General Doubly Robust Identification and Estimation *

Arthur Lewbel, Jin-Young Choi, and Zhuzhu Zhou
Boston College, Goethe University, and Boston College

original Nov. 2016, revised Mar. 2019
THIS VERSION IS PRELIMINARY AND INCOMPLETE

Abstract

Consider two parametric models. At least one is correctly specified, but we don't know which. Both models include a common vector of parameters. An estimator for this common parameter vector is called Doubly Robust (DR) if it's consistent no matter which model is correct. We provide a general technique for constructing DR estimators. Our General Doubly Robust (GDR) technique is a simple extension of the Generalized Method of Moments. We illustrate our GDR with a variety of models, including average treatment effect estimation. Our empirical application is instrumental variables estimation, where either one of two instrument vectors might be invalid.

JEL codes: C51, C36, C31, Keywords: Doubly Robust Estimation, Generalized Method of Moments, Instrumental Variables, Average Treatment Effects, Parametric Models

*Corresponding Author: Arthur Lewbel, Department of Economics, Maloney 315, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://www2.bc.edu/arthur-lewbel/

1 Introduction

Consider two different parametric models, which we will call G and H. One of these models is correctly specified, but we don't know which one (or both could be right). Both models include the same parameter vector α. An estimator $\widehat{\alpha}$ is called *Doubly Robust* (DR) if $\widehat{\alpha}$ is consistent no matter which model is correct.

We provide a general technique for constructing doubly robust (DR) estimators, which we call General Doubly Robust (GDR) estimation. The technique can be immediately extended to triply robust and general multiply robust models (where at least one of three or more proposed model is correct). Our GDR takes the form of a weighted average of Hansen's (1982) Generalized Method of Moments (GMM) based estimates of α, and has similar associated root-n asymptotics (albeit with some complications that when both models are correct).

The term double robustness was coined by Robins, Rotnitzky, and van der Laan (2000), but is based on Scharfstein, Rotnitzky, and Robins (1999) and the augmented inverse probability weighting average treatment effect estimator introduced by Robins, Rotnitzky, and Zhao (1994). In their application α is a population Average Treatment Effect (ATE). To summarize their application, suppose we have data consisting of n observations of a random vector Z. Let $\widetilde{G}(Z,\beta)$ be a proposed functional form for the expectation of an outcome given a binary treatment indicator and a vector of other observed covariates. Let G denote the model for α based on $\widetilde{G}$, that is, the expectation of the difference between $\widetilde{G}$ in the treatment group and the control group. Let $\widetilde{H}(Z,\gamma)$ be a proposed functional form for the propensity score, that is, the probability of being given treatment as a function of covariates. Then H is the model for the ATE α based on $\widetilde{H}$, i.e., the expected difference between propensity score weighted outcomes. A DR estimator $\widehat{\alpha}$ is then an estimator for the ATE α that is consistent if either $\widetilde{G}$ or $\widetilde{H}$ is (or both are) correctly specified. See, e.g., Słoczyński and Wooldridge (2018), Wooldridge (2007), Bang and Robins (2005), Rose and van der Laan (2014), Funk, Westreich, Wiesen, Stürmer, Brookhart, and Davidian (2011), Robins, Rotnitzky, and van der Laan (2000), and Scharfstein, Rotnitzky, and Robins (1999).

In this treatment effect example, as in most DR applications, one could consistently estimate α based on a nonparametric estimator of either the conditional outcome or the propensity score. That is, the functional forms of either $\widetilde{G}$ or $\widetilde{H}$ could be replaced with nonparametric estimators of these functions, which would then be substituted into the models G or H to consistently estimate α. But there are a number of potential problems associated with nonparametric estimation, e.g., it can be impractical at moderate sample sizes due to the curse of dimensionality, and it can be sensitive to the choice of tuning parameters like kernel and bandwidth choice.

The alternative to nonparametric estimation provided by DR estimators is to parameterize both $\widetilde{G}$ and $\widetilde{H}$. DR methods avoid the complications associated with nonparametric estimation, but still provide some insurance against misspecification, since only one of the two models G or H needs to be correctly specified, and the user doesn't need to know which one is correct. Our GDR estimator has these same benefits. Unlike nonparametric estimators, GDR requires no smoothing functions, tuning parameters, regularization, or penalty functions, and converges at the parametric root N rate. And unlike standard parametric models, GDR provides two chances instead of just one to correctly specify a functional form.

