0. Assumption 1 ii) will generally hold with $\zeta = 2$ if $\bar{g}(\gamma)$ is twice continuously Frechet differentiable. In that case Assumption 1 iii) become the $n^{1/4}$ rate condition familiar from Newey and McFadden (1994) and other works. The more general $1 < \zeta < 2$ case allows for the first Frechet derivative of $\bar{g}(\gamma)$ to satisfy a Lipschitz condition. In this case Assumption 1 iii) will require a convergence rate of $\hat{\gamma}$ that is faster than $n^{1/4}$.

We note that previous results suggest that $n^{1/4}$ convergence of $\hat{\gamma}$ may be stronger than is needed. As shown in Robins et. al. (2008) and Cattaneo and Jansson (2014) the variance terms in $\sqrt{n}\bar{g}(\hat{\gamma})$ are of the same order as the variance term of a nonparametric estimator, rather than being the order of $\sqrt{n}$ times those variance terms. The arguments for these weaker results are quite complicated so we do not attempt to give an account here. Instead we focus on the relatively simple conditions of Assumption 1.

Another component of an asymptotic normality result is convergence of the Jacobian term $\partial\tilde{g}(\beta)/\partial\beta$. The conditions we impose to account for the Jacobian term are standard. Let $\tilde{G}(\beta) = \partial\tilde{g}(\beta)/\partial\beta$ denote the derivative of the moment function.

ASSUMPTION 2: *There is a neighborhood $\mathcal{N}$ of β_0 such that i) $g(z_i, \beta, \gamma)$ is differentiable in β on N with probability approaching 1 ii) there is $\zeta' > 0$ and $d(z_i)$ with $E[d(z_i)] < \infty$ such that*

for $\beta \in N$ *and* $\|\gamma - \gamma_0\|$ *small enough*

$$\left\| \frac{\partial g(z_i, \beta, \gamma)}{\partial \beta} - \frac{\partial g(z_i, \beta_0, \gamma_0)}{\partial \beta} \right\| \leq d(z_i)(\|\beta - \beta_0\|^{\zeta'} + \|\gamma - \gamma_0\|^{\zeta'}),$$

iii) $E[\|\partial g(z_i, \beta_0, \gamma_0)/\partial \beta\|] < \infty$; *iv)* $\|\hat{\gamma}_{-\ell} - \gamma_0\| \xrightarrow{p} 0, (\ell = 1, ..., L)$.

Define

$$G = E[\partial g(z_i, \beta, \gamma_0)/\partial \beta|_{\beta=\beta_0}]$$

LEMMA 18: *If Assumption 2 is satisfied then for any* $\bar{\beta} \xrightarrow{p} \beta_0$, $\tilde{g}(\beta)$ *is differentiable at* $\bar{\beta}$ *with probability approaching one and* $\partial \tilde{g}(\bar{\beta})/\partial \beta \xrightarrow{p} G$.

With Lemmas 17 and 18 in place the asymptotic normality of semiparametric GMM follows in a standard way.

THEOREM 19: *If Assumptions 1 and 2 are satisfied,* $\hat{\beta} \xrightarrow{p} \beta_0$, $\hat{W} \xrightarrow{p} W$, $G'WG$ *is nonsingular, and* $E[\|g(z_i, \beta_0, \gamma_0)\|^2] < \infty$ *then for* $\Omega = E[g(z_i, \beta_0, \gamma_0)g(z_i, \beta_0, \gamma_0)']$,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), V = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}.$$

It is also useful to have a consistent estimator of the asymptotic variance of $\hat{\beta}$. As usual such an estimator can be constructed as

$$\hat{V} = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1},$$
$$\hat{G} = \frac{\partial \tilde{g}(\hat{\beta})}{\partial \beta}, \hat{\Omega} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i \in \mathcal{I}_\ell} g(z_i, \hat{\beta}, \hat{\gamma}_{-\ell})g(z_i, \hat{\beta}, \hat{\gamma}_{-\ell})'.$$

Note that this variance estimator ignores the estimation of γ, which works here because the moment conditions are locally robust. Its consistency will follow under the conditions of Theorem 19 and one additional condition that accounts for the presence of $\hat{\beta}$ in $\hat{\Omega}$.

