Locally Robust Semiparametric Estimation

By

Victor Chernozhukov, Juan Carlos Escancianoy,

Hidehiko Ichimura, Whitney K. Newey,

James R. Robins

April 2018

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION
DISCUSSION PAPER NO. 4

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION (CREPE)
THE UNIVERSITY OF TOKYO
http://www.crepe.e.u-tokyo.ac.jp/

Locally Robust Semiparametric Estimation

Victor Chernozhukov — *MIT*
Juan Carlos Escanciano — *Indiana University*
Hidehiko Ichimura — *University of Tokyo*
Whitney K. Newey — *MIT*
James R. Robins — *Harvard University*

April 2018

Abstract

We give a general construction of debiased/locally robust/orthogonal (LR) moment functions for GMM, where the derivative with respect to first step nonparametric estimation is zero and equivalently first step estimation has no effect on the influence function. This construction consists of adding an estimator of the influence function adjustment term for first step nonparametric estimation to identifying or original moment conditions. We also give numerical methods for estimating LR moment functions that do not require an explicit formula for the adjustment term.

LR moment conditions are important when the first step is machine learning. We derive LR moment conditions for dynamic discrete choice based on first step machine learning estimators of conditional choice probabilities.

We provide simple and general asymptotic theory for LR estimators based on sample splitting. This theory uses the additive decomposition of LR moment conditions into an identifying condition and a first step influence adjustment. Our conditions require only mean square consistency and a few (generally either one or two) readily interpretable rate conditions.

LR moment functions have the advantage of being less sensitive to first step estimation and so less biased. Some LR moment functions are also doubly robust meaning they hold if one first step is incorrect. We give novel classes of doubly robust moment functions and characterize double robustness. For doubly robust estimators our asymptotic theory only requires one rate condition.

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

1 Introduction

There are many economic parameters that depend on nonparametric or large dimensional first steps. Examples include dynamic discrete choice, games, average consumer surplus, and treatment effects. This paper shows how to construct moment functions for GMM estimators that are debiased/locally robust/orthogonal (LR), where moment conditions have a zero derivative with respect to the first step. We show that LR moment functions can be constructed by adding the influence function adjustment for first step estimation to the original moment functions. This construction can also be interpreted as a decomposition of LR moment functions into identifying moment functions and a first step influence function term. We use this decomposition to give simple and general conditions for root-n consistency and asymptotic normality, with different properties being assuming for the identifying and influence function terms. The conditions are easily interpretable mean square consistency and second order remainder conditions based on estimated moments that use cross-fitting (sample splitting). We also give numerical estimators of the influence function adjustment.

LR moment functions have several advantages. Debiased/LR/orthogonal moment conditions are important for root-n consistency when machine learning is used as a first step, as shown by Belloni, Chernozhukov, and Hansen (2014). LR moment functions can be used to construct debiased/double machine learning (DML) estimators by plugging in machine learning first step estimators, as in Chernozhukov et al. (2017, 2018).

We illustrate by deriving LR moment functions for dynamic discrete choice estimation based on conditional choice probabilities. We give a DML estimator for dynamic discrete choice that uses first step machine learning of conditional choice probabilities. We find that it performs well in a Monte Carlo example. Such structural models provide a potentially important application of DML, because of potentially high dimensional state spaces. Adding the first step influence adjustment term provides a general way of constructing LR moment conditions for structural models so that machine learning can be used for first step estimation of conditional choice probabilities, state transition distributions, and other unknown functions on which structural estimators depend.

LR moment conditions also have the advantage of being relatively insensitive to small variation away from the first step true function. This robustness property is appealing in many settings where it may be difficult to get the first step completely correct. Many interesting and useful LR moment functions have an additional property that they are doubly robust (DR), meaning moment conditions hold when one first step is not correct. We give novel classes of DR moment conditions, including for average linear functionals of conditional expectations and probability densities. The construction of adding the first step influence function adjustment to an identifying moment function is useful in obtaining these moment conditions. We also give necessary and sufficient conditions for a large class of moment functions to be DR. We find

DR moments have simpler and more general conditions for asymptotic normality, which helps motivate our consideration of DR moment functions as special cases of LR ones.

The reduced sensitivity to first step estimation of LR moments leads to substantial improvements in finite sample properties in many cases relative to just using the original moment conditions. For dynamic discrete choice we find large bias reductions, moderate variance increases and even reductions in some cases, and coverage probabilities substantially closer to nominal. For machine learning estimators of the partially linear model, Chernozhukov et al. (2017, 2018) found bias reductions so large that the LR estimator is root-n consistent but the estimator based on the original moment condition is not. Substantial improvements were also found for density weighted averages by Newey, Hsieh, and Robins (2004, NHR). The twicing kernel estimators in NHR are numerically equal to LR estimators based on an original kernel, as shown in Newey, Hsieh, Robins (1998), and the twicing kernel estimators were shown to have smaller mean square error in large samples. Also, a Monte Carlo example in NHR finds that the mean square error (MSE) of the LR estimator has a smaller minimum and is flatter as a function of bandwidth than the MSE of Powell, Stock, and Stoker's (1989) density weighted average derivative estimator. We expect similar finite sample improvements from LR moments will be present in other cases.

LR moment conditions have appeared in earlier work. They are semiparametric versions of Neyman (1959) C-alpha test scores for parametric models. Hasminskii and Ibragimov (1978) suggested LR estimation of functionals of a density and argued for their advantages over plug-in estimators. Pfanzagl and Wefelmeyer (1981) considered their use for improving asymptotic efficiency of functionals of distribution estimators. Bickel and Ritov (1988) gave a LR estimator of the integrated squared density that attains root-n consistency under minimal conditions. The Robinson (1988) semiparametric regression and Ichimura (1993) index regression estimators are LR. Newey (1990) showed that LR moment conditions can be obtained as residuals from projections on the tangent set in a semiparametric model. Newey (1994a) showed that derivatives of an objective function where the first step has been "concentrated out" are LR, including the efficient score of a semiparametric model. NHR (1998, 2004) gave estimators of averages that are linear in density derivative functionals with similarly fast remainder rates to Bickel and Ritov (1988). Doubly robust moment functions have been constructed by Robins, Rotnitzky, and Zhao (1994, 1995), Robins and Rotnitzky (1995), Scharfstein, Rotnitzky, and Robins (1999), Robins, Rotnitzky, and van der Laan (2000), Robins and Rotnitzky (2001), Graham (??), and Firpo and Rothe (2017). They are widely used for estimating treatment effects, e.g. Bang and Robins (2005). Van der Laan and Rubin (2006) developed targeted maximum likelihood to obtain a LR estimating equation based on the efficient influence function of a semiparametric model. Robins et. al. (2008, 2017) showed that efficient influence functions are LR, characterized some doubly robust moment conditions, and developed higher order influence functions that

can reduce bias. Belloni, Chernozhukov, and Wei (2013), Belloni, Chernozhukov, and Hansen (2014), Kandasamy et al. (2015), and Belloni, Chernozhukov, Fernandez-Val, and Hansen (2016) derived LR moments in several specific settings.

A main contribution of this paper is the construction of LR moment conditions from any moment condition and first step estimator that can result in a root-n consistent estimator of the parameter of interest. This construction is based on the limit of the first step when a data observation has a general distribution that allows for misspecification, similarly to Newey (1994). LR moment functions are constructed by adding to identifying moment functions the influence function of the true expectation of the identifying moment functions evaluated at the first step limit, i.e. by adding the influence function term that accounts for first step estimation. The addition of the influence adjustment "partials out" the first order effect of the first step on the moments. This construction of LR moments extends those cited above for first step density and distribution estimators to *any first step,* included nonparametric regression or instrumental variable estimators. Also, this construction is *estimator based* rather than model based as in van der Laan and Rubin (2006) and Robins et al. (2008, 2017). The construction depends only on the moment functions and the first step rather than a semiparametric model. Also, we use the fundamental Gateaux derivative definition of the influence function to show LR rather than an embedding in a regular semiparametric model.

The focus on the functional that is the true expected moments evaluated at the first step limit is the key to this construction. This focus should prove useful for constructing LR moments in many setting, including those where it has already been used to find the asymptotic variance of semiparametric estimators, such as Newey (1994a), Pakes and Olley (1995), Hahn (1998), Ai and Chen (2003), Hirano, Imbens, and Ridder (2003), Bajari, Hong, Krainer, and Nekipelov (2010), Bajari, Chernozhukov, Hong, and Nekipelov (2009), Hahn and Ridder (2013, 2016), and Ackerberg, Chen, Hahn, and Liao (2015), Liao and Ridder (??). LR moment functions can be constructed in each of these settings by adding the first step influence function derived for each case as an adjustment to the original, identifying moment functions.

Another contribution is the development of LR moment conditions for dynamic discrete choice. We derive the influence adjustment for first step estimation of conditional choice probabilities as in Hotz and Miller (1993). We find encouraging Monte Carlo results when various machine learning methods are used to construct the first step. We also give LR conditional moment restrictions that are based on orthogonal instruments.

An additional contribution is to provide general estimators of the influence adjustment term that can be used to construct LR moments without knowing their form. These methods estimate the adjustment term numerically, thus avoiding the need to know its form. It is beyond the scope of this paper to develop machine learning versions of these numerical estimators. Such estimators are developed by Chernozhukov, Newey, and Robins (2018) for average linear functionals of

conditional expectations.

Further contributions include novel classes of DR estimators, including linear functionals of nonparametric instrumental variables and density estimators, and a characterization of (necessary and sufficient conditions for) double robustness. We also give related, novel partial robustness results where original moment conditions are satisfied even when the first step is not the truth.

A main contribution is simple and general asymptotic theory for LR estimators that use cross-fitting in the construction of the average moments. This theory is based on the structure of LR moment conditions as an identifying moment condition depending on one first step plus an influence adjustment that can depend on an additional first step. We give a remainder decomposition that leads to mean square consistency conditions for first steps plus a few readily interpretable rate conditions. For DR estimators there is only one rate condition, on a product of sample remainders from two first step estimators, leading to particularly simple conditions. This simplicity motivates our inclusion of results for DR estimators. This asymptotic theory is also useful for existing moment conditions that are already known to be LR. Whenever the moment condition can be decomposed into an identifying moment condition depending on one first step and an influence function term that may depend on two first steps the simple and general regularity conditions developed here will apply.

LR moments reduce smoothing bias resulting from first step nonparametric estimation relative to original moment conditions. There are other sources of bias arising from nonlinearity of moment conditions in the first step and the empirical distribution. Cattaneo and Jansson (2017) and Cattaneo, Jansson, and Ma (2017) give useful bootstrap and jackknife methods for removing nonlinearity bias. Newey and Robins (2017) show that this bias may also be removed by cross fitting in some settings. We allow for cross-fitting in this paper.

Section 2 describes the general construction of LR moment functions for semiparametric GMM. Section 3 gives LR moment conditions for dynamic discrete choice. Section 4 shows how the first step adjustment term can be estimated. Section 5 gives novel classes of DR moment functions and characterizes double robustness. Section 6 gives an orthogonal instrument construction of LR moments based on conditional moment restrictions. Section 7 gives simple and general asymptotic theory for LR estimators.

2 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 F_0 of z_i.

To describe semiparametric GMM 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. We assume 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

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

where A^T denotes the transpose of a matrix A and B is the parameter space for β. Such estimators have been considered by, e.g. Andrews (1994), Newey (1994a), Newey and McFadden (1994), Pakes and Olley (1995), Chen and Liao (2015), and others.

Locally robust (LR) moment functions can be constructed by adding to the identifying or original moment functions $m(z, \beta, \gamma)$ the influence function adjustment for the first step estimator $\hat{\gamma}$. To describe this influence adjustment let $\gamma(F)$ denote the limit of $\hat{\gamma}$ when z_i has distribution F, where we restrict F only in that $\gamma(F)$ exists and possibly other regularity conditions are satisfied. That is, $\gamma(F)$ is the limit of $\hat{\gamma}$ under possible misspecification, similarly to Newey (1994). Let G be some other distribution and $F_\tau = (1 - \tau)F_0 + \tau G$ for $0 \leq \tau \leq 1$, where F_0 denotes the true distribution of z_i. We assume that G is chosen so that $\gamma(F_\tau)$ is well defined for $\tau > 0$ small enough and possibly other regularity conditions are satisfied, similarly to Ichimura and Newey (2015). The influence function adjustment will be the function $\phi(z, \beta, \gamma, \lambda)$ such that for all such G,

$$\frac{d}{d\tau} E[m(z_i, \beta, \gamma(F_\tau))] = \int \phi(z, \beta, \gamma_0, \lambda_0) G(dz), E[\phi(z_i, \beta, \gamma_0, \lambda_0)] = 0, \tag{2.2}$$

where λ is an additional nonparametric or large dimensional unknown object on which $\phi(z, \beta, \gamma, \lambda)$ depends and the derivative is from the right (i.e. for positive values of τ) and at $\tau = 0$. This equation is the well known definition of the influence function $\phi(z, \beta, \gamma_0, \lambda_0)$ of $\mu(F) = E[m(z_i, \beta, \gamma(F))]$ as the Gateaux derivative of $\mu(F)$, e.g. Huber (1981). The restriction of G so

that $\gamma(F_\tau)$ exists allows $\phi(z,\beta,\gamma_0,\lambda_0)$ to be the influence function when $\gamma(F)$ is only well defined for certain types of distributions, such as when $\gamma(F)$ is a conditional expectation or density. The function $\phi(z,\beta,\gamma,\lambda)$ will generally exist when $E[m(z_i,\beta,\gamma(F))]$ has a finite semiparametric variance bound. Also $\phi(z,\beta,\gamma,\lambda)$ will generally be unique because we are not restricting G very much. Also, $\phi(z,\beta,\gamma,\lambda)$ will be the influence adjustment term from Newey (1994a), as discussed in Ichimura and Newey (2017).

LR moment functions can be constructed by adding $\phi(z,\beta,\gamma,\lambda)$ to $m(z,\beta,\gamma)$ to obtain new moment functions

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

Let $\hat{\lambda}$ be a nonparametric or large dimensional estimator having limit $\lambda(F)$ when z_i has distribution F, with $\lambda(F_0) = \lambda_0$. Also let $\hat{\psi}(\beta) = \sum_{i=1}^n \psi(z_i,\beta,\hat{\gamma},\hat{\lambda})/n$. A LR GMM estimator can be obtained as

$$\hat{\beta} = \arg\min_{\beta\in B} \hat{\psi}(\beta)^T \hat{W} \hat{\psi}(\beta). \tag{2.4}$$

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 Ω of $\sqrt{n}\hat{\psi}(\beta_0)$. As we will further discuss, $\psi(z,\beta,\gamma,\lambda)$ being LR will mean that the estimation of γ and λ does not affect Ω, so that $\Omega = E[\psi(z_i,\beta_0,\gamma_0,\lambda_0)\psi(z_i,\beta_0,\gamma_0,\lambda_0)^T]$. An optimal $\hat{W}$ also gives an efficient estimator in the wider sense shown in Ackerberg, Chen, Hahn, and Liao (2014), making $\hat{\beta}$ efficient in a semiparametric model where the only restrictions imposed are equation (2.1).

The LR property we consider is that the derivative of the true expectation of the moment function with respect to the first step is zero, for a Gateaux derivative like that for the influence function in equation (2.2). Define $F_\tau = (1-\tau)F_0 + \tau G$ as before where G is such that both $\gamma(F_\tau)$ and $\lambda(F_\tau)$ are well defined. The LR property is that for all G as specified,

$$\frac{d}{d\tau} E[\psi(z_i,\beta,\gamma(F_\tau),\lambda(F_\tau))] = 0. \tag{2.5}$$

Note that this condition is the same as that of Newey (1994a) for the presence of $\hat{\gamma}$ an $\hat{\lambda}$ to have no effect on the asymptotic distribution, when each F_τ is a regular parametric submodel. Consequently, the asymptotic variance of $\sqrt{n}\hat{\psi}(\beta_0)$ will be Ω as in the last paragraph.

To show LR of the moment functions $\psi(z,\beta,\gamma,\lambda) = m(z,\beta,\gamma) + \phi(z,\beta,\gamma,\lambda)$ from equation (2.3) we use the fact that the second, zero expectation condition in equation (2.2) must hold for all possible true distributions. For any given β define $\mu(F) = E[m(z_i,\beta,\gamma(F))]$ and $\phi(z,F) = \phi(z,\beta,\gamma(F),\lambda(F))$.

THEOREM 1: *If i)* $d\mu(F_\tau)/d\tau = \int \phi(z,F_0)G(dz)$, *ii)* $\int \phi(z,F_\tau)F_\tau(dz) = 0$ *for all* $\tau \in [0,\bar{\tau})$, *and iii)* $\int \phi(z,F_\tau)F_0(dz)$ *and* $\int \phi(z,F_\tau)G(dz)$ *are continuous at* $\tau = 0$ *then*

$$\frac{d}{d\tau} E[\phi(z_i,F_\tau)] = -\frac{d\mu(F_\tau)}{d\tau}. \tag{2.6}$$

The proofs of this result and others are given in Appendix B. Assumptions i) and ii) of Theorem 1 require that both parts of equation (2.2) hold with the second, zero mean condition being satisfied when F_τ is the true distribution. Assumption iii) is a regularity condition. The LR property follows from Theorem 1 by adding $d\mu(F_\tau)/d\tau$ to both sides of equation (2.6) and noting that the sum of derivatives is the derivative of the sum. Equation (2.6) shows that the addition of $\phi(z,\beta,\gamma,\lambda)$ "partials out" the effect of the first step γ on the moment by "cancelling" the derivative of the identifying moment $E[m(z_i,\beta,\gamma(F_\tau))]$ with respect to τ. This LR result for $\psi(z,\beta,\gamma,\lambda)$ differs from literature in its Gateaux derivative formulation and in the fact that is not a semiparametric influence function but is the hybrid sum of an identifying moment function $m(z,\beta,\gamma)$ and an influence function adjustment $\phi(z,\beta,\gamma,\lambda)$.