An alternative approach to modeling if one thought that either G or H was correctly specified would be to engage in some form of model selection. Model selection has some disadvantages relative doubly robust methods, e.g., one needs to correct limiting distributions for pretest bias, tests for which model is superior can be inconclusive, and choosing just one model will be inefficient if both specifications are correct. In the context of GMM based models, selection methods like Andrews and Lu (2001) and Liao (2013) attempt to choose which moments from a set of possible moments are valid. This is somewhat different from our problem, which deals with collections of moments being valid or invalid.

Another alternative would be model averaging, which is generally not consistent unless both $\widetilde{G}$ and $\widetilde{H}$ happen to be correctly specified. Like DR, our GDR avoids these issues. However, our GDR estimator does take the form of a weighted average of GMM estimates of α, and so closely

resembles GMM model averaging. A number of model averaging estimators exist for GMM and related models. Kuersteiner and Okui (2010) apply Hansen's (2007) model averaging criterion for instruments in linear instrumental variables models. Averaging across instruments or moments in GMM models is also considered by Martins and Gabriel (2014), Sueishi (2013), and DiTraglia (2016). Unlike these, we do not use typical model averaging criteria like mean squared error or Bayes weights or information criteria to choose weights. Instead, weights are chosen to yield the DR consistency property.

The main drawback of existing DR estimators is that they are not generic, in that for each problem one needs to design a specific DR estimator, which can then only be used for that one specific application. Existing DR applications require that one find some clever way of expressing α as the mean of functions of both $\widetilde{G}(Z,\beta)$ and $\widetilde{H}(Z,\gamma)$ that happens to possess the DR property. In the ATE example, this expression is given by equation (8) below, which has the tricky DR property of equaling the true α if either $\widetilde{G}$ or $\widetilde{H}$ is correctly specified. No general method exists for finding or constructing such equations, and only few examples of such models are known in the literature. Perhaps the closest thing to a general method may be Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018), who provide a way of constructing locally robust estimators, which they show sometimes leads to double robustness. In contrast to these limited results, our GDR provides a simple general method of constructing estimators that have the DR property.

Existing DR applications express the parameter α as a function of $\widetilde{G}(Z,\beta)$, $\widetilde{H}(Z,\gamma)$, and Z, where $\widetilde{G}$ and $\widetilde{H}$ are conditional mean functions. We further generalize by assuming that the true value of α satisfies either $E[G(Z,\alpha,\beta)] = 0$ or $E[H(Z,\alpha,\gamma)] = 0$ for some known vector valued functions G and H. Our GDR estimator then consistently estimates α, despite not knowing which of these two sets of equalities actually holds, for any functions G and H that satisfy some regularity and identification conditions.

Unlike existing DR estimators, we do not need to find some clever, model specific way to combine these moments. All that is needed to apply our estimator is to know the functions G and H. For

example, for estimation of the average treatment effect α, the function G is just moments implied by the standard expression for α as the difference in expected outcomes between treated and control groups, while the function H consists of moments implied by the standard expression of α as the mean of propensity score weighted outcomes.

We do not claim that our GDR estimator is superior to existing DR estimators in applications where DR estimators are known to exist. Rather, our primary contribution is providing a general method for constructing estimators that possess the DR property in applications where no DR estimator is known. Also, our GDR estimator has an extremely simple numerical form, and an ordinary root N consistent, asymptotically normal limiting distribution.

In the next section we describe our GDR estimator. Section 3 then gives four examples of potential applications. In section 4 we prove consistency of the GDR estimator and provide limiting distribution theory. Later sections provide an empirical application, and discuss extensions, including to triply and other mulitply robust estimators.

2 The GDR Estimator

In this section we describe the GDR estimator (proof of consistency and limiting distribution theory is provided later). Let Z be a vector of observed random variables, let α, β and γ be vectors of parameters, and assume G and H are known functions. Assume a sample consisting of n iid observations z_i of the vector Z. The goal is root-n consistent estimation of α.

Let $g_0(\alpha, \beta) \equiv E\{G(Z, \alpha, \beta)\}$, $h_0(\alpha, \gamma) \equiv E\{H(Z, \alpha, \gamma)\}$, $\theta_0 \equiv \{\alpha_0, \beta_0, \gamma_0\}$, and $\theta \equiv \{\alpha, \beta, \gamma\}$.

Assumption A1: For a compact sets Θ_α, Θ_β, and Θ_γ, $\alpha_0 \in \Theta_\alpha$, $\beta_0 \in \Theta_\beta$, and $\gamma_0 \in \Theta_\gamma$. Let $\Theta = \Theta_\alpha \times \Theta_\beta \times \Theta_\gamma$.

Assumption A2: Either 1) $g_0(\alpha_0, \beta_0) = 0$, or 2) $h_0(\alpha_0, \gamma_0) = 0$, or both hold.