THEOREM 20: *If the conditions of Theorem 19 are satisfied and there is* $\tilde{d}(z)$ *with* $E[\tilde{d}(z_i)^2] < \infty$ *such that for* $\|\beta - \beta_0\|$ *and* $\|\beta - \beta_0\|$ *small enough*

$$\|g(z_i, \beta, \gamma) - g(z_i, \beta_0, \gamma_0)\| \leq \tilde{d}(z_i)(\|\beta - \beta_0\|^{\zeta'} + \|\gamma - \gamma_0\|^{\zeta'}),$$

then $\hat{V} \xrightarrow{p} V$.

In this Section we have used cross fitting to obtain relatively simple conditions for asymptotic normality of locally robust semiparametric estimators. It is also known that in some settings

some kinds of cross fitting improves the properties of semiparametric estimators. For linear kernel averages it is known that the leave one out method eliminates a bias term and leads to a reduction in asymptotic mean square error; e.g. see NHR and the references therein. Also Robins et. al. (2008) use cross fitting in higher order bias corrections. These results indicate the some kind of cross fitting can lead to estimators with improved properties. For reducing higher order bias and variance it may be desirable to let the number of groups grow with the sample size. That case is beyond the scope of this paper.

10 APPENDIX

We first give an alternative argument for Proposition 2 that is a special case of the proof of Theorem 2.2 of Robins et. al. (2008). As discussed above, $\phi(z, \beta_0, \gamma_0)$ is the influence function of the functional $\mu(F) = E[m(z_i, \beta_0, \gamma(F))]$. Because it is an influence function it has mean zero at all true distributions, i.e. $\int \phi(z, \beta_0, \gamma(F_0))F_0(dz) \equiv 0$ identically in F_0. Since a regular parametric model $\{F_\tau\}$ is just a subset of all true models, we have

$$\int \phi(z, \beta_0, \gamma(F_\tau))F_\tau(dz) \equiv 0,$$

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

$$\left.\frac{\partial E[\phi(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau}\right|_{\tau=0} = -E[\phi(z_i, \beta_0, \gamma_0)S(z_i)]. \tag{10.1}$$

Summing equations (3.3) and (10.1) we obtain

$$\begin{aligned}\left.\frac{\partial E[g(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau}\right|_{\tau=0} &= \left.\frac{\partial E[m(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau}\right|_{\tau=0} + \left.\frac{\partial E[\phi(z_i, \beta_0, \gamma(F_\tau))]}{\partial \tau}\right|_{\tau=0} \\ &= E[\phi(z_i, \beta_0, \gamma_0)S(z_i)] - E[\phi(z_i, \beta_0, \gamma_0)S(z_i)] = 0.\end{aligned}$$

Thus we see that the adjusted moment functions $g(z, \beta, \gamma)$ are locally robust.

Next we derive the form of the adjustment term when a first step is $E[\cdot|x, y = 1]$ for some binary variable y with where $y_i \in \{0, 1\}$. Consider a first step function of the form $\gamma_1(x) = E[w_i|x_i = x, y_i = 1]$. Let $\pi(x_i) = E[y_i|x_i]$. Note that $E[w_i|x_i, y_i = 1] = E[y_i w_i|x_i]/E[y_i|x_i]$ so that

$$\begin{aligned}\frac{\partial E_{F_\tau}[w_i|x_i, y_i = 1]}{\partial \tau} &= E[\pi(x_i)^{-1}\{y_i w_i - E[y_i w_i|x_i]\} - E[w_i|x_i, y_i = 1](y_i - \pi(x_i))\}S(z_i)|x_i] \\ &= E[\pi(x_i)^{-1}y_i\{w_i - E[w_i|x_i, y_i = 1]\}S(z_i)|x_i].\end{aligned}$$

Suppose that there is $\delta(x_i)$ such that

$$\begin{aligned}\frac{\partial E_{F_\tau}[m(z_i, \beta_0, \gamma_1(F_\tau))]}{\partial \tau} &= E[\delta(x_i)\frac{\partial E_{F_\tau}[w_i|x_i, y_i = 1]}{\partial \tau}] \\ &= E[\delta(x_i)\pi(x_i)^{-1}y_i\{w_i - E[w_i|x_i, y_i = 1]\}S(z_i)].\end{aligned}$$