Another zero Gateaux derivative property of LR moment functions is also useful. If the sets Γ and Λ of possible limits $\gamma(F)$ and $\lambda(F)$, respectively, are convex and $\gamma(F)$ and $\lambda(F)$ can vary separately from one another then LR moment functions will have the property that for any $\gamma\in\Gamma$, $\lambda\in\Lambda$, and $\bar\psi(\gamma,\lambda)=E[\psi(z_i,\beta_0,\gamma,\lambda)]$,

$$\frac{\partial}{\partial\tau}\bar\psi((1-\tau)\gamma_0+\tau\gamma,\lambda_0)=0,\frac{\partial}{\partial\tau}\bar\psi(\gamma_0,(1-\tau)\lambda_0+\tau\lambda)=0. \tag{2.7}$$

That is, the expected value of the LR moment function will have a zero partial Gateaux derivative with respect to each of the first steps γ and λ. This property will be useful for several results to follow and a stronger, Frechet derivative version will be useful in the asymptotic theory.

The approach of constructing LR moment functions by adding the influence adjustment differs from the model based approach of using as moment functions an efficient influence function or score for a semiparametric model. The approach here is *estimator based* rather than model based. The influence adjustment $\phi(z,\beta,\gamma,\lambda)$ is determined by the nonparametric estimator $\hat\gamma$ and the moment functions $m(z,\beta,\gamma)$ rather than some underlying semiparametric model. This estimator based approach has proven useful for deriving the influence function of a wide variety of semiparametric estimators, as mentioned in the Introduction. Here this estimator based approach provides a general way to construct LR moment functions. For any moment function $m(z,\beta,\gamma)$ and first step estimator $\hat\gamma$ a corresponding LR estimator can be constructed as in equations (2.3) and (2.4).

The addition of $\phi(z,\beta,\gamma,\lambda)$ does not affect identification of β because $\phi(z,\beta,\gamma_0,\lambda_0)$ has expectation zero for any β and true F_0. Consequently, the LR GMM estimator will have the same asymptotic variance as the original GMM estimator $\tilde\beta$ when $\sqrt{n}(\tilde\beta-\beta_0)$ is asymptotically normal, under appropriate regularity conditions. The addition of $\phi(z,\beta,\gamma,\lambda)$ will change other properties of the estimator. As discussed in Chernozhukov et al. (2017, 2018), it can even remove enough bias so that the LR estimator is root-n consistent and the original estimator is not. We further discuss this phenomena in Section 7.

If F_τ was modified so that τ is a function of a smoothing parameter, e.g. a bandwidth,

and τ gives the magnitude of the smoothing bias of $\gamma(F_\tau)$, then equation (2.5) is a small bias condition, being equivalent to

$$E[\psi(z_i, \beta_0, \gamma(F_\tau), \lambda(F_\tau))] = o(\tau).$$

Here $E[\psi(z_i, \beta_0, \gamma(F_\tau), \lambda(F_\tau))]$ is a bias in the moment condition resulting from smoothing and this bias shrinks faster than τ. In this sense LR GMM estimators have the small bias property considered in NHR. This interpretation is also one sense in which LR GMM is "debiased."

In some cases the original moment functions $m(z, \beta, \gamma)$ are already LR and the influence adjustment will be zero. An important class of moment functions that are LR are those where $m(z, \beta, \gamma)$ is the derivative with respect to β of an objective function where nonparametric parts have been concentrated out. That is, suppose that there is a function $q(z, \beta, \zeta)$ such that $m(z, \beta, \gamma) = \partial q(z, \beta, \zeta(\beta))/\partial\beta$ where $\zeta(\beta) = \arg\max_\zeta E[q(z_i, \beta, \zeta)]$, where γ includes $\zeta(\beta)$ and possibly additional functions. Proposition 2 of Newey (1994a) and Lemma 2.5 of Chernozhukov et al. (2018) then implies that $m(z, \beta, \gamma)$ will be LR. This class of moment functions includes various partially linear regression models where ζ represents a conditional expectation. It also includes the efficient score for a semiparametric model, Newey (1994a, pp. 1358-1359).

Cross fitting, that is also known as sample splitting, has often been used to improve the properties of semiparametric and machine learning estimators; e.g. see Bickel (1982), Schick (1986), and Powell, Stock, and Stoker (1989). Cross fitting is known to remove a source of bias and can be used to construct estimators with remainder terms that converge to zero as fast as known possible, as in NHR and Newey and Robins (2017). Cross fitting is also useful for double machine learning estimators, as outlined in Chernozhukov et. al. (2017, 2018). For these reasons we allow for cross-fitting, where sample moments have the form

$$\hat{\psi}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \psi(z_i, \beta, \hat{\gamma}_i, \hat{\lambda}_i),$$

with $\hat{\gamma}_i$ and $\hat{\lambda}_i$ being formed from observations other than the i^{th}. This kind of cross fitting removes an "own observation" bias term and is useful for showing root-n consistency when $\hat{\gamma}_i$ and $\hat{\lambda}_i$ are machine learning estimators.

One version of cross-fitting with good properties in examples in Chernozhukov et. al. (2018) can be obtained by partitioning the observation indices into L groups $I_\ell, (\ell = 1, ..., L)$, forming $\hat{\gamma}_\ell$ and $\hat{\lambda}_\ell$ from observations not in I_ℓ, and constructing

$$\hat{\psi}(\beta) = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} \psi(z_i, \beta, \hat{\gamma}_\ell, \hat{\lambda}_\ell). \tag{2.8}$$

Further bias reductions may be obtained in some cases by using different sets of observations for computing $\hat{\gamma}_\ell$ and $\hat{\lambda}_\ell$, leading to remainders that converge to zero as fast as known possible

in interesting cases; see Newey and Robins (2017). The asymptotic theory of Section 7 focuses on this kind of cross fitting.

As an example we consider a bound on average equivalent variation. Let $\gamma_0(x)$ denote the conditional expectation of quantity q conditional on $x = (p^T, y)$ where $p = (p_1, p_2^T)^T$ is a vector of prices and y is income. The object of interest is a bound on average equivalent variation for a price change from $\bar{p}_1$ to $\check{p}_1$ given by

$$\beta_0 = E[\int \ell(p_1, y_i)\gamma_0(p_1, p_{2i}, y_i)dp_1], \ell(p_1, y) = w(y)1(\bar{p}_1 \leq p_1 \leq \check{p}_1)\exp\{-B(p_1 - \bar{p}_1)\}],$$

where $w(y)$ is a function of income and B a constant. It follows by Hausman and Newey (2016) that if B is a lower (upper) bound on the income effect for all individuals then β_0 is an upper (lower) bound on the equivalent variation for a price change from $\bar{p}_1$ to $\check{p}_1$, averaged over heterogeneity, other prices p_{2i}, and income y_i. The function $w(y)$ allows for averages over income in specific ranges, as in Hausman and Newey (2017).

A moment function that could be used to estimate β_0 is

$$m(z, \beta, \gamma) = \int \ell(p_1, y)\gamma(p_1, p_2, y)dp_1 - \beta.$$

Note that

$$E[m(z_i, \beta_0, \gamma)] + \beta_0 = E[\int \ell(p_1, y_i)\gamma(p_1, p_{2i}, y_i)dp_1] = E[\lambda_0(x_i)\gamma(x_i)], \lambda_0(x) = \frac{\ell(p_1, y)}{f_0(p_1|p_2, y)},$$

where $f_0(p_1|p_2, y)$ is the conditional pdf of p_{1i} given p_{2i} and y_i. Then by Proposition 4 of Newey (1994) the influence function adjustment for any nonparametric estimator $\hat{\gamma}(x)$ of $E[q_i|x_i = x]$ is

$$\phi(z, \beta, \gamma, \lambda) = \lambda(x)[q - \gamma(x)].$$

Here $\lambda_0(x)$ is an example of an additional unknown function that is included in $\phi(z, \beta, \gamma, \lambda)$ but not in the original moment functions $m(z, \beta, \gamma)$. Let $\hat{\gamma}_i(x)$ be an estimator of $E[q_i|x_i = x]$ that can depend on i and $\hat{\lambda}_i(x)$ be an estimator of $\lambda_0(x)$, such as $\hat{f}_i(p_1|p_2, y)^{-1}\ell(p_1, y)$ for an estimator $\hat{f}_i(p_1|p_2, y)$. The LR estimator obtained by solving $\hat{\psi}(\beta) = 0$ for $m(z, \beta, \gamma)$ and $\phi(z, \beta, \gamma, \lambda)$ as above is

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\int \ell(p_1, y_i)\hat{\gamma}_i(p_1, p_{2i}, y_i)dp_1 + \hat{\lambda}_i(x_i)[q_i - \hat{\gamma}_i(x_i)]\right\}. \tag{2.9}$$

3 Machine Learning for Dynamic Discrete Choice

A challenging problem for estimating dynamic structural models is the dimensionality of state spaces. Machine learning addresses this problem through the use of model selection to estimate high dimensional choice probabilities. These choice probabilities estimators can then be used

in conditional choice probability (CCP) estimators of structural parameters, following Hotz and Miller (1993). In order for CCP estimators based on machine learning to be root-n consistent they must be based on orthogonal (i.e. LR) moment conditions, see DML. Adding the adjustment term provides the way to construct LR moment conditions from known moment conditions for CCP estimators. In this Section we do so for the Rust (1987) model of dynamic discrete choice.

We consider an agent choosing among J discrete alternative by maximizing the expected present discounted value of utility. We assume that the per-period utility function for an agent making choice j in period t is given by

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

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 much of the literature we assume that ϵ_t is i.i.d. over time with known CDF that has support R^J, is independent of and state process x_t, and x_t is first-order Markov.

To describe the agent's choice probabilities 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 $\bar{v}_j(x_t) = u_j(x_t, \beta_0) + \delta E[\bar{v}(x_{t+1})|x_t, j]$ the expected value function conditional on choice j. As in Rust (1987), in each period the agent is assumed to make the choice j that maximizes the expected present discounted value of utility $\bar{v}_j(x_t) + \epsilon_{jt}$. The probability of choosing j in period t is then

$$P_j(\bar{v}_t) = \Pr(\bar{v}_j(x_t) + \epsilon_{jt} \geq \bar{v}_k(x_t) + \epsilon_{kt}; k = 1, ..., J), \bar{v}_t = (\bar{v}_1(x_t), ..., \bar{v}_J(x_t))'. \tag{3.1}$$

These choice probabilities have a useful relationship to the structural parameters β when there is a renewal choice, where the conditional distribution of x_{t+1} given the renewal choice and x_t does not depend on x_t. Without loss of generality suppose that the renewal choice is $j = 1$. Let $\tilde{v}_{jt}$ denote $\tilde{v}_j(x_t) = \bar{v}_j(x_t) - \bar{v}_1(x_t)$, so that $\tilde{v}_{1t} \equiv 0$. As usual, subtracting $\bar{v}_{1t}$ from each $\bar{v}_{jt}$ in $P_j(\bar{v}_t)$ does not change the choice probabilities, so that they depend only on $\tilde{v}_t = (\tilde{v}_{2t}, ..., \tilde{v}_{Jt})$.

The renewal nature of $j = 1$ leads to a specific formula for $\tilde{v}_{jt}$ in terms of the per period utilities $u_{jt} = u_j(x_t, \beta_0)$ and the choice probabilities $P_t = P(\tilde{v}_t) = (P_1(\bar{v}_t), ...P_J(\bar{v}_t))'$. As in Hotz and Miller (1993), there is a function $\mathcal{P}^{-1}(P)$ such that $\tilde{v}_t = \mathcal{P}^{-1}(P_t)$. Let $H(P)$ denote the function such that

$$H(P_t) = E[\max_{1 \leq j \leq J} \{\mathcal{P}^{-1}(P_t)_j + \epsilon_{jt}\}|x_t] = E[\max_{1 \leq j \leq J} \{\tilde{v}_{jt} + \epsilon_{jt}\}|x_t].$$

For example, for multinomial logit $H(P_t) = .5772 - \ln(P_{1t})$. Note that by $j = 1$ being a renewal we have $E[\bar{v}_{t+1}|x_t, 1] = C$ for a constant C, so that

$$\bar{v}(x_t) = \bar{v}_{1t} + H(P_t) = u_{1t} + \delta C + H(P_t).$$

It then follows that

$$\bar{v}_{jt} = u_{jt} + \delta E[\bar{v}(x_{t+1})|x_t, j] = u_{jt} + \delta E[u_{1,t+1} + H(P_{t+1})|x_t, j] + \delta^2 C, (j = 1, ..., J).$$

Subtracting then gives

$$\tilde{v}_{jt} = u_{jt} - u_{1t} + \delta\{E[u_{1,t+1} + H(P_{t+1})|x_t, j] - E[u_{1,t+1} + H(P_{t+1})|1]\}. \tag{3.2}$$

This expression for the choice specific value function $\tilde{v}_{jt}$ depends only on $u_j(x_t, \beta)$, $H(P_{t+1})$, and conditional expectations given the state and choice, and so can be used to form semiparametric moment functions.

To describe those moment functions let $\gamma_1(x)$ denote the vector of possible values of the choice probabilities $E[y_t|x_t = x]$, where $y_t = (y_{1t}, ..., y_{Jt})'$. Also let $\gamma_j(x_t, \beta, \gamma_1), (j = 2, ..., J)$ denote a possible $E[u_1(x_{t+1}, \beta) + H(\gamma_1(x_{t+1}))|x_t, j]$ as a function of β, x_t and γ_1, and $\gamma_{J+1}(\beta, \gamma_1)$ a possible value of $E[u_1(x_t, \beta) + H(\gamma_1(x_{t+1}))|1]$. Then a possible value of $\tilde{v}_{jt}$ is given by

$$\tilde{v}_j(x_t, \beta, \gamma) = u_j(x_t, \beta) - u_1(x_t, \beta) + \delta[\gamma_j(x_t, \beta, \gamma_1) - \gamma_{J+1}(\beta, \gamma_1)], (j = 2, ..., J).$$

These value function differences are semiparametric, depending on the function γ_1 of choice probabilities and the conditional expectations γ_j, $(j = 2, ..., J)$. Let $\tilde{v}(x_t, \beta, \gamma) = (\tilde{v}_2(x_t, \beta, \gamma), ..., \tilde{v}_J(x_t, \beta, \gamma))'$, and $A(x_t)$ denote a matrix of functions of x_t with J columns. Semiparametric moment functions are given by

$$m(z, \beta, \gamma) = A(x)[y - P(\tilde{v}(x, \beta, \gamma))].$$

LR moment functions can be constructed by adding the adjustment term for the presence of the first step γ. This adjustment term is derived in Appendix A. It takes the form

$$\phi(z, \beta, \gamma, \lambda) = \sum_{j=1}^{J+1} \phi_j(z, \beta, \gamma, \lambda),$$

where $\phi_j(z, \beta, \gamma, \lambda)$ is the adjustment term for γ_j holding all other components γ fixed at their true values. To describe it define

$$\begin{aligned} &P_{\tilde{v}j}(\tilde{v}) = \partial P(\tilde{v})/\partial \tilde{v}_j, \ \pi_1 = \Pr(y_{t1} = 1), \ \lambda_{10}(x) = E[y_{1t}|x_{t+1} = x], \\ &\lambda_{j0}(x) = E[A(x_t)P_{\tilde{v}j}(\tilde{v}_t)\frac{y_{tj}}{P_j(\tilde{v}_t)}|x_{t+1} = x], (j = 2, ..., J). \end{aligned} \tag{3.3}$$

Then for $w_t = x_{t+1}$ and $z = (y, x, w)$ let

$$\phi_1(z, \beta, \gamma, \lambda) = -\delta \left(\sum_{j=2}^{J} \{\lambda_j(x) - E[A(x_t)P_{\tilde{v}j}(\tilde{v}_t)]\pi_1^{-1}\lambda_1(x)\} \right) [\partial H(\gamma_1(x))/\partial P]'\{y - \gamma_1(x)\}$$

$$\phi_j(z, \beta, \gamma, \lambda) = -\delta A(x)P_{\tilde{v}j}(\tilde{v}(x, \beta, \gamma))\frac{y_j}{P_j(\tilde{v}(x, \beta, \gamma))}\{u_1(w, \beta) + H(\gamma_1(w)) - \gamma_j(x, \beta, \gamma_1)\}, (j = 2, ..., J),$$

$$\phi_{J+1}(z, \beta, \gamma, \lambda) = \delta \left(\sum_{j=2}^{J} E[A(x_t)P_{\tilde{v}j}(\tilde{v}(x_t, \beta, \gamma))] \right) \pi_1^{-1} y_1\{u_1(w, \beta) + H(\gamma_1(w)) - \gamma_{J+1}(\beta, \gamma_1)\}.$$

THEOREM 2: *If the marginal distribution of x_t does not vary with t then LR moment functions for the dynamic discrete choice model are*

$$\psi(z,\beta,\gamma) = A(x_t)[y_t - P(\tilde{v}(x_t,\beta,\gamma))] + \sum_{j=1}^{J+1} \phi_j(z,\beta,\lambda).$$

The form of $\psi(z,\beta,\gamma)$ is amenable to machine learning. A machine learning estimator of the conditional choice probability vector $\gamma_{10}(x)$ is straightforward to compute and can then be used throughout the construction of the orthogonal moment conditions everywhere γ_1 appears. If $u_1(x,\beta)$ is linear in x, say $u_1(x,\beta) = x_1'\beta_1$ for subvectors x_1 and β_1 of x and β respectively, then machine learning estimators can be used to obtain $\hat{E}[x_{1,t+1}|x_t,j]$ and $\hat{E}[\hat{H}_{t+1}|x_j,j]$, $(j=2,...,J)$, and a sample average used to form $\hat{\gamma}_{J+1}(\beta,\hat{\gamma}_1)$. The value function differences can then be estimated as

$$\tilde{v}_j(x_t,\beta,\hat{\gamma}) = u_j(x_t,\beta) - u_1(x_t,\beta) + \hat{E}[x_{1,t+1}|x_t,j]'\beta_1 - \hat{E}[x_{1,t+1}|1]'\beta_1 + \hat{E}[\hat{H}_{t+1}|x_t,j] - \hat{E}[\hat{H}_{t+1}|1].$$

Furthermore, denominator problems can be avoided by using structural probabilities (rather than the machine learning estimators) in all denominator terms.