Assumption A2 says that either the G model is true or the H model is true (or both are true),

for some unknown true coefficient values α_0, β_0, and γ_0. This is a defining feature DR estimators, and hence of our GDR estimator.

Assumption A3: For any $\{\alpha, \beta, \gamma\} \in \Theta$, if $g_0(\alpha, \beta) = 0$ then $\{\alpha, \beta\} = \{\alpha_0, \beta_0\}$, and if $h_0(\alpha, \gamma) = 0$ then $\{\alpha, \gamma\} = \{\alpha_0, \gamma_0\}$.

Assumptions A2 and A3 are identification assumptions. They imply that if G is the true model, then the true values of the coefficients $\{\alpha_0, \beta_0\}$ are identified by $g_0(\alpha_0, \beta_0) = 0$, and if H is the true model, then the true values of the coefficients $\{\alpha_0, \gamma_0\}$ are identified by $h_0(\alpha_0, \gamma_0) = 0$. Assumption A3 rules out the existence of alternative pseudo-true values satisfying the 'wrong' moments, e.g., this assumption rules out having both $g_0(\alpha_0, \beta_0) = 0$ and $g_0(\alpha_1, \beta_1) = 0$ for some $\alpha_1 \neq \alpha_0$.

Note that Assumption A3 is a potentially strong restriction, and is not required by some existing DR estimators. As our examples later will illustrate, satisfying this assumption generally requires that parameters be over identified, which in turn typically means that the vector G contains more elements than the set $\{\alpha, \beta\}$, and that the vector H contains more elements than the set $\{\alpha, \gamma\}$. Otherwise, as in method of moments estimation, $g_0(\alpha, \beta) = 0$ and $h_0(\alpha, \gamma) = 0$ each have as many equations as unknowns, and so typically a pseudo-true solution will exist for whichever one is misspecified (if one is), thereby violating Assumption A3.

Define the following functions:

$$\widehat{g}(\alpha, \beta) \equiv \frac{1}{n} \sum_{i=1}^{n} G(Z_i, \alpha, \beta), \qquad \widehat{h}(\alpha, \gamma) \equiv \frac{1}{n} \sum_{i=1}^{n} H(Z_i, \alpha, \gamma),$$

$$\hat{Q}^g(\alpha, \beta) \equiv \widehat{g}(\alpha, \beta)' \widehat{\Omega}_g \widehat{g}(\alpha, \beta), \qquad \hat{Q}^h(\alpha, \gamma) \equiv \widehat{h}(\alpha, \gamma)' \widehat{\Omega}_h \widehat{h}(\alpha, \gamma),$$

where $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$ are positive definite matrices. Note that $\hat{Q}^g(\alpha, \beta)$ is the standard Hansen (1982) and Hansen and Singleton (1982) Generalized Method of Moments (GMM) objective function that would be used to estimate α and β if G were correctly specified. Similarly, $\hat{Q}^h(\alpha, \beta)$ is the GMM objective function that would be used to estimate α and γ if H were correctly specified. Define $\widehat{\alpha}_g$,

$\widehat{\beta}_g$, $\widehat{\alpha}_h$, and $\widehat{\gamma}_h$ by

$$\{\widehat{\alpha}_g, \widehat{\beta}_g\} = \arg \min_{\{\alpha,\beta\} \in \Theta_\alpha \times \Theta_\beta} \hat{Q}^g(\alpha, \beta) \quad \text{and} \quad \{\widehat{\alpha}_h, \widehat{\gamma}_h\} = \arg \min_{\{\alpha,\gamma\} \in \Theta_\alpha \times \Theta_\gamma} \hat{Q}^h(\alpha, \gamma). \tag{1}$$

So $\{\widehat{\alpha}_g, \widehat{\beta}_g\}$ is nothing more than the standard GMM estimate of model G, and $\{\widehat{\alpha}_h, \widehat{\gamma}_h\}$ is the standard GMM estimate of model H.

Define $\widehat{W}_g$ by

$$\widehat{W}_g \equiv \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)} \tag{2}$$

Our proposed GDR estimator $\widehat{\alpha}$ is then given by

$$\widehat{\alpha} = \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\widehat{\alpha}_h + \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\widehat{\alpha}_g}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)} = \widehat{W}_g \widehat{\alpha}_h + \left(1 - \widehat{W}_g\right) \widehat{\alpha}_g \tag{3}$$

So our GDR estimate $\widehat{\alpha}$ is simply a weighted average of the GMM estimates $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$, where the weights are proportional to the GMM objective functions $\hat{Q}^h$ and $\hat{Q}^h$.