Then taking the limit gives the following result:

PROPOSITION A1: *If there is* $\delta(x)$ *such that* $\partial E[m(z_i, \beta_0, \gamma_1(F_\tau))]/\partial\tau = E[\delta(x_i)\partial E_{F_\tau}[w_i|x_i, y_i = 1]/\partial\tau]$ *then the adjustment term is*

$$\phi(z, \beta, \gamma) = \delta(x)\pi(x)^{-1}y\{w - E[w_i|x_i = x, y_i = 1]\}.$$

Next we derive the adjustment term when a nonparametric regression is evaluated at a variable different than the one being conditioned on in the regression. Note that for $\delta(v)$ with

$$\begin{aligned}\frac{\partial E[\delta(v_i)E_{F_\tau}[y_i|x_i = x]|_{x=w_i}]}{\partial\tau} &= \frac{E[E[\delta(v_i)|w_i]E_{F_\tau}[y_i|x_i = x]|_{x=w_i}]}{\partial\tau} \\ &= \partial\int E[\delta(v_i)|w_i = w]E_{F_\tau}[y_i|x_i = w]f_w(w)dw/\partial\tau \\ &= \partial\int [f_w(x)/f_x(x)]E[\delta(v_i)|w_i = x]E_{F_\tau}[y_i|x_i = x]f_x(dx)/\partial\tau \\ &= E[\{f_x(x_i)^{-1}f_w(x_i)E[\delta(v_i)|w_i = w]|_{w=x_i}(y_i - E[y_i|x_i])\}S(z_i)]\end{aligned}$$

Taking limits gives

PROPOSITION A2: *If there is* $\delta(v)$ *such that* $\partial E[m(z_i, \beta_0, \gamma_1(F_\tau))]/\partial\tau = \partial E[\delta(v_i)E_{F_\tau}[y_i|x_i = x]|_{x=w_i}]/\partial\tau$ *then the adjustment term is*

$$\phi(z, \beta, \gamma) = f_x(x)^{-1}f_w(x)E[\delta(v_i)|w_i = x](y_i - E[y_i|x_i = x]).$$

Next we give the proofs for the asymptotic normality results.

Proof of Lemma 17: Let

$$\bar{g}(\gamma) = \int g(z, \beta_0, \gamma)F_0(dz),\ \hat{\Delta}_{i\ell} = g(z_i, \beta_0, \hat{\gamma}_{-\ell}) - \bar{g}(\hat{\gamma}_{-\ell}) - g(z_i, \beta_0, \gamma_0), (i \in \mathcal{I}_\ell), \bar{\Delta}_\ell = \frac{1}{n}\sum_{i\in\mathcal{I}_\ell}\hat{\Delta}_{i\ell}.$$

Also, let $Z_{-\ell}$ denote a vector of all observations z_i for $i \notin \mathcal{I}_\ell$. Note that by construction $E[\hat{\Delta}_{i\ell}|Z_{-\ell}] = 0$, so for any $i, j \in \mathcal{I}_\ell, i \neq j$, it follows by z_i and z_j independent conditional on $Z_{-\ell}$ that $E[\hat{\Delta}'_{i\ell}\hat{\Delta}_{j\ell}|Z_{-\ell}] = E[\hat{\Delta}'_{i\ell}|Z_{-\ell}]E[\hat{\Delta}_{j\ell}|Z_{-\ell}] = 0$ Furthermore

$$E[\left\|\hat{\Delta}_{i\ell}\right\|^2|Z_{-\ell}] \leq \int \|g(z.\beta_0, \hat{\gamma}_{-\ell}) - g(z, \beta_0, \gamma_0)\|^2 F_0(dz).$$

Therefore, for n_ℓ equal to the number of observations in group ℓ, Assumption 1 i) implies

$$E[\bar{\Delta}'_\ell\bar{\Delta}_\ell|Z_{-\ell}] = \frac{1}{n^2}\sum_{i\in\mathcal{I}_\ell}E[\left\|\hat{\Delta}_{i\ell}\right\|^2|Z_{-\ell}] \leq \frac{n_\ell}{n^2}\int \|g(z.\beta_0, \hat{\gamma}_{-\ell}) - g(z, \beta_0, \gamma_0)\|^2 F_0(dz) = o_p(n_\ell/n^2).$$