The challenging part of the machine learning for this estimator is the dependence on β of the reverse conditional expectations in $\lambda_1(x)$. It may be computationally prohibitive and possibly unstable to redo machine learning for each β. One way to to deal with this complication is to update β periodically, with more frequent updates near convergence. It is important that at convergence the β in the reverse conditional expectations is the same as the β that appears elsewhere FIX.

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 $\sum_{t=1}^{T-1} \psi(z_t,\beta,\gamma)/(T-1)$ to estimate β. It will be just as important to use LR moments for estimation with a single individual as it is with a cross section of individuals, although our asymptotic theory will not apply to that case.

Bajari, Chernozhukov, Hong, and Nekipelov (2009) derived the influence adjustment for 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.

We report a Monte Carlo study. The design of loosely like bus replacement application of Rust (1987). Here x_t has transition density

$$x_{t+1} = \begin{cases} x_t + N(.25,1)^2, y_t = 1, \\ x_t = 1 + N(.25,1)^2, y_t = 0. \end{cases}$$

We assume that the agent chooses y_t contingent on state to maximize

$$\sum_{t=1}^{\infty} \delta^{t-1}[y_t(\alpha\sqrt{x_t} + \varepsilon_t) + (1-y_t)RC], \alpha = -.3, RC = -4.$$

We estimate the conditional choice probability and conditional expectation of tomorrow given today using kernel and series nonparametric regression. The series approximation is taken inside the logit estimator of choice probability. We estimate adjustment term using series regression throughout; kernel estimation of adjustment term did not work well. Here are results for 1000 periods for a single decision maker

LR CCP Estimators, Dynamic Discrete Choice

	Bias		Std Err		95% Cov	
	α	RC	α	RC	α	RC
Two step kernel	-.24	.08	.08	.32	.01	.86
LR kernel	-.05	.02	.06	.32	.95	.92
Two step quad	-.00	.14	.049	.33	.91	.89
LR quad	-.00	.01	.085	.39	.95	.92
Logit Lasso	-.12	.25	.06	.28	.74	.84
LR Logit Lasso	-.09	.01	.08	.36	.93	.95
Random Forest	-.15	-.44	.09	.50	.91	.98
LR Ran. For.	.00	.00	.06	.44	1.0	.98
Boosted Trees	-.10	-.28	.08	.50	.99	.99
LR Boost Tr.	.03	.09	.07	.47	.99	.97

Here we find bias reduction and confidence intervals that for the most part have coverage that is closer to nominal. We also find smaller standard errors for the LR estimator in a number of cases.

4 Estimating the Adjustment Term

Construction of LR moment functions requires an estimator $\hat{\phi}(z, \beta)$ of the adjustment term. The form of $\phi(z, \beta, \gamma, \lambda)$ is known for some cases from the semiparametric estimation literature. Powell, Stock, and Stoker (1989) derived the adjustment term for density weighted average derivatives. Newey (1994a) gave the adjustment term for mean square projections (including conditional expectations), densities, and their derivatives. Hahn (1998) and Hirano, Imbens, and Ridder (2003) used those results to obtain the adjustment term for treatment effect estimators, where the LR estimator will be the doubly robust estimator of Robins, Rotnitzky, and Zhao (1994, 1995). Bajari, Hong, Krainer, and Nekipelov (2010) and Bajari, Chernozhukov, Hong, and Nekipelov (2009) derived adjustment terms in some game models. Hahn and Ridder (2013, 2016) derived adjustments in models with generated regressors including control functions. These prior results can be used to obtain LR estimators by adding the adjustment term with nonparametric estimators plugged in.

For new cases it may be necessary to derive the form of the adjustment term. Also, it is possible to numerically estimate the adjustment term based on series estimators and other nonparametric estimators. In this Section we describe how to construct estimators of the adjustment term in these ways.

4.1 Deriving the Formula for the Adjustment Term

One approach to estimating the adjustment term is to derive a formula for $\phi(z,\beta,\gamma,\lambda)$ and then plug-in $\hat{\gamma}$ and $\hat{\lambda}$ in that formula. A formula for $\phi(z,\beta,\gamma,\lambda)$ can be obtained as in Newey (1994a). Let $\gamma(F)$ be the limit of the nonparametric estimator $\hat{\gamma}$ when z_i has distribution F. Also, let F_τ denote a regular parametric model of distributions with $F_\tau = F_0$ at $\tau = 0$ and score (derivative of the log likelihood at $\tau = 0$) equal to $S(z)$. Then under certain regularity conditions $\phi(z,\beta,\gamma_0,\lambda_0)$ will be the unique solution to

$$\left.\frac{\partial \int m(z,\beta,\gamma(F_\tau))F_0(dz)}{\partial \tau}\right|_{\tau=0} = E[\phi(z_i,\beta,\gamma_0,\lambda_0)S(z_i)], E[\phi(z_i,\beta,\gamma_0,\lambda_0)] = 0, \tag{4.1}$$

as $\{F_\tau\}$ and the corresponding score $S(z)$ are allowed to vary over a family of parametric models where the set of scores for the family has mean square closure that includes all mean zero functions with finite variance. Equation (4.1) is a functional equation that can be solved to find the adjustment term, as was done in many of the papers cited in the previous paragraph.

The influence adjustment can be calculated by taking a limit of the Gateaux derivative as shown in Ichimura and Newey (2017). Let $\gamma(F)$ be the limit of $\hat{\gamma}$ when F is the true distribution of z_i, as before. Let G_z^h be a family of distributions that approaches a point mass at z as $h \longrightarrow 0$. If $\phi(z_i,\beta,\gamma_0,\lambda_0)$ is continuous in z_i with probability one then

$$\phi(z,\beta,\gamma_0,\lambda_0) = \lim_{h\longrightarrow 0}\left(\left.\frac{\partial E[m(z_i,\beta,\gamma(F_\tau^h))]}{\partial \tau}\right|_{\tau=0}\right), F_\tau^h = (1-\tau)F_0 + \tau G_z^h. \tag{4.2}$$

This calculation is more constructive than equation (4.1) in the sense that the adjustment term here is a limit of a derivative rather than the solution to a functional equation. Ichimura and Newey (2017) use this formula to derive the adjustment term when $\hat{\gamma}$ is a nonparametric instrumental variables (NPIV) estimator. In Section 6 we use those results to construct LR estimators when the first step is NPIV.

With a formula for $\phi(z,\beta,\gamma,\lambda)$ in hand from either solving the functional equation in equation (4.1) or calculating the limit of the derivative in equation (4.2), one can estimate the adjustment term by plugging estimators $\hat{\gamma}$ and $\hat{\lambda}$ into $\phi(z,\beta,\gamma,\lambda)$. This approach to estimating LR moments can used to construct LR moments for the average surplus described near the end of Section 2. There the adjustment term depends on the conditional density of p_{1i} given p_{2i} and y_i. Let $\hat{f}_\ell(p_1|p_2,y)$ be some estimator of the conditional pdf of p_{1i} given p_{2i} and y_i. Plugging

that estimator into the formula for $\lambda_0(x)$ gives $\hat{\lambda}_\ell(x) = \frac{\ell(p_1,y)}{\hat{f}_\ell(p_1|p_2,y)}$.This $\hat{\lambda}_\ell(x)$ can then be used in equation (2.9).

4.2 Estimating the Adjustment Term for First Step Series Estimators

Estimating the adjustment term is relatively straightforward when the first step is a series estimator. The adjustment term can be estimated by treating the first step estimator as if it were parametric and applying a standard formula for the adjustment term for parametric two-step estimators. Suppose that $\hat{\gamma}_\ell$ depends on the data through a $K \times 1$ vector $\hat{\zeta}_\ell$ of parameter estimators that has true value ζ_0. Let $m(z, \beta, \zeta)$ denote $m(z, \beta, \gamma)$ as a function of ζ. Suppose that there is a $K \times 1$ vector of functions $h(z, \zeta)$ such that $\hat{\zeta}_\ell$ satisfies

$$\frac{1}{\sqrt{\bar{n}_\ell}} \sum_{i \in \bar{I}_\ell} h(z_i, \hat{\zeta}_\ell) = o_p(1), \tag{4.3}$$

where $\bar{I}_\ell$ is a subset of observations, none which are included in I_ℓ, and $\bar{n}_\ell$ is the number of observations in $\bar{I}_\ell$. Then a standard calculation for parametric two-step estimators (e.g. Newey, 1984, and ??) gives the parametric adjustment term

$$\phi(z_i, \beta, \hat{\zeta}_\ell, \hat{\Psi}_\ell) = \hat{\Psi}_\ell(\beta) h(z_i, \hat{\zeta}_\ell), \hat{\Psi}_\ell(\beta) = -\sum_{j \in \bar{I}_\ell} \frac{\partial m(z_j, \beta, \hat{\zeta}_\ell)}{\partial \zeta} \left(\sum_{j \in \bar{I}_\ell} \frac{\partial h(z_j, \hat{\zeta}_\ell)}{\partial \zeta} \right)^{-1}, i \in I_\ell.$$

In many cases $\phi(z_i, \beta, \hat{\zeta}_\ell, \hat{\Psi}_\ell)$ is known to approximate the true adjustment term $\phi(z, \beta, \gamma_0, \lambda_0)$, as shown by Newey (1994a, 1997) and Ackerberg, Chen, and Hahn (2012) for estimating the asymptotic variance of functions of series estimators. Here this approximation is used for estimation of β instead of just for variance estimation. The estimated LR moment function will be

$$\psi(z_i, \beta, \hat{\zeta}_\ell, \hat{\Psi}_\ell) = m(z_i, \beta, \hat{\zeta}_\ell) + \phi(z_i, \beta, \hat{\zeta}_\ell, \hat{\Psi}_\ell). \tag{4.4}$$

We note that if $\hat{\zeta}_\ell$ were computed from the whole sample then $\hat{\phi}(\beta) = 0$. This degeneracy does not occur when cross-fitting is used, which removes "own observation" bias and is important for first step machine learning estimators, as noted in Section 2.

We can apply this approach to construct LR moment functions for an estimator of the average surplus bound example that is based on series regression. Here the first step estimator of $\gamma_0(x) = E[q_i|x_i = x]$ will be that from an ordinary least regression of q_i on a vector $a(x_i)$ of approximating functions. The corresponding $m(z, \beta, \zeta)$ and $h(z, \zeta)$ are

$$m(z, \beta, \zeta) = A(x)'\zeta - \beta, h(z, \zeta) = a(x)[q - a(x)'\zeta], A(x) = \int \ell(p_1, y) a(p_1, p_2, y) dp_1.$$

Let $\hat{\zeta}_\ell$ denote the least squares coefficients from regressing q_i on $a(x_i)$ for observations that are not included in I_ℓ. Then the estimator of the locally robust moments given in equation (4.4) is

$$\psi(z_i, \beta, \hat{\zeta}_\ell, \hat{\Psi}_\ell) = A(x_i)'\hat{\zeta}_\ell - \beta + \hat{\Psi}_\ell a(x_i)[q_i - a(x_i)'\hat{\zeta}_\ell],$$
$$\hat{\Psi}_\ell = \sum_{j \in \bar{I}_\ell} A(x_j)' \left(\sum_{j \in \bar{I}_\ell} a(x_j)a(x_j)' \right)^{-1}.$$

It can be shown similarly to Newey (1994a, p. 1369) that $\hat{\Psi}_\ell$ estimates the population least squares coefficients from a regression of $\lambda_0(x_i)$ on $a(x_i)$, so that $\hat{\lambda}_\ell(x_i) = \hat{\Psi}_\ell a(x_i)$ estimates $\lambda_0(x_i)$. In comparison the LR estimator described in the previous subsection was based on an explicit nonparametric estimator of $f_0(p_1|p_2, y)$, while this $\hat{\lambda}_\ell(x)$ implicitly estimates the inverse of that pdf via a mean-square approximation of $\lambda_0(x_i)$ by $\hat{\Psi}_\ell a(x_i)$.

Chernozhukov, Newey, and Robins (2018) give a machine learning method for choosing the functions to include in the vector $A(x)$. This method can be combined with machine learning methods for estimating $E[q_i|x_i]$ to construct a double machine learning estimator of average surplus, as shown in Chernozhukov, Hausman, and Newey (2018).

In parametric models moment functions like those in equation (4.4) have been used to "partial out" nuisance parameters ζ. For maximum likelihood they are the basis of Neyman's (1959) C-alpha test. Wooldridge (1991) generalized such moment conditions to nonlinear least squares and Lee (2005), Bera et. al (2010), and Chernozhukov et. al. (2015) to GMM. What is novel here is their use in construction of semiparametric estimators and the interpretation of the estimated LR moment functions $\psi(z_i, \beta, \hat{\zeta}_\ell, \hat{\Psi}_\ell)$ as the sum of an original moment function $m(z_i, \beta, \hat{\zeta}_\ell)$ and an adjustment term $\phi(z_i, \beta, \hat{\zeta}_\ell, \hat{\Psi}_\ell)$.

4.3 Estimating the Adjustment Term with First Step Smoothing

The adjustment term can be estimated in a general way that allows for kernel density, locally linear regression, and other kernel smoothing estimators for the first step. The idea is to differentiate with respect to the effect of the i^{th} observation on sample moments. Newey (1994b) used a special case of this approach to estimate the asymptotic variance of a functional of a kernel based semiparametric or nonparametric estimator. Here we extend this method to a wider class of first step estimators, such as locally linear regression, and apply it to estimating the adjustment term for construction of LR moments.

We will describe this estimator for case where γ is a vector of functions of a vector of variables x. Let $h(z, x, \gamma)$ be a vector of functions of a data observation z, x, and a possible realized value of γ (i.e. a vector of real numbers γ). Also let $\hat{h}_\ell(x, \gamma) = \sum_{j \in \bar{I}_\ell} h(z_j, x, \gamma)/\bar{n}_\ell$ be a sample average over a set of observations $\bar{I}_\ell$ not included in I_ℓ, where $\bar{n}_j$ is the number of observations

in $\bar{I}_j$. We assume that the first step estimator $\hat{\gamma}_\ell(x)$ solves

$$0 = \hat{h}_\ell(x, \gamma).$$

We suppress the dependence of h and $\hat{\gamma}$ on a bandwidth. For example for a pdf $\kappa(u)$ a kernel density estimator would correspond to $h(z_j, x, \gamma) = \kappa(x - x_j) - \gamma$ and a locally linear regression would be $\hat{\gamma}_1(x)$ for

$$h(z_j, x, \gamma) = \kappa(x - x_j) \begin{pmatrix} 1 \\ x - x_j \end{pmatrix} [y_j - \gamma_1 - (x - x_j)'\gamma_2].$$

To measure the effect of the i^{th} observation on $\hat{\gamma}$ let $\hat{\gamma}^\xi_{\ell i}(x)$ be the solution to

$$0 = \hat{h}_\ell(x, \gamma) + \xi \cdot h(z_i, x, \gamma).$$

This $\hat{\gamma}^\xi_{\ell i}(x)$ is the value of the function obtained from adding the contribution $\xi \cdot h(z_i, x, \gamma)$ of the i^{th} observation. An estimator of the adjustment term can be obtained by differentiating the average of the original moment function with respect to ξ at $\xi = 0$. This procedure leads to an estimated locally robust moment function given by

$$\psi(z_i, \beta, \hat{\gamma}_\ell) = m(z_i, \beta, \hat{\gamma}_\ell) + \frac{\partial}{\partial \xi} \frac{1}{\bar{n}_\ell} \sum_{j \in \bar{I}_\ell} m(z_j, \beta, \hat{\gamma}^\xi_{\ell i}(\cdot)) \Bigg|_{\xi=0}.$$

This estimator is a generalization of the influence function estimator for kernels in Newey (1994b).

5 Double and Partial Robustness

The zero derivative condition in equation (2.5) is an appealing robustness property in and of itself. A zero derivative means that the expected moment functions remain closer to zero than τ as τ varies away from zero. This property can be interpreted as local insensitivity 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 (DR) moment conditions. DR moment conditions have expectation zero if one first stage component is incorrect. DR moment conditions allow two chances for the moment conditions to hold, an appealing robustness feature. Also, DR moment conditions have simpler conditions for asymptotic normality than general LR moment functions

as discussed in Section 7. Because many interesting LR moment conditions are also DR we consider double robustness.

LR moments that are constructed by adding the adjustment term for first step estimation provide candidates for DR moment functions. The derivative of the expected moments with respect to each first step will be zero, a necessary condition for DR. The condition for moments constructed in this way to be DR is the following:

ASSUMPTION 1: *There are sets* Γ *and* Λ *such that for all* $\gamma \in \Gamma$ *and* $\lambda \in \Lambda$

$$E[m(z_i, \beta_0, \gamma)] = -E[\phi(z_i, \beta_0, \gamma, \lambda_0)], E[\phi(z_i, \beta_0, \gamma_0, \lambda)] = 0.$$

This condition is just the definition of DR for the moment function $\psi(z, \beta, \gamma) = m(z, \beta, \gamma) + \phi(z, \beta, \gamma, \lambda)$, pertaining to specific sets Γ and Λ.