The intuition behind our GDR estimator is straightforward. Suppose model model H is wrong and model G is right, so $E\left[H(Z, \alpha, \gamma)\right] \neq 0$ for any α and γ, and $E\left[G(Z, \alpha_0, \beta_0)\right] = 0$. Then $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ goes in probability to zero while the limiting value of $\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ is nonzero. So the weight on $\widehat{\alpha}_h$ in equation (3) will go to zero, and the weight on $\widehat{\alpha}_g$ will be nonzero. As a result, $\widehat{\alpha}$ will have the same probability limit as $\widehat{\alpha}_g$, and since model G is right, this probability limit will be α_0. The same logic applies if model H is right and G is wrong, switching the roles of g and h, and the roles of β and γ. Finally, if both models are right, then $\widehat{\alpha}$ is just a weighted average of consistent estimators of α_0, and so is consistent no matter what values the weights take on. We therefore obtain the double robustness property that, whichever model is right, $\widehat{\alpha} \to^p \alpha_0$.

Regarding the weighting matrices $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$, note that the usual GMM weight matrices might not be optimal (in terms of efficiency) for the third step in our GDR estimator. However, there are advantages to just using the standard GMM weight matrices, which we discuss in section 4. As a result, in equation (3) we let $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$ be the standard GMM weighting matrix estimates obtained from the first step of two step GMM, we take $\hat{Q}^g(\alpha, \beta)$ and $\hat{Q}^h(\alpha, \gamma)$ to be the second step objective

functions in the standard two step GMM estimator, and we take $\{\widehat{\alpha}_g, \widehat{\beta}_g\}$ and $\{\widehat{\alpha}_h, \widehat{\gamma}_h\}$ to be the standard two step GMM estimates of each model.

3 GDR Examples

Before proceeding to show consistency and deriving the limiting distribution of the GDR estimator, we consider four example applications. The first two examples show how GDR could be used in place of existing DR applications. The second two examples are new applications for which no existing DR estimator were known.

3.1 Average Treatment Effect

Harking back to the earliest DR estimators like Robins, Rotnitzky, and van der Laan (2000), Scharfstein, Rotnitzky, and Robins (1999), and Robins, Rotnitzky, and Zhao (1994), here we describe the construction of DR estimates of average treatment effects, as in, e.g., Bang and Robins (2005), Funk, Westreich, Wiesen, Stürmer, Brookhart, and Davidian (2011), Rose and van der Laan (2014), Lunceford and Davidian (2004), Słoczyński and Wooldridge (2018) and Wooldridge (2007). We then show how this model could alternatively be estimated using our GDR construction. Note that other DR estimators of treatment effects also exist, e.g., Lee and Lee (2018).

The assumption in this application is that either the conditional mean of the outcome or the propensity score of treatment is correctly parametrically specified. Let $Z = \{Y, T, X\}$ where Y is an outcome, T is a binary treatment indicator, and X is a J vector of other covariates (including a constant). The average treatment effect we wish to estimate is

$$\alpha = E\{E(Y|T=1,X) - E(Y|T=0,X)\}. \tag{4}$$

As is well known, an alternative propensity score weighted expression for the same average treatment effect is

$$\alpha = E\left\{\frac{YT}{E(T|X)} - \frac{Y(1-T)}{1-E(T|X)}\right\}. \tag{5}$$

Let $\widetilde{G}(T, X, \beta)$ be the proposed functional form of the conditional mean of the outcome, for some K vector of parameters β. So if $\widetilde{G}$ is correctly specified, then $\widetilde{G}(T, X, \beta) = E(Y|T, X)$. Similarly, let $\widetilde{H}(X, \gamma)$ be the proposed functional form of the propensity score for some J vector of parameters γ, so if $\widetilde{H}$ is correctly specified, then $\widetilde{H}(X, \gamma) = E(T|X)$.

One standard estimator of α, based on equation (4), consists of first estimating β by least squares, minimizing the sample average of $E[\{Y - \widetilde{G}(T, X, \beta)\}^2]$, and then estimating α as the sample average of $\widetilde{G}(1, X, \beta) - \widetilde{G}(0, X, \beta)$. This estimator is equivalent to GMM estimation of α and β, using the vector of moments

$$E\left[\begin{array}{c} \{Y - \widetilde{G}(T, X, \beta)\} r_1(T, X) \\ \alpha - \{\widetilde{G}(1, X, \beta) - \widetilde{G}(0, X, \beta)\} \end{array}\right] = 0 \tag{6}$$

for some vector valued function $r_1(T, X)$. Least squares estimation of β specifically chooses $r_1(T, X)$ to equal $\partial \widetilde{G}(T, X, \beta)/\partial\beta$, but alternative functions could be used, corresponding to, e.g., weighted least squares estimation, or to the score functions associated with a maximum likelihood based estimator of β, given a parameterization for the error terms $Y - \widetilde{G}(T, X, \beta)$. Note that to identify the K vector β, the function $r_1(T, X)$ needs to be a $\widetilde{K}$ vector for some $\widetilde{K} \geq K$. The problem with this estimator is that in general α will not be consistently estimated if the functional form of $\widetilde{G}(T, X, \beta)$ is not the correct specification of $E(Y|T, X)$.