Standard arguments then imply that for each ℓ we have $\bar{\Delta}_\ell = o_p(\sqrt{n_\ell}/n)$. It then follows that

$$\sqrt{n}\left[\tilde{g}(\beta_0) - \frac{1}{n}\sum_{i=1}^{n} g(z_i, \beta_0, \gamma_0) - \bar{g}(\hat{\gamma})\right] = \sqrt{n}\sum_{\ell=1}^{L}\bar{\Delta}_\ell = o_p(\sqrt{n_\ell/n}) \xrightarrow{p} 0.$$

It also follows by Assumption 1 ii) and iii) that

$$\sqrt{n}\left\|\bar{g}(\hat{\gamma})\right\| \leq \sqrt{n}C\left\|\hat{\gamma} - \gamma_0\right\|^{\zeta} \xrightarrow{p} 0.$$

The conclusion then follows by the triangle inequality. *Q.E.D.*

Proof of Lemma 18: Let $\tilde{G}(\beta) = \partial\tilde{g}(\beta)/\partial\beta$ when the derivative exists and $\hat{G} = \partial g(z_i, \beta_0, \gamma_0)/\partial\beta$. By the law of large numbers and Assumption 2 iii), $\hat{G} \xrightarrow{p} G$. Also, by Assumption 2 i), ii), iii) $\tilde{G}(\bar{\beta})$ is well defined with probability approaching one $\sum_{i=1}^{n} d(z_i)/n = O_p(1)$ by the Markov inequality, and by the triangle inequality,

$$\begin{aligned}\left\|\tilde{G}(\bar{\beta}) - \hat{G}\right\| &\leq \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in\mathcal{I}_\ell}\{d(z_i)(\left\|\bar{\beta} - \beta_0\right\|^{\zeta'} + \left\|\hat{\gamma}_\ell - \gamma_0\right\|^{\zeta'})\} \\ &\leq \left(\frac{1}{n}\sum_{i=1}^{n} d(z_i)\right)(\left\|\bar{\beta} - \beta_0\right\|^{\zeta'} + \sum_{\ell=1}^{L}\left\|\hat{\gamma}_\ell - \gamma_0\right\|^{\zeta'}) = O_p(1)o_p(1) \xrightarrow{p} 0.\end{aligned}$$

The conclusion then follows by the triangle inequality. Q.E.D.

The proofs of Theorems 19 and 20 are standard and so we omit them.

Acknowledgements

Whitney Newey gratefully acknowledges support by the NSF. Helpful comments were provided by M. Cattaneo, J. Hahn, M. Jansson, Z. Liao, J. Robins, R. Moon, A. de Paula, J.M. Robin, participants in seminars at Cornell, Harvard-MIT, UCL, and USC.

REFERENCES

Ackerberg, D., X. Chen, and J. Hahn (2012): "A Practical Asymptotic Variance Estimator for Two-step Semiparametric Estimators," *The Review of Economics and Statistics* 94: 481–498.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic Efficiency of Semiparametric Two-Step GMM," *The 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.

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

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.

Bang, and J.M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics* 61, 962–972.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica* 80, 2369–2429.

Belloni, A., V. Chernozhukov, and Y. Wei (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," arXiv preprint arXiv:1304.3969.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2016): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, forthcoming..

Bera, A.K., G. Montes-Rojas, and W. Sosa-Escudero (2010): "General Specification Testing with Locally Misspecified Models," *Econometric Theory* 26, 1838–1845.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag, New York.

Bickel, P.J. and Y. Ritov (2003): "Nonparametric Estimators Which Can Be "Plugged-in," *Annals of Statistics* 31, 1033-1053.

Cattaneo, M.D., and M. Jansson (2014): "Bootstrapping Kernel-Based Semiparametric Estimators," working paper, Berkeley.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34, 1987, 305–334.

Chamberlain, G. (1992): "Efficiency Bounds for Semiparametric Regression," *Econometrica* 60, 567–596.

Chen, X. and X. Shen (1997): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica* 66, 289-314.

Chen, X., O.B. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica* 71, 1591-1608.

Chen, X., and A. Santos (2015): "Overidentification in Regular Models," working paper.

Chernozhukov, V., C. Hansen, and M. Spindler (2015): "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics* 7: 649–688.