The construction of adding the adjustment term to an identifying or original moment function leads to several novel classes of DR moment conditions. One such class has a first step that satisfies a conditional moment restriction

$$E[y_i - \gamma_0(w_i)|x_i] = 0, \tag{5.1}$$

where w_i is potentially endogenous and x_i is a vector of instrumental variables. This condition is the nonparametric instrumental variable (NPIV) restriction as in Newey and Powell (1989, 2003) and Newey (1991). A first step conditional expectation where $\gamma_0(x_i) = E[y_i|x_i]$ is included as special case with $w_i = x_i$. Ichimura and Newey (2017) showed that the adjustment term for this step takes the form $\phi(z, \gamma, \lambda) = \lambda(x)[y - \gamma(w)]$ so $m(z, \beta, \gamma) + \lambda(x)[y - \gamma(x)]$ is a candidate for a DR moment function. A sufficient condition for DR is:

ASSUMPTION 2: *i) Equation (5.1) is satisfied; ii)* $\Lambda = \{\lambda(x) : E[\lambda(x_i)^2] < \infty\}$ *and* $\Gamma = \{\gamma(x) : E[\lambda(x_i)^2] < \infty\}$; *iii) there is* $v(w)$ *with* $E[v(w_i)^2] < \infty$ *such that* $E[m(z_i, \beta_0, \gamma)] = E[v(w_i)\{\gamma(w_i) - \gamma_0(w_i)\}]$ *for all* $\gamma \in \Gamma$*; iv) there is* $\lambda_0(x)$ *such that* $v(w_i) = E[\lambda_0(x_i)|w_i]$*; and v)* $E[y_i^2] < \infty$.

By the Riesz representation theorem condition iii) is necessary and sufficient for $E[m(z_i, \beta_0, \gamma)]$ to be a mean square continuous functional of γ with representer $v(w)$. Condition iv) is an additional condition giving continuity in the reduced form difference $E[\gamma(w_i) - \gamma_0(w_i)|x_i]$, as further discussed in Ichimura and Newey (2017). Under this condition

$$\begin{aligned} E[m(z_i, \beta_0, \gamma)] &= E[E[\lambda_0(x_i)|w_i]\{\gamma(w_i) - \gamma_0(w_i)\}] = E[\lambda_0(x_i)\{\gamma(w_i) - \gamma_0(w_i)\}] \\ &= -E[\phi(z_i, \gamma, \lambda_0)], \quad E[\phi(z_i, \gamma_0, \lambda)] = E[\lambda(x_i)\{y_i - \gamma_0(w_i)\}] = 0. \end{aligned}$$

Thus Assumption 2 implies Assumption 1 so that we have

THEOREM 3: *If Assumption 2 is satisfied then* $m(z,\beta,\gamma)+\lambda(x)\{y-\gamma(w)\}$ *is doubly robust.*

There are many interesting, novel examples of DR moment conditions that are special cases Theorem 3. The average surplus bound is an example where $y_i = q_i$, $w_i = x_i$, x_i is the observed vector of prices and income, $\Lambda = \Gamma$ is the set of all measurable functions of x_i with finite second moment, and $\gamma_0(x) = E[y_i|x_i = x]$. Let x_1 denote p_1 and x_2 the vector of other prices and income, so that $x = (x_1, x_2')'$. Also let $f_0(x_1|x_2)$ denote the conditional pdf of p_1 given x_2 and $\ell(x) = \ell(p_1, y)$ for income y. Let $m(z,\beta,\gamma) = \int \ell(p_1,x_2)\gamma(p_1,x_2)dp_1 - \beta$ as before. Multiplying and dividing through by $f_0(p_1|x_2)$ gives, for all $\gamma, \lambda \in \Gamma$ and $\lambda_0(x) = f_0(x_1|x_2)^{-1}\ell(x)$,

$$E[m(z_i,\beta_0,\gamma)] = E[\int \ell(p_1,x_{2i})\gamma(p_1,x_{2i})dp_1]-\beta_0 = E[E[\lambda_0(x_i)\gamma(x_i)|x_{2i}]]-\beta_0 = E[\lambda_0(x_i)\{\gamma(x_i)-\gamma_0(x_i)\}].$$

Theorem 3 then implies that the LR moment function for average surplus $m(z,\beta,\gamma)+\lambda(x)[q-\gamma(x)]$ is DR. A corresponding DR estimator $\hat\beta$ is given in equation (2.9).

The surplus bound is an example of a parameter where $\beta_0 = E[g(z_i,\gamma_0)]$ for some linear functional $g(z,\gamma)$ of γ and for γ_0 satisfying the conditional moment restriction of equation (5.1). For the surplus bound $g(z,\gamma) = \int \ell(p_1,x_2)\gamma(p_1,x_2)dp_1$. If Assumption 2 is satisfied then choosing $m(z,\beta,\gamma) = g(z,\gamma)-\beta$ a DR moment condition is $g(z,\gamma)-\beta+\lambda(x)[y-\gamma(w)]$. A corresponding DR estimator is

$$\hat\beta = \frac{1}{n}\sum_{i=1}^{n}\{g(z_i,\hat\gamma_i) + \hat\lambda_i(x_i)[y_i - \hat\gamma_i(w_i)]\}, \tag{5.2}$$

where $\hat\gamma_i(w)$ and $\hat\lambda_i(x)$ are estimators of $\gamma_0(w)$ and $\lambda_0(x)$ respectively. An estimator $\hat\gamma_i$ can be constructed by nonparametric regression when $w_i = x_i$ or NPIV in general. A series estimator $\hat\lambda_i(x)$ can be constructed similarly to the surplus bound example in Section 3.2. For $w_i = x_i$ Newey and Robins (2017) give such series estimators of $\hat\lambda_i(x)$ and Chernozhukov, Newey, and Robins (2018) show how to choose the approximating functions for $\hat\lambda_i(x_i)$ by machine learning . Simple and general conditions for root-n consistency and asymptotic normality of $\hat\beta$ that allow for machine learning are given in Section 7.

Novel examples of the DR estimator in equation (5.2) $w_i = x_i$ are given by Newey and Robins (2017) and Chernozhukov, Newey, and Robins (2018). Also Appendix B gives a generalization to $\gamma(w)$ and $\gamma(x)$ that satisfy orthogonality conditions more general than conditional moment restrictions and novel examples of those. A novel example with $w_i \neq x_i$ is a weighted average derivative of $\gamma_0(w)$ satisfying equation (5.1). Here $g(z,\gamma) = \bar v(w)\partial\gamma(w)/\partial w$ for some weight function $\bar v(w)$. Let $f_0(w)$ be the pdf of w_i and $v(w) = -f_0(w)^{-1}\partial[\bar v(w)f_0(w)]/\partial w$, assuming that derivatives exist. Assume that $\bar v(w)\gamma(w)f_0(w)$ is zero on the boundary of the support of w_i.Integration by parts then gives Assumption 2 iii). Assume also that there exists $\lambda_0 \in \Lambda$ with $v(w_i) = E[\lambda_0(x_i)|w_i]$. Then for estimators $\hat\gamma_i$ and $\hat\lambda_i$ a DR estimator of the weighted average

derivative is

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\{\bar{v}(w_i)\frac{\partial\hat{\gamma}_i(w_i)}{\partial w} + \hat{\lambda}_i(x_i)[y_i - \hat{\gamma}_i(w_i)]\},$$

This is a DR version of the weighted average derivative estimator of Ai and Chen (2007). A special case of this example is the DR moment condition for the weighted average derivative in the exogenous case where $w_i = x_i$ given in Firpo and Rothe (2016).

Theorem 3 includes as special cases existing DR moment functions when $w_i = x_i$, including the mean with randomly missing data given by Robins and Rotnitzky (1995), the class of DR estimators in Robins et al. (2008), and the DR estimators of Firpo and Rothe (2016). We illustrate for the mean with missing data. Let $w = x$, $x = (a, u)$ for an observed data indicator $a \in \{0, 1\}$ and covariates u, $m(z, \beta, \gamma) = \gamma(1, u) - \beta$, and $\lambda_0(x) = a/\Pr(a_i = 1|u_i = u)$. Here it is well known that

$$E[m(z_i, \beta_0, \gamma)] = E[\gamma(1, u_i)] - \beta_0 = E[\lambda_0(x_i)\{\gamma(x_i) - \gamma_0(x_i)\}] = -E[\lambda_0(x_i)\{y_i - \gamma(x_i)\}].$$

Then DR of the moment function $\gamma(1, w) - \beta + \lambda(x)[y - \gamma(x)]$ of Robins and Rotnitzky (1995) follows by Proposition 5.

Another novel class of DR moment conditions are those where the first step γ is a pdf of a function x of the data observation z. By Proposition 5 of Newey (1994a), the adjustment term for such a first step is $\phi(z, \beta, \gamma, \lambda) = \lambda(x) - \int \lambda(u)\gamma(u)du$ for some possible λ. A sufficient condition for the DR as in Assumption 1 is:

ASSUMPTION 3: *x_i has pdf $\gamma_0(x)$ and for $\Gamma = \{\gamma : \gamma(x) \geq 0, \int \gamma(x)dx = 1\}$ there is $\lambda_0(x)$ such that for all $\gamma \in \Gamma$,*

$$E[m(z_i, \beta_0, \gamma)] = \int \lambda_0(x)\{\gamma(x) - \gamma_0(x)\}dx.$$

Note that for $\phi(z, \gamma, \lambda) = \lambda(x) - \int \lambda(\tilde{x})\gamma(\tilde{x})d\tilde{x}$ it follows from Assumption 3 that $E[m(z_i, \beta_0, \gamma)] = -E[\phi(z_i, \gamma, \lambda_0)]$ for all $\gamma \in \Gamma$. Also, $E[\phi(z_i, \gamma_0, \lambda)] = E[\lambda(x_i)] - \int \lambda(\tilde{x})\gamma_0(\tilde{x})dx = 0$. Then Assumption 1 is satisfied so we have:

THEOREM 4: *If Assumption 3 is satisfied then $m(z, \beta, \gamma) + \lambda(x) - \int \lambda(\tilde{x})\gamma(\tilde{x})d\tilde{x}$ is DR.*

The integrated squared density $\beta_0 = \int \gamma_0(x)^2 dx$ is an example for $m(z, \beta, \gamma) = \gamma(x) - \beta$, $\lambda_0 = \gamma_0$, and

$$\psi(z, \beta, \gamma, \lambda) = \gamma(x) - \beta + \lambda(x) - \int \lambda(\tilde{x})\gamma(\tilde{x})dx.$$

This DR moment function seems to be novel. Another example is the density weighted average derivative (DWAD) of Powell, Stock, and Stoker (1989), where $m(z, \beta, \gamma) = -2y \cdot \partial\gamma(x)/\partial x - \beta$.

Let $\delta(x_i) = E[y_i|x_i]\gamma_0(x_i)$. Assuming that $\delta(u)\gamma(u)$ is zero on the boundary and differentiable, integration by parts gives

$$E[m(z_i, \beta_0, \gamma)] = -2E[y_i \partial\gamma(x_i)/\partial x] - \beta_0 = \int [\partial\delta(\tilde{x})/\partial x]\{\gamma(\tilde{x}) - \gamma_0(\tilde{x})\}du,$$

so that Assumption 3 is satisfied with $\lambda_0(x) = \partial\delta(x)/\partial x$. Then by Theorem 4

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\{-2\frac{\partial\hat{\gamma}_i(x_i)}{\partial x} + \frac{\partial\hat{\delta}_i(x_i)}{\partial x} - \int \frac{\partial\hat{\delta}_i(\tilde{x})}{\partial x}\hat{\gamma}_i(\tilde{x})d\tilde{x}\}$$

is a DR estimator. It was shown in NHR (1998) that the Powell, Stock, and Stoker (1989) estimator with a twicing kernel is numerically equal to a leave one out version of this estimator for the original (before twicing) kernel. Thus the DR result for $\hat{\beta}$ gives an interpretation of the twicing kernel estimator as a DR estimator.

The expectation of the DR moment functions of both Theorem 3 and 4 are affine in γ and λ holding the other fixed at the truth. This property of DR moment functions is general, as shown by the following characterization of DR moment functions:

THEOREM 5: *If Γ and Λ are linear then $\psi(z, \beta, \gamma, \lambda)$ is DR if and only if*

$$\partial E[\psi(z_i, \beta_0, (1-\tau)\gamma_0 + \tau\gamma, \lambda_0)]|_{\tau=0} = 0, \partial E[\psi(z_i, \beta_0, \gamma_0, (1-\tau)\lambda_0 + \tau\lambda)]|_{\tau=0} = 0,$$

and $E[\psi(z_i, \beta_0, \gamma, \lambda_0)]$ and $E[\psi(z_i, \beta_0, \gamma_0, \lambda)]$ are affine in γ and λ respectively.

The zero derivative condition of this result is a Gateaux derivative, componentwise version of LR. Thus, we can focus a search for DR moment conditions to those that are LR. Also, a DR moment function must have an expectation that is affine in each of γ and λ while the other is held fixed at the truth. It is sufficient for this condition that $\psi(z_i, \beta_0, \gamma, \lambda)$ be affine in each of γ and λ while the other is held fixed. This property can depend on how γ and λ are specified. For example the missing data DR moment function $m(1, u) - \beta + \pi(u)^{-1}a[y - \gamma(x)]$ is not affine in the propensity score $\pi(u) = \Pr(a_i = 1|u_i = u)$ but is in $\lambda(x) = \pi(u)^{-1}a$.

In general Theorem 5 motivates the construction of DR moment functions by adding the adjustment term to obtain a LR moment function that will then be DR if it is affine in γ and λ separately. It is interesting to note that in the NPIV setting of Theorem 3 and the density setting of Theorem 4 that the adjustment term is always affine in γ and λ. It then follows from Theorem 5 that in those settings LR moment conditions are precisely those where $E[m(z_i, \beta_0, \gamma)]$ is affine in γ. Robins and Rotnitzky (2001) gave conditions for existence of DR moment conditions in semiparametric models. Theorem 5 is complementary to those results in giving a complete characterization of DR moments when Γ and Λ are linear.

Assumptions 2 and 3 both specify that $E[m(z_i, \beta_0, \gamma)]$ is continuous in an integrated squared deviation norm. These continuity conditions are linked to finiteness of the semiparametric

variance bound for the functional $E[m(z_i, \beta_0, \gamma)]$, as discussed in Newey and McFadden (1994) for Assumption 2 with $w_i = x_i$ and for Assumption 3. For Assumption 2 with $w_i \neq x_i$ Severini and Tripathi (2012) showed for $m(z, \beta, \gamma) = v(w)\gamma(w) - \beta$ with known $v(w)$ that the existence of $\lambda_0(w)$ with $v(w_i) = E[\lambda_0(x_i)|w_i]$ is necessary for existence of a root-n consistent estimator of β. Thus the conditions of Assumption 2 are also linked to necessary conditions for root-n consistent estimation when $w_i \neq x_i$.

Partial robustness refers to settings where $E[m(z_i, \beta_0, \bar{\gamma})] = 0$ for some $\bar{\gamma} \neq \gamma_0$. The novel DR moment conditions given here lead to novel partial robustness results as we now demonstrate in the conditional moment restriction setting of Assumption 2. When $\lambda_0(x)$ in Assumption 2 is restricted in some way there may exist $\tilde{\gamma} \neq \gamma_0$ with $E[\lambda_0(x_i)\{y_i - \tilde{\gamma}(w_i)\}] = 0$. Then

$$E[m(z_i, \beta_0, \tilde{\gamma})] = -E[\lambda_0(x_i)\{y_i - \tilde{\gamma}(w_i)\}] = 0.$$

Consider the average derivative $\beta_0 = E[\partial\gamma_0(w_i)/\partial w_r]$ where $m(z, \beta, \gamma) = \partial\gamma(w)/\partial w_r - \beta$ for some r. Let $\delta = (E[a(x_i)p(w_i)'])^{-1}E[a(x_i)y_i]$ be the limit of the linear IV estimator with right hand side variables $p(w)$ and the same number of instruments $a(x)$. The following is a partial robustness result giving conditions for the average derivative of the linear IV estimator to equal the true average derivative:

THEOREM 6: If $-\partial \ln f_0(w)/\partial w_r = c'p(w)$ for a constant vector c, $E[p(w_i)p(w_i)']$ is nonsingular, and $E[a(x_i)|w_i = w] = \Pi p(w)$ for a square nonsingular Π then for $\delta = (E[a(x_i)p(w_i)'])^{-1}E[a(x_i)y_i]$,

$$E[\partial\{p(w_i)'\delta\}/\partial w_r] = E[\partial\gamma_0(w_i)/\partial w_r].$$

This result shows that if the density score is a linear combination of the right-hand side variables $p(w)$ used by linear IV, the conditional expectation of the instruments $a(x_i)$ given w_i is a nonsingular linear combination of $p(w)$, and $p(w)$ has a nonsingular second moment matrix then the average derivative of the linear IV estimator is the true average derivative. This is a generalization to NPIV of Stoker's (1986) result that linear regression coefficients equal the average derivatives when the regressors are multivariate Gaussian.