An alternative common estimator of α, based on equation (5), consists of first estimating γ by least squares, minimizing the sample average of $E[\{T - \widetilde{H}(X, \gamma)\}^2]$, and then estimating α as the sample average of $\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)}$. This estimator is equivalent to GMM estimation of α and γ, using the vector of moments

$$E\left[\begin{array}{c} \{T - \widetilde{H}(X, \gamma)\} r_2(X) \\ \alpha - \left\{\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)}\right\} \end{array}\right] = 0 \tag{7}$$

for some $\widetilde{J}$ vector valued function $r_2(X)$. As above, least squares estimation of γ sets $r_2(X)$ equal to $\partial \widetilde{H}(X, \gamma)/\partial\gamma$, but as above alternative functions could be chosen for $r_2(X)$. To identify the J vector γ, the function $r_2(X)$ needs to be a $\widetilde{J}$ vector for some $\widetilde{J} \geq J$. With this estimator,

in general α will not be consistently estimated if the functional form of $\widetilde{H}(X,\gamma)$ is not the correct specification of $E(T|X)$.

A doubly robust estimator like that of Bang and Robins (2005) and later authors assumes α can be expressed as

$$\alpha = E\left\{\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)} + \frac{T-\widetilde{H}(X,\gamma)}{\widetilde{H}(X,\gamma)}\widetilde{G}(1,X,\beta) - \frac{T-\widetilde{H}(X,\gamma)}{1-\widetilde{H}(X,\gamma)}\widetilde{G}(0,X,\beta)\right\}. \quad (8)$$

Observe that if $\widetilde{H}(X,\gamma) = E(T|X)$, then the first two terms in the above expectation equal equation (5) and the second two terms have mean zero. By rearranging terms, equation (8) can be rewritten as

$$\alpha = E\left[\widetilde{G}(1,X,\beta) - \widetilde{G}(0,X,\beta) + \frac{T}{\widetilde{H}(X,\gamma)}\{Y-\widetilde{G}(1,X,\beta)\} - \frac{1-T}{1-\widetilde{H}(X,\gamma)}\{Y-\widetilde{G}(0,X,\beta)\}\right]. \quad (9)$$

Rewriting the equation this way, it can be seen that if $\widetilde{G}(T,X,\beta) = E(Y|T,X)$, then the first two terms in equation (9) equal equation (4), and the second two terms have mean zero. This shows that equation (8) or equivalently (9) is doubly robust, in that it equals the average treatment effect α if either $\widetilde{G}(T,X,\beta)$ or $\widetilde{H}(X,\gamma)$ is correctly specified. The GMM estimator associated with this doubly robust estimator estimates α, β, and γ, using the moments

$$E\begin{bmatrix} \{Y-\widetilde{G}(T,X,\beta)\}r_1(T,X) \\ \{T-\widetilde{H}(X,\gamma)\}r_2(X) \\ \alpha - \left\{\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)} + \frac{T-\widetilde{H}(X,\gamma)}{\widetilde{H}(X,\gamma)}\widetilde{G}(1,X,\beta) - \frac{T-\widetilde{H}(X,\gamma)}{1-\widetilde{H}(X,\gamma)}\widetilde{G}(0,X,\beta)\right\} \end{bmatrix} = 0. \quad (10)$$

Construction of this doubly robust estimator required finding an expression like equation (8) that is special to the problem at hand. In general, finding such expressions for any particular problem may be difficult or impossible.