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

Firpo, S. and C. Rothe (2016): "Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions," working paper.

Hasminskii, R.Z. and I.A. Ibragimov (1978): "On the Nonparametric Estimation of Functionals," *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics*, 41-51.

Hausman, J.A., and W.K. Newey (2016): "Individual Heterogeneity and Average Welfare," *Econometrica* 84, 1225-1248.

Hotz, V.J. and R.A. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies* 60, 497-529.

Ichimura, H., and S. Lee (2010): "Characterization of the Asymptotic Distribution of Semiparametric M-Estimators," *Journal of Econometrics* 159, 252–266.

Ichimura, H. and W.K. Newey (2016): "The Influence Function of Semiparametric Estimators," CEMMAP working paper.

Lee, Lung-fei (2005): "A $C(\alpha)$-type Gradient Test in the GMM Approach," working paper.

Newey, W.K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics* 5, 99-135.

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. (1999): "Consistency of Two-Step Sample Selection Estimators Despite Misspecification of Distribution," *Economics Letters* 63, 129-132.

Newey, W.K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. Engle, and D. McFadden, pp. 2113-2241. North Holland.

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

NEWEY, W.K., AND J.L. POWELL (2003) "Instrumental Variable Estimation of Nonparametric Models," *Econometrica* 71, 1565-1578.

NEWEY, W.K., F. HSIEH, AND J.M. ROBINS (1998): "Undersmoothing and Bias Corrected Functional Estimation," MIT Dept. of Economics working paper 72, 947-962.

NEWEY, W.K., F. HSIEH, AND J.M. ROBINS (2004): "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica* 72, 947-962.

NEYMAN, J. (1959): "Optimal Asymptotic Tests of Composite Statistical Hypotheses," *Probability and Statistics, the Harald Cramer Volume*, ed., U. Grenander, New York, Wiley.

PAKES, A. AND G.S. OLLEY (1995): "A Limit Theorem for a Smooth Class of Semiparametric Estimators," *Journal of Econometrics* 65, 295-332.

POWELL, J.L., J.H. STOCK, AND T.M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

ROBINS, J.M., A. ROTNITZKY, AND L.P. ZHAO (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association* 89: 846–866.

ROBINS, J.M. AND A. ROTNITZKY (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association* 90:122–129.

ROBINS, J.M., A. ROTNITZKY, AND L.P. ZHAO (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association* 90,106–121.

ROBINS, J.M.,AND A. ROTNITZKY (2001): Comment on "Semiparametric Inference: Question and an Answer Likelihood" by P.A. Bickel and J. Kwon, *Statistica Sinica* 11, 863-960.

ROBINS, J.M., A. ROTNITZKY, AND M. VAN DER LAAN (2000): "Comment on 'On Profile Likelihood' by S. A. Murphy and A. W. van der Vaart, *Journal of the American Statistical Association* 95, 431-435.

ROBINS, J., M. SUED, Q. LEI-GOMEZ, AND A. ROTNITZKY (2007): "Comment: Performance of Double-Robust Estimators When Inverse Probability' Weights Are Highly Variable," *Statistical Science* 22, 544–559.

ROBINS, J.M., L. LI, E. TCHETGEN, AND A. VAN DER VAART (2008) "Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals," *IMS Collections Probability and Statistics: Essays in Honor of David A. Freedman, Vol 2,* 335-421.

RUST, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica* 55, 999-1033.

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

SCHARFSTEIN D.O., A. ROTNITZKY, AND J.M. ROBINS (1999): Rejoinder to "Adjusting For Nonignorable Drop-out Using Semiparametric Non-response Models," *Journal of the American Statistical Association* 94, 1135-1146.

SEVERINI, T. AND G. TRIPATHI (2006): "Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors," *Econometric Theory* 22, 258-278.

STOKER, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1482.

TAMER, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *Review of Economic Studies* 70, 147-165.

VAN DER VAART, A.W. (1991): "On Differentiable Functionals," *The Annals of Statistics,* 19, 178-204.

VAN DER VAART, A.W. (1998): *Asymptotic Statistics,* Cambride University Press, Cambridge, England.

WOOLDRIDGE, J.M. (1991): "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," *Journal of Econometrics* 47, 5-46.