DR moment conditions can be used to identify parameters of interest. Under Assumption 1 β_0 may be identified from

$$E[m(z_i, \beta_0, \bar{\gamma})] = -E[\phi(z_i, \beta_0, \bar{\gamma}, \lambda_0)]$$

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

THEOREM 7: *If Assumption 1 is satisfied, λ_0 is identified, and for some $\bar{\gamma}_1$ the equation $E[\psi(z_i, \beta, \bar{\gamma}_1, \lambda_0)] = 0$ has a unique solution then β_0 is identified as that solution.*

Applying this result to the NPIV setting of Assumption 2 gives an explicit formula for certain functionals of $\gamma_0(w)$ without requiring that the completeness identification condition of Newey and Powell (1989, 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 $\lambda_0(x)$ to $v(w) = E[\lambda_0(x)|w]$ will be identified if such a solution exists. Plugging in $\bar{\gamma}_1 = 0$ in the equation $E[\psi(z_i, \beta_0, \bar{\gamma}_1, \lambda_0)] = 0$ gives

COROLLARY 8: *If $v(w_i)$ is identified and there exists $\lambda_0(x_i)$ such that $v(w_i) = E[\lambda_0(x_i)|w_i]$ then $\beta_0 = E[v(w_i)\gamma_0(w_i)]$ is identified as $\beta_0 = E[\lambda_0(x_i)y_i]$.*

Note that this result holds without the completeness condition. Identification of $\beta_0 = E[v(w_i)\gamma_0(w_i)]$ for known $v(w_i)$ with $v(w_i) = E[\lambda_0(x_i)|w_i]$ follows from Severini and Tripathi (2006). Corollary 8 extends that analysis to the case where $v(w_i)$ is only identified but not necessarily known and links it to DR moment conditions. Santos (2011) gives a related formula for a parameter $\beta_0 = \int \tilde{v}(w)\lambda_0(w)dw$. 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.

6 Conditional Moment Restrictions

Models of conditional moment restrictions that depend on unknown functions are important in econometrics. In such models the nonparametric components may be determined simultaneously with the parametric components. In this setting it is useful to work directly with the instrumental variables to obtain LR moment conditions rather than a first step influence adjustment. For that reason we focus in this Section on constructing LR moments by orthogonalizing the instrumental variables.

Our orthogonal instruments framework is based on based on conditional moment restrictions of the form

$$E[\rho_j(z_i, \beta_0, \gamma_0)|x_{ji}] = 0, (j = 1, ..., J), \tag{6.1}$$

where each $\rho_j(z, \beta, \gamma)$ is a scalar residual and x_j are instruments that may differ across j. This model is considered by Chamberlain (1992) and Ai and Chen (2003, 2007) when x_j is the same for each j and for Ai and Chen (2012) when the set of x_j includes x_{j-1}. We allow the residual vector $\rho(z, \beta, \gamma)$ to depend on the entire function γ and not just its value at some function of the observed data z_i.

In this framework we consider LR moment functions having the form

$$\psi(z, \beta, \gamma, \lambda) = \lambda(x)\rho(z, \beta, \gamma), \tag{6.2}$$

where $\lambda(x) = [\lambda_1(x_1), ..., \lambda_J(x_J)]$ is a matrix of instrumental variables with j^{th} column given by $\lambda_j(x_j)$. We will define orthogonal instruments to be those that make $\psi(z, \beta, \gamma, \lambda)$ locally robust. To define orthogonal instrumental variables we assume that γ is allowed to vary over a linear set Γ as F varies. For each $\Delta \in \Gamma$ let

$$\bar{\rho}_\gamma(x, \Delta) = (\frac{\partial E[\rho_1(z_i, \beta_0, \gamma_0 + \tau\Delta)|x_1]}{\partial \tau}, ..., \frac{\partial E[\rho_J(z_i, \beta_0, \gamma_0 + \tau\Delta)|x_J]}{\partial \tau})'.$$

This $\bar{\rho}_\gamma(x, \Delta)$ is the Gateaux derivative with respect to γ of the conditional expectation of the residuals in the direction Δ. We characterize $\lambda_0(x)$ as orthogonal if

$$E[\lambda_0(x_i)\bar{\rho}_\gamma(x_i, \Delta)] = 0 \text{ for all } \Delta \in \Gamma.$$

We assume that $\bar{\rho}_\gamma(x, \Delta)$ is linear in Δ and consider the Hilbert space of vectors of random vectors $a(x) = (a_1(x_1), ..., a_J(x_J))$ with inner product $\langle a, b\rangle = E[a(x_i)'b(x_i)]$. Let $\bar{\Lambda}_\gamma$ denote the closure of the set $\{\bar{\rho}_\gamma(x, \Delta) : \Delta \in \Gamma\}$ in that Hilbert space. Orthogonal instruments are those where each row of $\lambda_0(x)$ is orthogonal to $\bar{\Lambda}_\gamma$. They can be interpreted as instrumental variables where the effect of estimation of γ has been partialed out. When $\lambda_0(x)$ is orthogonal then $\psi(z, \beta, \gamma, \lambda) = \lambda(x)\rho(z, \beta, \gamma)$ is LR:

THEOREM 9: *If each row of $\lambda_0(x)$ is orthogonal to $\bar{\Lambda}_\gamma$ then the moment functions in equation (6.2) are LR.*

We also have a DR result:

THEOREM 10: *If each row of $\lambda_0(x)$ is orthogonal to $\bar{\Lambda}_\gamma$ and $\rho(z, \beta, \gamma)$ is affine in $\gamma \in \Gamma$ then the moment functions in equation (6.2) are DR for $\Lambda = \{\lambda(x) : E[\lambda(x_i)'\rho(z_i, \beta_0, \gamma_0)'\rho(z_i, \beta_0, \gamma_0)\lambda(x_i)]$.*

There are many ways to construct orthogonal instruments. For instance, given a $r \times (J-1)$ matrix of instrumental variables $\lambda(x)$ one could construct corresponding orthogonal ones $\lambda_0(x_i)$ as the matrix where each row of $\lambda(x)$ is replaced by the residual from the least squares projection of the corresponding row of $\lambda(x)$ on $\bar{\Lambda}_\gamma$. For local identification of β we also require that

$$rank(\partial E[\psi(z_i, \beta, \gamma_0)]/\partial\beta|_{\beta=\beta_0}) = \dim(\beta). \tag{6.3}$$

A model where β_0 is identified from semiparametric conditional moment restrictions with common instrumental variables is a special case where x_{ji} is the same for each j. In this case there is a way of constructing orthogonal instruments that leads to an efficient estimator of β_0. Let $\Sigma(x_i)$ 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 $\bar{\Lambda}_\gamma$ is closed in this inner product by $\Sigma(x_i)^{-1}$ bounded. Let $\tilde{\lambda}_k^\Sigma(x_i, \lambda)$ denote the

residual from the least squares projection of the k^{th} row $\lambda(x)'e_k$ of $\lambda(x)$ on $\bar{\Lambda}_\gamma$ with the inner product $\langle a, b\rangle_\Sigma$. Then for all $\Delta \in \Gamma$,

$$E[\tilde{\lambda}_k^\Sigma(x_i,\lambda)'\Sigma(x_i)^{-1}\bar{\rho}_\gamma(x_i,\Delta)] = 0,$$

so that for $\tilde{\lambda}^\Sigma(x_i,\lambda) = [\tilde{\lambda}_1^\Sigma(x_i,\lambda),...,\tilde{\lambda}_r^\Sigma(x_i,\lambda)]$ the instrumental variables $\tilde{\lambda}^\Sigma(x_i,\lambda)\Sigma(x_i)^{-1}$ are orthogonal. Also, $\tilde{\lambda}^\Sigma(x_i,\lambda)$ can be interpreted as the solution to

$$\min_{\{D(x):D(x)'e_k\in\bar{\Lambda}_\gamma,k=1,...,r\}} tr(E[\{\lambda(x_i)-D(x_i)\}\Sigma(x_i)^{-1}\{\lambda(x_i)-D(x_i)\}'])$$

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

$$\lambda_0^*(x) = \tilde{\lambda}^{\Sigma^*}(x,\lambda_\beta)\Sigma^*(x)^{-1}, \lambda_\beta(x_i) = \left.\frac{\partial E[\rho(z_i,\beta,\gamma_0)|x_i]}{\partial\beta}\right|'_{\beta=\beta_0}, \Sigma^*(x_i) = Var(\rho_i|x_i), \rho_i = \rho(z_i,\beta_0,\gamma_0).$$

THEOREM 11: *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

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

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.

This framework includes moment restrictions with a NPIV first step γ satisfying $E[\rho(z_i,\gamma_0)|x_i] = 0$ where we can specify $\rho_1(z,\beta,\gamma) = m(z,\beta,\gamma)$, $x_{1i} = 1$, $\rho_2(z,\beta,\gamma) = \rho(z,\gamma)$, and $x_{2i} = x_i$. It generalizes that setup by allowing for more residuals $\rho_j(z,\beta,\gamma)$, $(j \geq 3)$ and all the residuals to depend on β.

7 Asymptotic Theory

In this Section we give simple and general asymptotic theory for LR estimators that incorporate the cross-fitting of equation (2.8). Throughout we use the structure of LR moment functions that are the sum $\psi(z,\beta,\gamma,\lambda) = m(z,\beta,\gamma)+\phi(z,\beta,\gamma,\lambda)$ of an identifying or original moment function $m(z,\beta,\gamma)$ depending on a first step function γ and an influence adjustment term $\phi(z,\beta,\gamma,\lambda)$ that can depend on an additional first step λ. The asymptotic theory will apply to any moment function that can be decomposed into a function of a single nonparametric estimator and a

function of two nonparametric estimators. This structure and LR leads to particularly simple and general conditions.

The conditions we give are composed of mean square consistency conditions for first steps and one, two, or three rate conditions for quadratic remainders. We will only use one quadratic remainder rate for DR moment conditions, involving faster than $1/\sqrt{n}$ convergence of products of estimation errors for $\hat{\gamma}$ and $\hat{\lambda}$. When $E[m(z_i, \beta_0, \gamma) + \phi(z_i, \beta_0, \gamma, \lambda_0)]$ is not affine in γ we will impose a second rate rate condition that involves faster than than $n^{-1/4}$ convergence of $\hat{\gamma}$. When $E[\phi(z_i, \gamma_0, \lambda)]$ is also not affine in λ we will impose a third rate condition that involves faster than $n^{-1/4}$ convergence of $\hat{\lambda}$. Most adjustment terms $\phi(z, \beta, \gamma, \lambda)$ of which we are aware, including for first step conditional moment restrictions and densities, have $E[\phi(z_i, \beta_0, \gamma_0, \lambda)]$ affine in λ, so that faster $n^{-1/4}$ convergence of $\hat{\lambda}$ will not be required by our conditions. It will suffice for most LR estimators of which we are aware to have faster than $n^{-1/4}$ convergence of $\hat{\gamma}$ and faster than $1/\sqrt{n}$ convergence of the product of estimation errors for $\hat{\gamma}$ and $\hat{\lambda}$, with only the latter condition imposed for DR moment functions. We also impose some additional conditions for convergence of the Jacobian of the moments and sample second moments that give asymptotic normality and consistent asymptotic variance estimation for $\hat{\beta}$.

An important intermediate result for asymptotic normality is

$$\sqrt{n}\hat{\psi}(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i, \beta_0, \gamma_0, \lambda_0) + o_p(1), \tag{7.1}$$

where $\hat{\psi}(\beta)$ is the cross-fit, sample, LR moments of equation (2.8). This result will mean that the presence of the first step estimators has no effect on the limiting distribution of the moments at the true β_0. To formulate conditions for this result we decompose the difference between the left and right-hand side into several remainders. Let $\phi(z, \gamma, \lambda) = \phi(z, \beta_0, \gamma, \lambda)$, $\bar{\phi}(\gamma, \lambda) = E[\phi(z_i, \gamma, \lambda)]$, and $\bar{m}(\gamma) = E[m(z_i, \beta_0, \gamma)]$, so that $\bar{\psi}(\gamma, \lambda) = \bar{m}(\gamma) + \bar{\phi}(\gamma, \lambda)$ Then adding and subtracting terms gives

$$\sqrt{n}[\hat{\psi}(\beta_0) - \sum_{i=1}^{n} \psi(z_i, \beta_0, \gamma_0, \lambda_0)/n] = \hat{R}_1 + \hat{R}_2 + \hat{R}_3 + \hat{R}_4, \tag{7.2}$$

where

$$\hat{R}_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(z_i,\beta_0,\hat{\gamma}_i) - m(z_i,\beta_0,\gamma_0) - \bar{m}(\hat{\gamma}_i)] \tag{7.3}$$
$$+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\phi(z_i,\hat{\gamma}_i,\lambda_0) - \phi(z_i,\gamma_0,\lambda_0) - \bar{\phi}(\hat{\gamma}_i,\lambda_0) + \phi(z_i,\gamma_0,\hat{\lambda}_i) - \phi(z_i,\gamma_0,\lambda_0) - \bar{\phi}(\gamma_0,\hat{\lambda}_i)],$$
$$\hat{R}_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\phi(z_i,\hat{\gamma}_i,\hat{\lambda}_i) - \phi(z_i,\hat{\gamma}_i,\lambda_0) - \phi(z_i,\gamma_0,\hat{\lambda}_i) + \phi(z_i,\gamma_0,\lambda_0)],$$
$$\hat{R}_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\psi}(\hat{\gamma}_i,\lambda_0), \quad \hat{R}_4 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\phi}(\gamma_0,\hat{\lambda}_i),$$

We specify regularity conditions sufficient for each of $\hat{R}_1$, $\hat{R}_2$, $\hat{R}_3$, and $\hat{R}_4$ to converge in probability to zero so that equation (7.1) will hold. The remainder term $\hat{R}_1$ is a stochastic equicontinuity term as in Andrews (1994). We give mean square consistency conditions for $\hat{R}_1 \xrightarrow{p} 0$ in Assumption 3.

The remainder term $\hat{R}_2$ is a second order remainder that involves both $\hat{\gamma}$ and $\hat{\lambda}$. When the influence adjustment is $\phi(z,\gamma,\lambda) = \lambda(x)[y-\gamma(w)]$, as for conditional moment restrictions, then

$$\hat{R}_2 = \frac{-1}{\sqrt{n}}\sum_{i=1}^{n}[\hat{\lambda}_i(x_i) - \lambda_0(x_i)][\hat{\gamma}_i(w_i) - \gamma_0(w_i)].$$

$\hat{R}_2$ will converge to zero when the product of convergence rates for $\hat{\lambda}_i(x_i)$ and $\hat{\gamma}_i(w_i)$ is faster than $1/\sqrt{n}$. However, that is not the weakest possible condition. Weaker conditions for locally linear regression first steps are given by Firpo and Rothe (2015) and for series regression first steps by Newey and Robins (2017). These weaker conditions still require that the product of biases of $\hat{\lambda}_i(x_i)$ and $\hat{\gamma}_i(w_i)$ of converge to zero faster $1/\sqrt{n}$ but have weaker conditions for variance terms. We allow for these weaker conditions by allowing $\hat{R}_2 \xrightarrow{p} 0$ as a regularity condition. Assumption 5 gives these conditions.

We will have $\hat{R}_3 = \hat{R}_4 = 0$ in the DR case of Assumption 1, where $\hat{R}_1 \xrightarrow{p} 0$ and $\hat{R}_2 \xrightarrow{p} 0$ will suffice for equation (7.1). In non DR cases LR leads to $\bar{\psi}(\gamma,\lambda_0) = \bar{m}(\gamma) + \bar{\phi}(\gamma,\lambda_0)$ having a zero functional derivative with respect to γ at γ_0 so that $\hat{R}_3 \xrightarrow{p} 0$ when $\hat{\gamma}_i$ converges to γ_0 at a fast enough, feasible rate. For example if $\bar{\psi}(\gamma,\lambda_0)$ is twice continuously Frechet differentiable in a neighborhood of γ_0 for a norm $\|\cdot\|$, with zero Frechet derivative at γ_0. Then

$$\left|\hat{R}_3\right| \leq C\sum_{\ell=1}^{L}\sqrt{n}\,\|\hat{\gamma}_\ell - \gamma_0\|^2 \xrightarrow{p} 0$$

when $\|\hat{\gamma} - \gamma_0\| = o_p(n^{-1/4})$. Here $\hat{R}_3 \xrightarrow{p} 0$ when each $\hat{\gamma}_\ell$ converges to γ_0 faster than $n^{-1/4}$. It may be possible to weaken this condition by bias correcting $m(z,\beta,\hat{\gamma})$, as by the bootstrap

in Cattaneo and Jansson (2017), the jackknife in Cattaneo Ma and Jansson (2017), and cross-fitting in Newey and Robins (2017). Consideration of such bias corrections for $m(z, \beta, \hat{\gamma})$ is beyond the scope of this paper.

In many cases $\hat{R}_4 = 0$ even though the moment conditions are not DR. For example that is true when $\hat{\gamma}$ is a pdf or when γ_0 estimates the solution to a conditional moment restriction. In such cases mean square consistency, $\hat{R}_2 \xrightarrow{p} 0$, and faster than $n^{-1/4}$ consistency of $\hat{\gamma}$ suffices for equation (7.1); no convergence rate for $\hat{\lambda}$ is needed. The simplification that $\hat{R}_4 = 0$ seems to be the result of λ being a Riesz representer for the linear functional that is the derivative of $\bar{m}(\gamma)$ with respect to γ. Such a Riesz representer will enter $\bar{\phi}(\lambda, \gamma_0)$ linearly, leading to $\hat{R}_4 = 0$. When $\hat{R}_4 \neq 0$ then $\hat{R} \xrightarrow{p} 0$ will follow from twice Frechet differentiability of $\bar{\phi}(\lambda, \gamma_0)$ in λ and faster than $n^{-1/4}$ convergence of $\hat{\lambda}$.

All of the conditions can be easily checked for a wide variety of machine learning and conventional nonparametric estimators. There are well known conditions for mean square consistency for many conventional and machine learning methods. Rates for products of estimation errors are also know for many first step estimators as are conditions for $n^{-1/4}$ consistency. Thus, the simple conditions we give here are general enough to apply to a wide variety of first step estimators.