In contrast, our proposed GDR estimator does not require any such creativity. All that is required for constructing the GDR for this problem is to know the two alternative standard estimators, based on equations (4) and (5), expressed in GMM form, i.e., equation (6) and equation (7). Just define $G(Z,\alpha,\beta)$ to be the vector of functions given in equation (6) and define $H(Z,\alpha,\gamma)$ to be the

vector of functions given in equation (7). That is,

$$G(Z,\alpha,\beta)=\begin{bmatrix}\{Y-\widetilde{G}\,(T,X,\beta)\}r_1\,(T,X)\\ \alpha-\{\widetilde{G}\,(1,X,\beta)-\widetilde{G}\,(0,X,\beta)\}\end{bmatrix} \tag{11}$$

and

$$H(Z,\alpha,\gamma)=\begin{bmatrix}\{T-\widetilde{H}\,(X,\gamma)\}r_2\,(X)\\ \alpha-\left\{\frac{YT}{\widetilde{H}(X,\gamma)}-\frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)}\right\}\end{bmatrix}. \tag{12}$$

These functions can then be plugged into the expressions in the previous section to obtain our GDR estimator, equation (3), without having to find an expression like equation (8) with its difficult to satisfy properties.

The vector $r_2\,(X)$ can include any functions of X as long as the corresponding moments $E\{H(Z,\alpha,\gamma)\}$ exist. To satisfy Assumption A3, we will want to choose $r_2\,(X)$ to include $\widetilde{J}$ elements where $\widetilde{J}$ is strictly greater than J. What we require is that, if the propensity score is incorrectly specified, then there is no α,γ (in the set of permitted values) that satisfies the moments $E\{H(Z,\alpha,\gamma)\}=0$, while, if the propensity score is correctly specified, then the only α,γ that satisfies $E\{H(Z,\alpha,\gamma)\}=0$ is α_0,γ_0. By the same logic, we will want to choose the $\widetilde{K}$ vector $r_1\,(T,X)$ to include strictly more than K elements. For efficiency, it could be sensible to let $r_2\,(X)$ and $r_1\,(T,X)$ include $\partial\widetilde{H}\,(X,\gamma)\,/\partial\gamma$ and $\partial\widetilde{G}\,(T,X,\beta)\,/\partial\beta$, respectively.

3.2 An Instrumental Variables Additive Regression Model

Okui, Small, Tan, and Rubins (2012) propose a DR estimator for an instrumental variables (IV) additive regression model. The model is the additive regression

$$Y=M(W,\alpha)+\widetilde{G}(X)+U, \tag{13}$$

$$E(Q\mid X)=\widetilde{H}(X),$$

$$E\,(U\mid X,Q)=0, \tag{14}$$

where Y is an observed outcome variable, W is a S vector of observed exogenous covariates, X is a J vector of observed confounders, and Q is a $K\geq J$ vector of observed instruments. Note

that this model has features that are unusual for instrumental variables estimation, in particular, the assumption that $E\left(U \mid X, Q\right) = 0$ is stronger than the usual $E\left(U \mid Q\right) = 0$ assumption. The function $M(W, \alpha)$ is assumed to be correctly parameterized, and the goal is estimation of α.

Okui, Small, Tan, and Rubins (2012) construct a DR estimator assuming that, in addition to the above, either $\widetilde{G}(X) = \widetilde{G}(X, \beta)$ is correctly parameterized, or that $\widetilde{H}(X) = \widetilde{H}(X, \gamma)$ is correctly parameterized. Let $Z = \{Y, W, X, Q\}$, and let $r_1(X)$ and $r_2(X)$ be vectors of functions chosen by the user. Define $G(\alpha, \beta, Z)$ and $H(\alpha, \gamma, Z)$ by

$$G(Z, \alpha, \beta) = \begin{bmatrix} \{Y - M(W, \alpha) - \widetilde{G}(X, \beta)\} r_1(X) \\ \{Y - M(W, \alpha) - \widetilde{G}(X, \beta)\} Q \end{bmatrix} \tag{15}$$

and

$$H(Z, \alpha, \gamma) = \begin{bmatrix} \{Q - \widetilde{H}(X, \gamma)\} r_2(X) \\ \{Y - M(W, \alpha)\}\{Q - \widetilde{H}(X, \gamma)\} \end{bmatrix}. \tag{16}$$

Okui, Small, Tan, and Rubins (2012) take $r_1(X) = \partial \widetilde{G}(X, \beta)/\partial \beta$ and $r_2(X) = \partial \widetilde{H}(X, \gamma)/\partial \gamma$. If $\widetilde{G}(X, \beta)$ is correctly specified, then $E\{G(Z, \alpha, \beta)\} = 0$, while if $\widetilde{H}(X, \gamma)$ is correctly specified then $E\{H(Z, \alpha, \gamma)\} = 0$.