The first formal assumption of this Section is sufficient for $\hat{R}_1 \xrightarrow{p} 0$.

ASSUMPTION 4: *For each* $\ell = 1, ..., L$, *i) Either* $m(z, \beta_0, \gamma)$ *does not depend on* z *or* $\int \{m(z, \beta_0, \hat{\gamma}_\ell) - m(z, \beta_0, \gamma_0)\}^2 F_0(dz) \xrightarrow{p} 0$, *ii)* $\int \{\phi(z, \hat{\gamma}_\ell, \lambda_0) - \phi(z, \gamma_0, \lambda_0)\}^2 F_0(dz) \xrightarrow{p} 0$, *and* $\int \{\phi(z, \gamma_0, \hat{\lambda}_\ell) - \phi(z, \gamma_0, \lambda_0)\}^2 F_0(dz) \xrightarrow{p} 0$;

The cross-fitting used in the construction of $\hat{\psi}(\beta_0)$ is what makes the mean-square consistency conditions of Assumption 4 sufficient for $\hat{R}_1 \xrightarrow{p} 0$. The next condition is sufficient for $\hat{R}_2 \xrightarrow{p} 0$.

ASSUMPTION 5: *For each* $\ell = 1, ..., L$, *either i)*

$$\sqrt{n} \int \max_j |\phi_j(z, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi_j(z, \gamma_0, \hat{\lambda}_\ell) - \phi_j(z, \hat{\gamma}_\ell, \lambda_0) + \phi_j(z, \gamma_0, \lambda_0)| F_0(dz) \xrightarrow{p} 0$$

or ii) $\hat{R}_2 \xrightarrow{p} 0$.

As previously discussed, this condition allows for just $\hat{R}_2 \xrightarrow{p} 0$ in order to allow the conditions of Firpo and Rothe (2015) and Newey and Robins (2017). The first result of this Section shows that Assumptions 4 and 5 are sufficient for equation (7.1) when the moment functions are DR.

LEMMA 12: *If Assumption 1 is satisfied, with probability approaching one* $\hat{\gamma} \in \Gamma$, $\hat{\lambda} \in \Lambda$, *and Assumptions 4 and 5 are satisfied then equation (7.1) is satisfied.*

An important class of DR estimators are those from equation (5.2). The following result gives conditions for asymptotic linearity of these estimators:

THEOREM 13: *If a) Assumptions 2 and 4 i) are satisfied with* $\hat{\gamma} \in \Gamma$ *and* $\hat{\lambda} \in \Lambda$ *with probability approaching one; b)* $\lambda_0(x_i)$ *and* $E[\{y_i - \gamma_0(w_i)\}^2|x_i]$ *are bounded; c) for each* $\ell = 1, ..., L$, $\int [\hat{\gamma}_\ell(w) - \gamma_0(w)]^2 F_0(dz) \xrightarrow{p} 0$, $\int [\hat{\lambda}_\ell(x) - \lambda_0(x)]^2 F_0(dz) \xrightarrow{p} 0$, *and either*

$$\sqrt{n}\left\{\int [\hat{\gamma}_\ell(w) - \gamma_0(w)]^2 F_0(dw)\right\}^{1/2}\left\{\int [\hat{\lambda}_\ell(x) - \lambda_0(x)]^2 F_0(dx)\right\}^{1/2} \xrightarrow{p} 0$$

or

$$\frac{1}{\sqrt{n}}\sum_{i \in I_\ell}\{\hat{\gamma}_\ell(w_i) - \gamma_0(w_i)\}\{\hat{\lambda}_\ell(x_i) - \lambda_0(x_i)\} \xrightarrow{p} 0;$$

then

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[g(z_i, \gamma_0) - \beta_0 + \lambda_0(x_i)\{y_i - \gamma_0(w_i)\}] + o_p(1).$$

The conditions of this result are simple, general, and allow for machine learning first steps. Conditions i) and ii) just require mean square consistency of the first step estimators $\hat{\gamma}$ and $\hat{\lambda}$. The only convergence rate condition is iii), that requires a product of estimation errors for the two first steps go to zero faster than $1/\sqrt{n}$. This condition allows a trade-off in convergence rates between the two first steps, and can be satisfied even when one of the two rates is not very fast. This trade-off can be important when $\lambda_0(x)$ is not continuous in one of the components of x, as in the surplus bound example. Discontinuity in x can limit that rate at which $\lambda_0(x)$ can be estimated. This result extends the results of Chernozhukov et al. (2018) and Farrell (2015) for DR estimators of treatment effects to the whole novel class of DR estimators from equation (5.2) with machine learning first steps. In interesting related work, Athey et al. (2016) show root-n consistent estimation of an average treatment effect is possible under very weak conditions on the propensity score - allowing for the possibility that the propensity score may not be consistently estimated - under strong sparsity of the regression function such. Thus, for machine learning the conditions here and in Athey et al. (2016) are complementary and one may prefer either depending on whether or not the regression function can be estimated extremely well based on a sparse method. The results here apply to many more DR moment conditions.

DR moment conditions have the special feature that $\hat{R}_3$ and $\hat{R}_4$ in Proposition 4 are equal to zero. For estimators that are not DR we impose that $\hat{R}_3$ and $\hat{R}_4$ converge to zero.

ASSUMPTION 6: *For each* $\ell = 1, ..., L$, *i)* $\sqrt{n}\bar{\psi}(\hat{\gamma}_\ell, \lambda_0) \xrightarrow{p} 0$ *and ii)* $\sqrt{n}\bar{\phi}(\gamma_0, \hat{\lambda}_\ell) \xrightarrow{p} 0$.

Assumption 6 requires that $\hat{\gamma}$ converge to γ_0 fast enough but places no restrictions on the convergence rate of $\hat{\lambda}$ when $\bar{\phi}(\gamma_0, \hat{\lambda}_\ell) = 0$.

LEMMA 14: *If Assumptions 4-6 are satisfied then equation (7.1) is satisfied.*

Assumptions 4-6 are based on the decomposition of LR moment functions into an identifying part and influence function. These conditions are different than other previous work in semiparametric estimation, as in Andrews (1994), Newey (1994), Newey and McFadden (1994), Chen, Linton, and van Keilegom (2003), Ichimura and Lee (2010), Escanciano et al. (2016), and Chernozhukov et al. (2018), that are not based on this decomposition. The conditions extend Chernozhukov et. al. (2018) to many more DR estimators and to estimators that are nonlinear in $\hat{\gamma}$ but only require a convergence rate for $\hat{\gamma}$ and not for $\hat{\lambda}$.

Another component of an asymptotic normality result is convergence of the Jacobian term $\partial\hat{\psi}(\beta)/\partial\beta$ to $M = E[\partial\psi(z_i,\beta,\gamma_0,\lambda_0)/\partial\beta|_{\beta=\beta_0}]$. We impose the following condition for this purpose.

ASSUMPTION 7: *M exists and there is a neighborhood $\mathcal{N}$ of β_0 and $\|\cdot\|$ such that i) for each ℓ, $\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0$, $\left\|\hat{\lambda}_\ell - \lambda_0\right\| \xrightarrow{p} 0$; ii)* for all $\|\gamma-\gamma_0\|$ and $\|\lambda-\lambda_0\|$ small enough $\psi(z_i,\beta,\gamma,\lambda)$ *is differentiable in β on $\mathcal{N}$ with probability approaching 1 iii) 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*

$$\left\|\frac{\partial\psi(z_i,\beta,\gamma,\lambda)}{\partial\beta} - \frac{\partial\psi(z_i,\beta_0,\gamma,\lambda)}{\partial\beta}\right\| \leq d(z_i)\left\|\beta-\beta_0\right\|^{\zeta'};$$

iii) For each $\ell = 1,...,L$, j, and k, $\int \left|\partial\psi_j(z,\beta_0,\hat{\gamma}_\ell,\hat{\lambda}_\ell)/\partial\beta_k - \partial\psi_j(z,\beta_0,\gamma_0,\lambda_0)/\partial\beta_k\right| F_0(dz) \xrightarrow{p} 0$,

The following intermediate result gives Jacobian convergence.

LEMMA 15: *If Assumption 7 is satisfied then for any $\bar{\beta} \xrightarrow{p} \beta_0$, $\hat{\psi}(\beta)$ is differentiable at $\bar{\beta}$ with probability approaching one and $\partial\hat{\psi}(\bar{\beta})/\partial\beta \xrightarrow{p} M$.*

With these results in place in place the asymptotic normality of semiparametric GMM follows in a standard way.

THEOREM 16: *If Assumptions 4-7 are satisfied, $\hat{\beta} \xrightarrow{p} \beta_0$, $\hat{W} \xrightarrow{p} W$, $M'WM$ is nonsingular, and $E[\|\psi(z_i,\beta_0,\gamma_0,\lambda_0)\|^2] < \infty$ then for $\Omega = E[\psi(z_i,\beta_0,\gamma_0,\lambda_0)\psi(z_i,\beta_0,\gamma_0,\lambda_0)']$,*

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

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

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

Note that this variance estimator ignores the estimation of γ and λ which works here because the moment conditions are LR. The following result gives conditions for consistency of $\hat{V}$.

THEOREM 17: *If Assumptions 4 and 7 are satisfied with $E[b(z_i)^2] < \infty$, $M'WM$ is nonsingular, and*

$$\int \left\| \phi(z, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi(z, \gamma_0, \hat{\lambda}_\ell) - \phi(z, \hat{\gamma}_\ell, \lambda_0) + \phi(z, \gamma_0, \lambda_0) \right\|^2 F_0(dz) \xrightarrow{p} 0$$

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

In this Section we have used cross-fitting and a decomposition of moment conditions into identifying and influence adjustment components to formulate simple and general conditions for asymptotic normality of LR GMM estimators. 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.

8 Appendix A: Proofs of Theorems

Proof of Theorem 1: By ii) and iii),

$$0 = (1-\tau) \int \phi(z, F_\tau) F_0(dz) + \tau \int \phi(z, F_\tau) G(dz).$$

Dividing by τ and solving gives

$$\frac{1}{\tau} \int \phi(z, F_\tau) F_0(dz) = - \int \phi(z, F_\tau) G(dz) + \int \phi(z, F_\tau) F_0(z).$$

Taking limits as $\tau \longrightarrow 0$, $\tau > 0$ and using i) gives

$$\frac{d}{d\tau} \int \phi(z, F_\tau) F_0(dz) = - \int \phi(z, F_0) G(dz) + 0 = -\frac{d\mu(F_\tau)}{d\tau}.Q.E.D.$$

Proof of Theorem 2: We begin by deriving ϕ_1, the adjustment term for the first step CCP estimation. We use the definitions given in the body of the paper. We also let

$$P_{\tilde{v}j}(\tilde{v}) = \partial P(\tilde{v})/\partial \tilde{v}_j, \ \pi_1 = \Pr(y_{t1} = 1), \ \lambda_{10}(x) = E[y_{1t} | x_{t+1} = x],$$
$$\lambda_{j0}(x) = E[A(x_t) P_{\tilde{v}j}(\tilde{v}_t) \frac{y_{tj}}{P_j(\tilde{v}_t)} | x_{t+1} = x], (j = 2, ..., J).$$

Consider a parametric submodel as described in Section 4 and let $\gamma_1(x,\tau)$ denote the conditional expectation of y_t given x_t under the parametric submodel. Note that for $\tilde{v}_t = \tilde{v}(x_t)$,

$$
\begin{aligned}
&E[A(x_t)P_{\tilde{v}j}(\tilde{v}_t)\frac{\partial E[H(\gamma_1(x_{t+1},\tau))|x_t, y_{tj}=1]}{\partial \tau}] \\
&= \frac{\partial}{\partial \tau}E[A(x_t)P_{vj}(\tilde{v}_t)\frac{y_{tj}}{P_j(\tilde{v}_t)}H(\gamma_1(x_{t+1},\tau))] \\
&= \frac{\partial}{\partial \tau}E[E[A(x_t)P_{vj}(\tilde{v}_t)\frac{y_{tj}}{P_j(\tilde{v}_t)}|x_{t+1}]H(\gamma_1(x_{t+1},\tau))] \\
&= \frac{\partial}{\partial \tau}E[\lambda_{j0}(x_{t+1})H(\gamma_1(x_{t+1},\tau))] = \frac{\partial}{\partial \tau}E[\lambda_{j0}(x_t)H(\gamma_1(x_t,\tau))] \\
&= E[\lambda_{j0}(x_t)\frac{\partial H(\gamma_{10}(x_t))'}{\partial P}\frac{\partial \gamma_1(x_t,\tau)}{\partial \tau}] = E[\lambda_{j0}(x_t)\frac{\partial H(\gamma_{10}(x_t))'}{\partial P}\{y_t - \gamma_{10}(x_t)\}S(z_t)].
\end{aligned}
$$

where the last (sixth) equality follows as in Proposition 4 of Newey (1994a), and the fourth equality follows by equality of the marginal distributions of x_t and x_{t+1}. Similarly, for $\pi_1 = \Pr(y_{t1}=1)$ and $\lambda_{10}(x) = E[y_{1t}|x_{t+1}=x]$ we have

$$
\begin{aligned}
\frac{\partial E[H(\gamma_1(x_{t+1},\tau))|y_{t1}=1]}{\partial \tau} &= \frac{\partial E[\pi_1^{-1}y_{1t}H(\gamma_1(x_{t+1},\tau))]}{\partial \tau} = \frac{\partial E[\pi_1^{-1}\lambda_{10}(x_{t+1})H(\gamma_1(x_{t+1},\tau))]}{\partial \tau} \\
&= \frac{\partial E[\pi_1^{-1}\lambda_{10}(x_t)H(\gamma_1(x_t,\tau))]}{\partial \tau} \\
&= E[\pi_1^{-1}\lambda_{10}(x_t)\frac{\partial H(\gamma_{10}(x_t))'}{\partial P}\{y_t - \gamma_{10}(x_t)\}S(z_t)]
\end{aligned}
$$

Then combining terms gives

$$
\begin{aligned}
&\frac{\partial E[m(z_t,\beta_0,\gamma_1(\tau),\gamma_{-10})]}{\partial \tau} \\
&= -\delta\sum_{j=2}^{J}\{E[A(x_t)P_{vj}(\tilde{v}_t)\frac{\partial E[H(\gamma_1(x_{t+1},\tau))|x_t,y_{tj}=1]}{\partial \tau}] \\
&- E[A(x_t)P_{vj}(\tilde{v}_t)]\frac{\partial E[H(\gamma_1(x_{t+1},\tau))|y_{t1}=1]}{\partial \tau}\} \\
&= -\delta\sum_{j=2}^{J}E[\{\lambda_{j0}(x_t) - E[A(x_t)P_{\tilde{v}j}(\tilde{v}_t)]\pi_1^{-1}\lambda_{10}(x_t)\}\frac{\partial H(\gamma_{10}(x_t))'}{\partial P}\{y_t - \gamma_{10}(x_t)\}S(z_t)] \\
&= E[\phi_1(z_t,\beta_0,\gamma_0,\lambda_0)S(z_t)].
\end{aligned}
$$

Next, we show the result for $\phi_j(z,\beta,\gamma,\lambda)$ for $2 \le j \le J$. As in the proof of Proposition 4 of Newey (1994a), for any w_t we have

$$
\frac{\partial}{\partial \tau}E[w_t|x_t,y_{tj}=1,\tau] = E[\frac{y_{tj}}{P_j(\tilde{v}_t)}\{w_t - E[w_t|x_t,y_{tj}=1]\}S(z_t)|x_t].
$$

It follows that

$$\begin{aligned}\frac{\partial E[m(z_t,\beta_0,\gamma_j(\tau),\gamma_{-j,0})]}{\partial\tau} &= -\delta E[A(x_t)P_{vj}(\tilde{v}_t)\frac{\partial E[u_{1,t+1}+H_{t+1}|x_t,y_{tj}=1,\tau]}{\partial\tau}] \\ &= -\delta\frac{\partial}{\partial\tau}E[E[A(x_t)P_{vj}(\tilde{v}_t)\{u_{1,t+1}+H_{t+1}\}|x_t,y_{tj}=1,\tau]]. \\ &= -\delta E[A(x_t)P_{vj}(\tilde{v}_t)\frac{y_{tj}}{P_j(\tilde{v}_t)}\{u_{1,t+1}+H_{t+1}-\gamma_{j0}(x_t,\beta_0,\gamma_1)\}S(z_t)] \\ &= E[\phi_j(z_t,\beta_0,\gamma_0,\lambda_0)S(z_t)],\end{aligned}$$

showing that the formula for ϕ_j is correct. The proof for ϕ_{J+1} follows similarly. *Q.E.D.*

Proof of Theorem 3: Given in text.

Proof of Theorem 4: Given in text.

Proof of Theorem 5: Let $\bar{\psi}(\gamma,\lambda) = E[\psi(z_i,\beta_0,\gamma,\lambda)]$. Suppose that $\psi(z,\beta,\gamma,\lambda)$ is DR. Then for any $\gamma\neq\gamma_0,\gamma\in\Gamma$ we have

$$0=\bar{\psi}(\gamma,\lambda_0)=\bar{\psi}(\gamma_0,\lambda_0)=\bar{\psi}((1-\tau)\gamma_0+\tau\gamma,\lambda_0).$$

for any τ. Therefore for any τ,

$$\bar{\psi}((1-\tau)\gamma_0+\tau\gamma,\lambda_0)=0=(1-\tau)\bar{\psi}(\gamma_0,\lambda_0)+\tau\bar{\psi}(\gamma,\lambda_0),$$

so that $\bar{\psi}(\gamma,\lambda_0)$ is affine in γ. Also by the previous equation $\bar{\psi}((1-\tau)\gamma_0+\tau\gamma,\lambda_0)=0$ identically in τ so that

$$\frac{\partial}{\partial\tau}\bar{\psi}((1-\tau)\gamma_0+\tau\gamma,\lambda_0)=0,$$

where the derivative with respect to τ is evaluated at $\tau=0$. Applying the same argument switching of λ and γ we find that $\bar{\psi}(\gamma_0,\lambda)$ is affine in λ and $\partial\bar{\psi}(\gamma_0,(1-\tau)\lambda_0+\tau\lambda)/\partial\tau=0$.

Next suppose that $\bar{\psi}(\gamma,\lambda_0)$ is affine γ and $\partial\bar{\psi}((1-\tau)\gamma_0+\tau\gamma,\lambda_0)/\partial\tau=0$. Then by $\bar{\psi}(\gamma_0,\lambda_0)=0$, for any $\gamma\in\Gamma$,

$$\begin{aligned}\bar{\psi}(\gamma,\lambda_0) &= \partial[\tau\bar{\psi}(\gamma,\lambda_0)]/\partial\tau=\partial[(1-\tau)\bar{\psi}(\gamma_0,\lambda_0)+\tau\bar{\psi}(\gamma,\lambda_0)]/\partial\tau \\ &= \partial\bar{\psi}((1-\tau)\gamma_0+\tau\gamma,\lambda_0)/\partial\tau=0.\end{aligned}$$

Switching the roles of γ and λ it follows analogously that $\bar{\psi}(\gamma_0,\lambda)=0$ for all $\lambda\in\Lambda$, so $\bar{\psi}(\gamma,\lambda)$ is doubly robust. *Q.E.D.*

Proof of Theorem 6: Let $\lambda_0(x)=-c'\Pi^{-1}a(x)$ so that $E[\lambda_0(x_i)|w_i]=-c'\Pi^{-1}\Pi p(w_i)=-c'p(w_i)$.Then integration by parts gives

$$\begin{aligned}E[m(z_i,\beta_0,\tilde{\gamma})] &= E[c'p(w_i)\{\tilde{\gamma}(w_i)-\gamma_0(w_i)\}]=-E[\gamma_0(x_i)\{\tilde{\gamma}(w_i)-\gamma_0(w_i)\}] \\ &= E[\gamma_0(x_i)\{y_i-\tilde{\gamma}(w_i)\}]=-c'\Pi^{-1}E[a(x_i)\{y_i-\tilde{\gamma}(w_i)\}]=0.Q.E.D.\end{aligned}$$

Proof of Theorem 7: If γ_{20} is identified then $m(z, \beta, \bar{\gamma}_1, \gamma_{20})$ is identified for every β. By DRρ

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

at $\beta = \beta_0$ and by assumption this is the only β where this equation equation is satisfied. Q.E.D.

Proof of Corollary 8: Given in text.

Proof of Theorem 9: Note that for $\rho_i = \rho(z_i, \beta_0, \gamma_0)$,

$$\bar{\psi}(\gamma_0, (1-\tau)\lambda_0 + \tau\lambda)] = (1-\tau)E[\lambda_0(x_i)\rho_i] + \tau E[\lambda(x_i)\rho_i] = 0. \tag{8.1}$$

Differentiating gives the second equality in eq. (2.7). Also, for $\Delta = \gamma - \gamma_0$,

$$\frac{\partial \bar{\psi}((1-\tau)\gamma_0 + \tau\gamma, \lambda_0)}{\partial \tau} = E[\lambda_0(x_i)\bar{\rho}(x_i, \Delta)] = 0,$$

giving the first equality in eq. (2.7). *Q.E.D.*

Proof of Theorem 10: The first equality in eq. (8.1) of the proof of Theorem 9 shows that $\bar{\psi}(\gamma_0, \lambda)$ is affine in λ. Also,

$$\bar{\psi}((1-\tau)\gamma_0+\tau\gamma, \lambda_0) = E[\lambda_0(x_i)\{(1-\tau)\rho(z_i, \beta_0, \gamma_0)+\tau\rho(z_i, \beta_0, \gamma)\}] = (1-\tau)\bar{\psi}(\gamma_0, \lambda_0)+\tau\bar{\psi}(\gamma, \lambda_0),$$

so that $\bar{\psi}(\gamma, \lambda_0)$ is affine in γ. The conclusion then follows by Theorem 5. *Q.E.D.*

Proof of Theorem 11: To see that $\tilde{\lambda}^{\Sigma^*}(x_i, \lambda^*)\Sigma^*(x_i)^{-1}$ minimizes the asymptotic variance note that for any orthogonal instrumental variable matrix $\lambda_0(x)$, by the rows of $\lambda_\beta(x_i) - \tilde{\lambda}^{\Sigma^*}(x_i, \lambda_\beta)$ being in $\bar{\Lambda}_\gamma$,

$$M = E[\lambda_0(x_i)\lambda_\beta(x_i)'] = E[\lambda_0(x_i)\tilde{\lambda}^{\Sigma^*}(x_i, \lambda_\beta)'] = E[\lambda_0(x_i)\rho_i\rho_i'\Sigma^*(x_i)^{-1}\tilde{\lambda}^{\Sigma^*}(x_i, \lambda_\beta)'].$$

Since the instruments are orthogonal the asymptotic variance matrix of GMM estimator with $\hat{W} \xrightarrow{p} W$ is the same as if $\hat{\gamma} = \gamma_0$. Define $m_i = M'W\lambda_0(x_i)\rho_i$ and $m_i^* = \tilde{\lambda}^{\Sigma^*}(x_i, \lambda_\beta)\Sigma^*(x_i)^{-1}\rho_i$. The asymptotic variance of the GMM estimator for orthogonal instruments $\lambda_0(x)$ is

$$(M'WM)^{-1}M'WE[\lambda_0(x_i)\rho_i\rho_i'\lambda_0(x_i)']WM(M'WM)^{-1} = (E[m_i m_i^{*\prime}])^{-1}E[m_i m_i'](E[m_i m_i^*])^{-1\prime}.$$

The fact that this matrix is minimized in the positive semidefinite sense for $m_i = m_i^*$ is well known, e.g. see Newey and McFadden (1994). *Q.E.D.*

The following result is useful for the results of Section 7:

LEMMA A1: *If Assumption 4 is satisfied then* $\hat{R}_1 \xrightarrow{p} 0$. *If Assumption 5 is satisfied then* $\hat{R}_2 \xrightarrow{p} 0$.

Proof: Define $\hat{\Delta}_{i\ell} = m(z_i, \hat{\gamma}_\ell) - m(z_i, \gamma_0) - \bar{m}(\hat{\gamma}_\ell)$ for $i \in I_\ell$ and let Z_ℓ^c denote the observations z_i for $i \notin I_\ell$. Note that $\hat{\gamma}_\ell$ depends only on Z_ℓ^c. By construction and independence of Z_ℓ^c and $z_i, i \in I_\ell$ we have $E[\hat{\Delta}_{i\ell}|Z_\ell^c] = 0$. Also by independence of the observations, $E[\hat{\Delta}_{i\ell}\hat{\Delta}_{j\ell}|Z_\ell^c] = 0$ for $i, j \in I_\ell$. Furthermore, for $i \in I_\ell$ $E[\hat{\Delta}_{i\ell}^2|Z_\ell^c] \leq \int [m(z, \hat{\gamma}_\ell) - m(z, \gamma_0)]^2 F_0(dz)$. Then we have

$$E[\left(\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}\hat{\Delta}_{i\ell}\right)^2 |Z_\ell^c] = \frac{1}{n}E[\left(\sum_{i\in I_\ell}\hat{\Delta}_{i\ell}\right)^2 |Z_\ell^c] = \frac{1}{n}\sum_{i\in I_\ell} E[\hat{\Delta}_{i\ell}^2|Z_\ell^c]$$
$$\leq \int [m(z, \hat{\gamma}_\ell) - m(z, \gamma_0)]^2 F_0(dz) \xrightarrow{p} 0.$$

The conditional Markov inequality then implies that $\sum_{i\in I_\ell}\hat{\Delta}_{i\ell}/\sqrt{n} \xrightarrow{p} 0$. The analogous results also holds for $\hat{\Delta}_{i\ell} = \phi(z_i, \hat{\gamma}_\ell, \lambda_0) - \phi(z_i, \gamma_0, \lambda_0) - \bar{\phi}(\hat{\gamma}_\ell, \lambda_0)$ and $\hat{\Delta}_{i\ell} = \phi(z_i, \gamma_0, \hat{\lambda}_\ell) - \phi(z_i, \gamma_0, \lambda_0) - \bar{\phi}(\gamma_0, \hat{\lambda}_\ell)$. Summing across these three terms and across $\ell = 1, ..., L$ gives the first conclusion.

For the second conclusion, note that under the first hypothesis of Assumption 5,

$$E[\left|\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}[\phi_j(z_i, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi_j(z_i, \gamma_0, \hat{\lambda}_\ell) - \phi_j(z_i, \hat{\gamma}_\ell, \lambda_0) + \phi_j(z_i, \gamma_0, \lambda_0)]\right| |Z_\ell^c]$$
$$\leq \frac{1}{\sqrt{n}}\sum_{i\in I_\ell} E[\left|\phi_j(z_i, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi_j(z_i, \gamma_0, \hat{\lambda}_\ell) - \phi_j(z_i, \hat{\gamma}_\ell, \lambda_0) + \phi_j(z_i, \gamma_0, \lambda_0)\right| |Z_\ell^c]$$
$$\leq \sqrt{n}\int \left|\phi_j(z, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi_j(z, \gamma_0, \hat{\lambda}_\ell) - \phi_j(z, \hat{\gamma}_\ell, \lambda_0) + \phi_j(z_i, \gamma_0, \lambda_0)\right| F_0(dz) \xrightarrow{p} 0,$$

so $\hat{R}_2 \xrightarrow{p} 0$ follows by the conditional Markov and triangle inequalities. The second hypothesis of Assumption 5 is just $\hat{R}_2 \xrightarrow{p} 0$. *Q.E.D.*

Proof of Lemma 12: By Assumption 1 and the hypotheses that $\hat{\gamma}_i \in \Gamma$ and $\hat{\lambda}_i \in \Lambda$ we have $\hat{R}_3 = \hat{R}_4 = 0$. By Lemma A1 we have $\hat{R}_1 \xrightarrow{p} 0$ and $\hat{R}_2 \xrightarrow{p} 0$. The conclusion then follows by the triangle inequality. *Q.E.D.*

Proof of Theorem 13: Note that for $\varepsilon = y - \gamma_0(w)$

$$\phi(z, \hat{\gamma}, \lambda_0) - \phi(z, \gamma_0, \lambda_0) = \lambda_0(x)[\hat{\gamma}(w) - \gamma_0(w)],$$
$$\phi(z, \gamma_0, \hat{\lambda}) - \phi(z, \gamma_0, \lambda_0) = [\hat{\lambda}(x) - \lambda_0(x)]\varepsilon,$$
$$\phi(z, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi(z, \gamma_0, \hat{\lambda}_\ell) - \phi(z, \hat{\gamma}_\ell, \lambda_0) + \phi_j(z, \gamma_0, \lambda_0) = -[\hat{\lambda}(x) - \lambda_0(x)][\hat{\gamma}(x) - \gamma_0(x)].$$

The first part of Assumption 4 ii) then follows by

$$\int [\phi(z, \hat{\gamma}_\ell, \lambda_0) - \phi(z, \gamma_0, \lambda_0)]^2 F_0(dz) = \int \lambda_0(x)^2[\hat{\gamma}(w) - \gamma_0(w)]^2 F_0(dz)$$
$$\leq C\int [\hat{\gamma}(w) - \gamma_0(w)]^2 F_0(dz) \xrightarrow{p} 0.$$

The second part of Assumption 4 ii) follows by

$$\begin{aligned}\int [\phi(z,\gamma_0,\hat{\lambda}_\ell) - \phi(z,\gamma_0,\lambda_0)]^2 F_0(dz) &= \int [\hat{\lambda}_\ell(x) - \lambda_0(x)]^2 \varepsilon^2 F_0(dz) \\ &= \int \left[\hat{\lambda}_\ell(x) - \lambda_0(x)\right]^2 E[\varepsilon^2|x] F_0(dz) \\ &\leq C \int \left[\hat{\lambda}_\ell(x) - \lambda_0(x)\right]^2 F_0(dz) \xrightarrow{p} 0.\end{aligned}$$

Next, note that by the Cauchy-Schwartz inequality,

$$\begin{aligned}&\sqrt{n} \int |\phi(z,\hat{\gamma}_\ell,\hat{\lambda}_\ell) - \phi(z,\gamma_0,\hat{\lambda}_\ell) - \phi(z,\hat{\gamma}_\ell,\lambda_0) + \phi(z,\gamma_0,\lambda_0)| F_0(dz) \\ &= \sqrt{n} \int \left|[\hat{\lambda}_\ell(x) - \lambda_0(x)][\hat{\gamma}_\ell(w) - \gamma_0(w)]\right| F_0(dx) \\ &\leq \sqrt{n} \{\int [\hat{\lambda}_\ell(x) - \lambda_0(x)]^2 F_0(dx)\}^{1/2} \{\int [\hat{\gamma}_\ell(w) - \gamma_0(w)]^2 F_0(dw)\}^{1/2}.\end{aligned}$$

Then the first rate condition of Assumption 5 holds under the first rate condition of Theorem 13 while the second condition of Assumption 5 holds under the last hypothesis of Theorem 13. Then eq. (7.1) holds by Lemma 12, and the conclusion by rearranging the terms in eq. (7.1). *Q.E.D.*

Proof of Lemma 14: Follows by Lemma A1 and the triangle inequality. *Q.E.D.*

Proof of Lemma 15: Let $\hat{M}(\beta) = \partial\hat{\psi}(\beta)/\partial\beta$ when the derivative exists, $\tilde{M}_\ell = n^{-1}\sum_{i\in I_\ell} \partial\psi(z_i,\beta_0,\hat{\gamma}_\ell,\hat{\lambda}$ and $\bar{M}_\ell = n^{-1}\sum_{i\in I_\ell} \partial\psi(z_i,\beta_0,\gamma_0,\lambda_0)/\partial\beta$. By the law of large numbers, and Assumption 5 iii), $\sum_{\ell=1}^L \bar{M}_\ell \xrightarrow{p} M$. Also, by condition iii) for each j and k,

$$E[|\tilde{M}_{\ell jk} - \bar{M}_{\ell jk}||Z^\ell] \leq \int \left|\partial\psi_j(z,\beta_0,\hat{\gamma}_\ell,\hat{\lambda}_\ell)/\partial\beta_k - \partial\psi_j(z,\beta_0,\gamma_0,\lambda_0)/\partial\beta_k\right| F_0(dz) \xrightarrow{p} 0.$$

Then by the conditional Markov inequality, for each ℓ,

$$\tilde{M}_\ell - \bar{M}_\ell \xrightarrow{p} 0.$$

It follows by the triangle inequality that $\sum_{\ell=1}^L \tilde{M}_\ell \xrightarrow{p} M$. Also, with probability approaching one we have for any $\bar{\beta} \xrightarrow{p} \beta_0$

$$\left\|\hat{M}(\bar{\beta}) - \sum_{\ell=1}^L \tilde{M}_\ell\right\| \leq \left(\frac{1}{n}\sum_{i=1}^n d(z_i)\right) \left\|\bar{\beta} - \beta_0\right\|^{\zeta'} = O_p(1)o_p(1) \xrightarrow{p} 0.$$

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

Proof of Theorem 16: The conclusion follows in a standard manner from the conclusions of Lemmas 14 and 15. *Q.E.D.*

Proof of Theorem 17: Let $\hat{\psi}_i = \psi(z_i, \hat{\beta}, \hat{\gamma}_\ell, \hat{\lambda}_\ell)$ and $\psi_i = \psi(z_i, \beta_0, \gamma_0, \lambda_0)$. By standard arguments (e.g. Newey, 1994), it suffices to show that $\sum_{i=1}^n \left\|\hat{\psi}_i - \psi_i\right\|^2 / n \xrightarrow{p} 0$. Note that

$$\hat{\psi}_i - \psi_i = \sum_{j=1}^5 \hat{\Delta}_{ji}, \hat{\Delta}_{1i} = \psi(z_i, \hat{\beta}, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \psi(z_i, \beta_0, \hat{\gamma}_\ell, \hat{\lambda}_\ell), \hat{\Delta}_{2i} = m(z_i, \beta_0, \hat{\gamma}_\ell) - m(z_i, \beta_0, \gamma_0),$$
$$\hat{\Delta}_{3i} = \phi(z_i, \hat{\gamma}_\ell, \lambda_0) - \phi(z_i, \gamma_0, \lambda_0), \hat{\Delta}_{4i} = \phi(z_i, \gamma_0, \hat{\lambda}_\ell) - \phi(z_i, \gamma_0, \lambda_0),$$
$$\hat{\Delta}_{5i} = \phi(z_i, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi(z_i, \hat{\gamma}_\ell, \lambda_0) - \phi(z_i, \gamma_0, \hat{\lambda}_\ell) + \phi(z_i, \gamma_0, \lambda_0).$$

By standard arguments it suffices to show that for each j and ℓ,

$$\frac{1}{n} \sum_{i \in I_\ell} \left\|\hat{\Delta}_{ji}\right\|^2 \xrightarrow{p} 0. \tag{8.2}$$