To get their doubly robust estimator, Okui, Small, Tan, and Rubins (2012) first specify $\widetilde{G}(X_i, \beta)$ and $\widetilde{H}(X_i, \gamma)$, then estimate $\hat{\gamma}$ by the moment:

$$E(Q|X_i) = \widetilde{H}(X_i, \gamma)$$

and then estimate α and β by minimizing a quadratic form of $\hat{B}(\alpha, \beta; \hat{\gamma})$, where

$$\hat{B}(\alpha, \beta; \hat{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \{Y_i - M(W_i, \alpha) - \widetilde{G}(X_i, \beta)\}\{Q_i - \widetilde{H}(X_i, \hat{\gamma})\} \\ \{Y_i - M(W_i, \alpha) - \widetilde{G}(X_i, \beta)\} r_1(X_i) \end{bmatrix}.$$

In place of the Okui, Small, Tan, and Rubins (2012) DR construction, we could estimate this model using the GDR estimator, equation (3), with G and H given by equations (15) and (16). To satisfy Assumption A3, $r_1(X)$ needs to include more than $S + J - K$ elements, and $r_2(X)$ needs to include more than J elements. So, e.g., we would want to include at least one more function of X

into $r_1(X)$ and $r_2(X)$, in addition to the functions $\partial\widetilde{G}(X,\beta)/\partial\beta$ and $\partial\widetilde{H}(X,\gamma)/\partial\gamma$ used by Okui, Small, Tan, and Rubins (2012).

3.3 Preference Parameter Estimates

One of the original applications of GMM estimation was the estimation of marginal utility parameters and of pricing kernels. See, e.g., Hansen and Singleton (1982) or Cochrane (2001). Consider a lifetime utility function of the form

$$u_\tau = E\left\{\sum_{t=0}^{T} b^t R_t U\left(C_t, X_t, \rho\right) \mid W_\tau\right\}$$

where u_τ is expected discounted lifetime utility in time period τ, b is the subjective rate of time preference, R_t is the time t gross returns from a traded asset, U is the single period utility function, C_t is observable consumption expenditures in time t, X_t is a vector of other observable covariates that affect utility, ρ is a vector of utility parameters, and W_τ is a vector of variables that are observable in time period τ. Maximization of this expected utility function under a life time budget constraining yields Euler equations of the form

$$E\left[\left\{bR_{t+1}\frac{U'\left(C_{t+1}, X_{t+1}, \rho\right)}{U'\left(C_t, X_t, \rho\right)} - 1\right\} \mid W_\tau\right] = 0 \tag{17}$$

where $U'\left(C_t, X_t, \rho\right)$ denotes $\partial U\left(C_t, X_t, \rho\right)/\partial C_t$. If the functional form of U' is known, then this equation provides moments that allow b and ρ to be estimated using GMM. But suppose we have two different possible specifications of U', and we do not know which specification is correct. Then our GDR estimator can be immediately applied, replacing the expression in the inner parentheses in equation (17) with $G(Z,\alpha,\beta)$ or $H(Z,\alpha,\gamma)$ to represent the two different specifications. Here α would represent parameters that are same in either specification, including the subjective rate of time preference b.

To give a specific example, a standard specification of utility is constant relative risk aversion with habit formation, where utility takes the form

$$U\left(C_t, X_t, \rho\right) = \frac{\left\{C_t - M\left(X_t\right)\right\}^{1-\rho} - 1}{1-\rho}$$

where X_t is a vector of lagged values of C_t, the parameter ρ is coefficient of risk aversion, and the function $M(X_t)$ is the habit function. See, e.g., Campbell and Cochrane (1999) or Chen and Ludvigson (2009). While this general functional form has widespread acceptance and use, there is considerable debate about the correct functional form for M, including whether X_t should include the current value of C_t or just lagged values. See, e.g., the debate about whether habits are internal or external as discussed in the above papers. Rather than take a stand on which habit model is correct, we could estimate the model by GDR.

To illustrate, suppose that with internal habits the function $M(X_t)$ would be given by $\widetilde{G}(X_t, \beta)$, where $\widetilde{G}$ is the internal habits functional form. Similarly, suppose with external habits $M(X_t)$ would be given by $\widetilde{H}(X_t, \gamma)$ where $\widetilde{H}$ is the external habits specification. Then, based on equation (17), we could define $G(Z, \alpha, \beta)$ and $H(Z, \alpha, \gamma)$ by

$$G(Z,\alpha,\beta) = \left[bR_{t+1} \frac{\left\{ C_{t+1} - \widetilde{G}(X_{t+1},\beta) \right\}^{-\rho}}{\left\{ C_t - \widetilde{G}(X_t,\beta) \right\}^{-\rho}} - 1 \right] W_\tau$$

$$H(Z,\alpha,\gamma) = \left[bR_{t+1} \frac{\left\{ C_{t+1} - \widetilde{H}(X_{t+1},\gamma) \right\}^{-\rho}}{\left\{ C_t - \widetilde{H}(X_t,\gamma) \right\}^{-\rho}} - 1 \right] W_\tau$$

In this example, we would have $\alpha = (b, \rho)$, and so would consistently estimate the discount rate b and the coefficient risk aversion ρ, no matter which habit model is correct. To help satisfy Assumption A3, we would generally want W_τ to have more elements than (α, β) and more than (α, γ).