For $j = 1$ it follows by a mean value expansion and Assumption 7 with $E[b(z_i)^2] < \infty$ that

$$\frac{1}{n} \sum_{i \in I_\ell} \left\|\hat{\Delta}_{1i}\right\|^2 = \frac{1}{n} \sum_{i \in I_\ell} \left\|\frac{\partial}{\partial \beta} \psi(z_i, \bar{\beta}, \hat{\gamma}_\ell, \hat{\lambda}_\ell)(\hat{\beta} - \beta)\right\|^2 \leq \frac{1}{n} \left(\sum_{i \in I_\ell} b(z_i)^2\right) \left\|\hat{\beta} - \beta\right\|^2 \xrightarrow{p} 0,$$

where $\bar{\beta}$ is a mean value that actually differs from row to row of $\partial \psi(z_i, \bar{\beta}, \hat{\gamma}_\ell, \hat{\lambda}_\ell)/\partial \beta$. For $j = 2$ note that by Assumption 4,

$$E[\frac{1}{n} \sum_{i \in I_\ell} \left\|\hat{\Delta}_{2i}\right\|^2 | Z^\ell] \leq \int \|m(z, \beta_0, \hat{\gamma}_\ell) - m(z, \beta_0, \gamma_0)\|^2 F_0(dz) \xrightarrow{p} 0,$$

so eq. (8.2) holds by the conditional Markov inequality. For $j = 3$ and $j = 4$ eq. (8.2) follows similarly. For $j = 5$, it follows by the hypotheses of Theorem 17 that

$$E[\frac{1}{n} \sum_{i \in I_\ell} \left\|\hat{\Delta}_{5i}\right\|^2 | Z^\ell] \leq \int \left\|\phi(z, \hat{\gamma}_\ell, \hat{\lambda}_\ell) - \phi(z, \gamma_0, \hat{\lambda}_\ell) - \phi(z, \hat{\gamma}_\ell, \lambda_0) + \phi(z, \gamma_0, \lambda_0)\right\|^2 F_0(dz) \xrightarrow{p} 0.$$

Then eq. (8.2) holds for $j = 5$ by the conditional Markov inequality. *Q.E.D.*

9 Appendix B: Doubly Robust Moment Functions for Orthogonality Conditions

It is interesting that the general condition $E[\lambda(x_i)\{y_i - \gamma_0(w_i)\}] = 0$ for all $\lambda \in \Gamma_2$ of Assumption 2 is like an identification condition for γ_0. For example, if Γ_2 is all functions of x_i with finite mean square then that condition is $E[y_i - \gamma_0(w_i)|x_i] = 0$, the nonparametric conditional moment restriction of Newey and Powell (2003) and Newey (1991). The first condition of Assumption 2

also has an interesting interpretation. Suppose that Γ_2 is a linear mean-square closed set and let $\Pi(\cdot|\Gamma_2)(x_i)$ denote the orthogonal projection on Γ_2. Then the condition is

$$\begin{aligned} E[m(z_i,\beta_0,\gamma)] &= -E[\lambda_0(x_i)\{y_i-\gamma(w_i)\}] = E[\lambda_0(x_i)\{\Pi(\gamma(w_i)|\Gamma_2)(x_i)-\Pi(\gamma_0(w_i)|\Gamma_2)(x_i)\}] \\ &= E[\lambda_0(x_i)\{\Pi(\gamma(w_i)-\gamma_0(w_i)|\Gamma_2)(x_i)\}]. \end{aligned}$$

Here we see that $E[m(z_i,\beta_0,\gamma)]$ is a linear, mean-square continuous function of $\Pi(\gamma(w_i)-\gamma_0(w_i)|\Gamma_2)(x_i)$. The Riesz representation theorem will also imply that if $E[m(z_i,\beta_0,\gamma)]$ is a linear, mean-square continuous function of $\Pi(\gamma(w_i)-\gamma_0(w_i)|\Gamma_2)(x_i)$ then $\lambda_0(x)$ exists satisfying the first part of Assumption 2. For the case where $w_i = x_i$ this mean-square continuity condition is necessary for existence of a root-n consistent estimator, as in Newey (1994) and Newey and McFadden (1994) We conjecture that when w_i need not equal x_i this condition generalizes Severini and Tripathi's (??) necessary condition for existence of a root-n consistent estimator of β_0.

From Proposition 5 it follows that the LR moment function of equation (**??**) will be DR when $m(z,\beta_0,\gamma_1)$ and $\rho(z,\gamma_1)$ are affine in γ_1. We can characterize DR moment conditions directly in a straightforward way. 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

$$E[m(z_i,\beta_0,\gamma_1)] = -E[\gamma^*(x_i)E[\rho(z_i,\gamma_1)|x_i]] = -E[\gamma^*(x_i)\rho(z_i,\gamma_1)], \gamma_1 \in \Gamma. \tag{9.1}$$

Let $\lambda_0(x)$ be any function such that $\lambda_0(x_i)-\gamma^*(x_i)$ is orthogonal to Π and $\psi(z,\beta,\gamma) = m(z,\beta,\gamma_1)+\gamma_2(x)\rho(z,\gamma_1)$. Then $E[\psi(z_i,\beta_0,\gamma_1,\lambda_0)]=0$ by the previous equation. It also follows that $E[\psi(z_i,\beta_0,\gamma_0,\lambda)] = 0$ by $E[\rho(z_i,\gamma_0)|x_i]=0$. Therefore $\psi(z,\beta,\gamma_1,\lambda)$ is DR, showing the following result:

PROPOSITION 9: *If $E[m(z_i,\beta_0,\gamma_1)]$ and $E[\rho(z_i,\gamma_1)|x_i]$ are affine in $\gamma_1\in\Gamma$ with Γ linear and $E[m(z_i,\beta_0,\gamma_1)]$ is a mean square continuous functional of $E[\rho(z_i,\gamma_1)|x_i]$ then there is $\lambda_0(x)$ such that $\psi(z,\beta,\gamma_1,\lambda) = m(z,\beta,\gamma_1)+\lambda(x)\rho(z,\gamma_1)$ is DR.*

It is interesting to note that λ_0 such that $E[\psi(z,\beta_0,\gamma_1,\lambda_0)] = 0$ for all $\gamma_1\in\Gamma$ is not unique when Π does not include all functions of x, which is the overidentified case of Chen and Santos (2015). This nonuniqueness can occur when there are multiple ways to estimate the first step γ_0 using the conditional moment restrictions $E[\rho(z_i,\gamma_0)|x_i]=0$. As discussed in Ichimura and Newey (2017), the different $\lambda_0(x_i)$ correspond to different first step estimators, with $\lambda_0(x_i)=\gamma^*(x_i)$ corresponding to the NP2SLS estimator.

An important class of DR moment conditions are those from the linear nonparametric IV setting in Newey and Powell (1989, 2003) and Newey (1991) where

$$\rho(z,\gamma_1) = y-\gamma_1(w), E[y_i-\gamma_0(w_i)|x_i] = E[\rho(z_i,\gamma_0)|x_i] = 0. \tag{9.2}$$

Consider a moment function $m(z,\beta,\gamma_1)$ such that $E[m(z_i,\beta_0,\gamma_1)]$ is affine in γ_1 and mean square continuous as a function of $\gamma_1(w_i)$. Then there exists $v(w_i)$ such that

$$E[m(z_i,\beta_0,\gamma_1)] = E[v(w_i)\{\gamma_1(w_i)-\gamma_0(w_i)].$$

Suppose also that there is $\bar\gamma(x)$ such that $v(w_i) = E[\bar\gamma(x_i)|w_i]$. Then we have

$$\begin{aligned} E[m(z_i,\beta_0,\gamma_1)] &= E[v(w_i)\{\gamma_1(w_i)-\gamma_0(w_i)\}] = E[E[\bar\gamma(x_i)|w_i]\{\gamma_1(w_i)-\gamma_0(w_i)\}] \\ &= E[\bar\gamma(x_i)\{\gamma_1(w_i)-\gamma_0(w_i)\}] = -E[\bar\gamma(x_i)E[\rho(z_i,\gamma_1)|x_i]] \\ &= -E[\bar\gamma(x_i)\rho(z_i,\gamma_1)] \end{aligned}$$

It then follows by Proposition 9 or by inspection that $\psi(z,\beta,\gamma) = m(z,\beta,\gamma_1)+\gamma_2(x)\{y-\gamma_1(w)\}$ is DR for $\lambda_0(x) = \bar\gamma(x_i)$. Interestingly, when $m(z,\beta,\gamma_1) = v(w)\gamma_1(w)-\beta$, 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 DR moment condition can always be constructed when this necessary condition is satisfied. Also, similarly to the above, the $\lambda_0(x)$ may not be unique.

PROPOSITION 10: *If $E[m(z_i,\beta_0,\gamma_1)]$ is affine and mean square continuous in γ_1, with Riesz representer $v(w_i)$, equation (9.2) is satisfied for $\rho(z,\gamma_1) = y-\gamma_1(w)$, and there is $\bar\gamma(x)$ such that $v(w_i) = E[\bar\gamma(x_i)|w_i]$ then $\psi(z,\beta,\gamma_1,\gamma_2) = m(z,\beta,\gamma_1)+\gamma_2(x)[y-\gamma_1(w)]$ is DR for any λ_0 with $\lambda_0(x)-\bar\gamma(x)$ orthogonal to Π.*

The partial robustness results of the last Section can be extended to conditional moment restrictions. Note that by the conditional mean zero restriction in equation (??), $\gamma_0(w)$ has the property that $E[\{y_i-\gamma_0(w_i)\}a(x_i)] = 0$ for all $a(x_i)$ with $E[a(x_i)^2] < \infty$, that is $y_i-\gamma_0(w_i)$ is orthogonal to the set $\mathcal{A}$ of functions of x_i with finite mean square. Let $\mathcal{A}^*$ be a closed linear subset of $\mathcal{A}$, such as a finite dimensional subset, and let γ_0 be such that $E[\{y_i-\gamma_0(w_i)\}a^*(x_i)] = 0$ for all $a^*\in\mathcal{A}^*$. For example, if $\mathcal{A}^*$ is finite dimensional with basis $A(x)$ then one could construct $\gamma_0(w) = p(w)'\delta^*$ for the population instrument variable coefficients $\delta^* = (E[A(x_i)p(w_i)'])^{-1}E[A(x_i)y_i]$ when $E[A(x_i)p(w_i)']$ is nonsingular. Then if there is $v(w_i)$ with $E[m(z_i,\beta_0,\gamma_1)] = E[v(w_i)\{\gamma_1(w_i)-\gamma_0(w_i)\}]$ and $\alpha^*(x_i)\in\mathcal{A}^*$ with $v(w_i) = E[\alpha^*(x_i)|w_i]$ we have

$$\begin{aligned} E[m(z_i,\beta_0,\gamma_1^*)] &= E[v(w_i)\{\gamma_1^*(w_i)-\gamma_0(w_i)\}] = E[E[\alpha^*(x_i)|w_i]\{\gamma_1^*(w_i)-\gamma_0(w_i)\}] \\ &= E[\alpha^*(x_i)\{\gamma_1^*(w_i)-\gamma_0(w_i)\}] = -E[\alpha^*(x_i)\{y_i-\gamma_1^*(w_i)\}]+E[\alpha^*(x_i)\{y_i-\gamma_0(w_i)\}] = 0. \end{aligned}$$

Thus we have the following result:

PROPOSITION 10A: *If $E[m(z_i,\beta_0,\gamma_1)]$ is affine and mean square continuous in γ_1, with Riesz representer $v(w_i)$, equation (9.2) is satisfied for $\rho(z,\gamma_1) = y-\gamma_1(w)$, γ_1^** satisfies $E[a^*(x_i)\{y_i-$

$\gamma_1^*(w_i)\}] = 0$ for all $a^* \in \mathcal{A}^*$, *and there exists* $\alpha^* \in \mathcal{A}^*$ with $v(w_i) = E[\alpha^*(x_i)|w_i]$, *then* $E[m(z_i, \beta_0, \gamma_1^*)] = 0$.

As an example we can show that when w_i and x_i have the same dimension, w_i is Gaussian and $E[x_i|w_i]$ is linear in w_i, linear instrumental variables (IV) estimates the average derivative of the structural function $\gamma_0(x_i)$. This result extends Stoker (1986) to instrumental variables. Assuming w_i and x_i include a constant as their last element and $E[x_i w_i']$ is nonsingular let $\delta^* = (E[x_i w_i'])^{-1} E[x_i y_i]$ denote the limit of the linear IV estimator with right-hand side variables w and instruments x.

i

PROPOSITION 10B: *If the* $\gamma_0(w_i)$ *satisfies* $E[\{y_i - \gamma_0(w_i)\}|x_i] = 0$, *the nonconstant elements of* w_i *are multivariate Gaussian with nonsingular variance matrix,* $E[x_i|w_i] = \Pi w_i$, *and* Π *is nonsingular, then* $E[x_i w_i']$ *is nonsingular and for any nonconstant element* w_{ir} *of* w_i

$$\delta_r^* = E[\partial \gamma_0(w_i)/\partial w_r].$$

Proof of Proposition 10b: Then by iterated expectations $E[x_i w_i'] = E[E[x_i|w_i]w_i'] = \Pi E[w_i w_i']$ is nonsingluar. Let $m(z, \beta, \gamma) = \partial \gamma(w)/\partial w_r - \beta$ for any nonconstant element w_r of w. Then by integration by parts and w_i being Gaussian there is a constant vector c such that

$$E[m(z_i, \beta_0, \gamma_1)] = E[v(w_i)\{\gamma_1(w_i) - \gamma_0(w_i)\}], v(w) = -\partial f_0(w)/\partial w_r = c'w.$$

Let $\mathcal{A}^*$ denote the set of linear combinations of x_i and $\gamma_1^*(w_i) = \delta^{*\prime} w_i$. The normal equations for linear IV imply $E[a^*(x_i)\{y_i - \gamma_1^*(w_i)\}] = 0$ for all $a^* \in \mathcal{A}^*$. Also, by $E[x_i|w_i] = \Pi w_i$ and Π nonsingular it follows that for $\alpha^*(x_i) = c'\Pi^{-1} x_i \in \mathcal{A}^*$,

$$E[\alpha^*(x_i)|w_i] = c'\Pi^{-1} E[x_i|w_i] = c'\Pi^{-1}\Pi w_i = c'w_i = v(w_i).$$

Then by Proposition 11 we have $E[m(z_i, \beta_0, \gamma_1^*)] = 0$, i.e.

$$E[\partial \gamma_0(w_i)/\partial w_r] = \beta_0 = E[\partial \gamma_1^*(w_i)/\partial w_r] = \delta_r^*. Q.E.D.$$

DR moment conditions can be used to identify parameters of interest. In general, if $\psi(z, \beta, \gamma_1, \gamma_2)$ is DR and λ_0 is identified then β_0 may be identified from

$$E[\psi(z_i, \beta_0, \bar{\gamma}_1, \lambda_0)] = 0,$$

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

Acknowledgements

Whitney Newey gratefully acknowledges support by the NSF. This paper was presented at Econometric Society meetings and the Vanderbilt conference. Helpful comments were provided by M. Cattaneo, B. Deaner, J. Hahn, M. Jansson, Z. Liao, A. Pakes, R. Moon, A. de Paula, V. Semenova, and participants in seminars at Cambridge, Columbia, Cornell, Harvard-MIT, UCL, USC, Yale, and Xiamen. B. Deaner provided capable research assistance.

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.

Athey, S., G. Imbens, and S. Wager (2017): "Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing," *Journal of the Royal Statistical Society, Series B,* forthcoming.

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. (1982): "On Adaptive Estimation," *Annals of Statistics* 10, 647-671.

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 (2017): "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," *Econometrica*, forthcoming.

Cattaneo, M.D., M. Jansson, and X. Ma (2017): "Two-step Estimation and Inference with Possibly Many Included Covariates," working paper.

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 Z. Liao (2015):

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

Chernozhukov, V., and C. Hansen (2004): ?? *Econometrica* .

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., G.W. Imbens and W.K. Newey (2007): "Instrumental Variable Identification and Estimation of Nonseparable Models," *Journal of Econometrics* 139, 4-14.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins (2018): "Debiased/Double Machine Learning for Treatment and Structural Parameters," with V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and J. Robins, *Econometrics Journal ??.*

7.

Escanciano, J-C., D. Jacho-Cha´vez, and A. Lewbel (2016): "Identification and Estimation of Semiparametric Two Step Models", *Quantitative Economics* 7, 561-589.

Firpo, S. and C. Rothe (2015): "Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions," working paper.

Graham, B.W. (): ??.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66, 315-331.

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

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.

Hausman, J.A., and W.K. Newey (2017): "Nonparametric Welfare Analysis," *Annual Review of Economics* 9, 521–546.

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

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

Huber, P. (1981): *Robust Statistics,* New York: Wiley.

Ichimura, H. (1993): "Estimation of Single Index Models," *Journal of Econometrics* 58, 71-120.

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

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

Kandasamy, K., A. Krishnamurthy, B. P´oczos, L. Wasserman, J.M. Robins (2015): "Influence Functions for Machine Learning: Nonparametric Estimators for Entropies, Divergences and Mutual Informations," ArXiv.

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

Murphy, K.M. and R.H. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics* 3, 370-379.

Newey, W.K. (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters* 14, 201-206.

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. (1994a): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.

NEWEY, W.K. (1994b): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory* 10, 233-253.

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

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, 1988.

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.

Pfanzagle, W. (1981) ??

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.

ROBINSON, P.M. (1988): "'Root-N-consistent Semiparametric Regression," *Econometrica* 56, 931-954.

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.

SCHICK, A. (1986): "On Asymptotically Efficient Estimation in Semiparametric Models," *Annals of Statistics* 14, 1139-1151.

Severini, T. A„ and W. H. Wong

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 Laan, M. and Rubin (2006); ??

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.

VAN DER VAART, A.W. (2014): "Higher Order Tangent Spaces and Influence Functions," Statistical Science 29, 679–686.

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.