3.4 Alternative Sets of Instruments

Consider a parametric model

$$Y = M(W, \alpha) + \epsilon$$

where Y is an outcome, W is vector of observed covariates, M is a known functional form, α is a vector of parameters to be estimated, and ϵ is an unobserved error term. Let R and Q denote two

different vectors of observed covariates that are candidate instruments. One may be unsure if either R or Q are valid instrument vectors are not, where validity is defined as being uncorrelated with ϵ.

We may then define model G by $E(\epsilon R) = 0$, so $G(Z, \alpha) = \{Y - M(W, \alpha)\} R$ and define model H by $E(\epsilon Q) = 0$, so $H(Z, \alpha) = \{Y - M(W, \alpha)\} Q$. With these definition we can then immediately apply the GDR estimator. In this case both β and γ are empty, but more generally, the variables R and Q could themselves be functions of covariates and of parameters β and γ, respectively.

A simple example that we consider in our Monte Carlo analysis is where $M(W, \alpha) = \alpha' W$, so the G model consists of the moments $E[(Y - \alpha' W) R] = 0$ and the H model is the moments $E[(Y - \alpha' W) Q] = 0$. The overientification condition of Assumption A3 is generally satisfied when Q and R each have more elements than W. Note this simple example has no β or γ parameters.

Next consider a richer example, which that includes some parameters other than α. This example, which we later empirically apply, is based on a model of Lewbel (2012). Suppose $Y = X'\alpha_x + S\alpha_s + \epsilon$, where X is a K-vector of observed exogenous covariates (including a constant term) satisfying $E(\epsilon X)$, and S is an endogenous or mismeasured covariate that is correlated with ϵ. The goal is estimation of the set of coefficients $\alpha = \{\alpha_x, \alpha_s\}$.

A standard instrumental variables based estimator for this model would consist of finding one or more covariates L such that $E(\epsilon L) = 0$. Then the set of instruments R would be defined by $R = \{X, L\}$. The equivalent GMM estimator would be based on the moments $E\{G(Z, \alpha)\} = 0$ where $G(Z, \alpha)$ is given by the stacked vectors

$$G(Z, \alpha) = \left\{ \begin{array}{c} X\left(Y - X'\alpha_x - S\alpha_s\right) \\ L\left(Y - X'\alpha_x - S\alpha_s\right) \end{array} \right\}. \tag{18}$$

A special case of this estimator (corresponding to a specific choice of the GMM weighting matrix) is standard linear two stage least squares estimation. The main difficulty with applying this estimator is that one must find one or more covariates L to serve as instruments. Defining L have more than one element results in more moments than parameters, helping to satisfy Assumption A3.

To illustrate, consider Engel curve estimation (see Lewbel 2008 for a short survey, and references

therein). Suppose Y is a consumer's expenditures on food, X is a vector of covariates that affect the consumer's tastes, and S is the consumer's total consumption expenditures (i.e., their total budget which must be allocated between food and non-food expenditures). Suppose, as is commonly the case, that S is observed with some measurement error. Then a possible and commonly employed set of instruments L consist of functions of the consumer's income. However, validity of functions of income as instruments for total consumption depends on an assumption of separability between the consumer's decisions on savings and their food expenditure decision, which may or may not be valid.

An alternative method of obtaining potential instruments is by exploiting functional form related assumptions. Lewbel (2012) shows that, under some conditions (including standard assumptions regarding classical measurement error), one may construct a set of potential instruments using the following procedure: Linearly regress S on X, and obtain the residuals from that regression. Define a vector of instruments P to be demeaned X (excluding the constant) times these residuals. This constructed vector P, along with X, then comprises the set of instruments used to construct a GMM estimator. This estimator is implemented in the STATA module IVREG2H by Baum and Schaffer (2012).

Let X_c denote the vector X with the constant removed. Algebraically, we can write the instruments obtained in this way as $R = \{X, P\}$ where $P = (X_c - \gamma_1)(S - X'\gamma_2)$, and where the vectors γ_1 and γ_2 in turn satisfy $E(X_c - \gamma_1) = 0$ and $E\{X(S - X'\gamma_2)\} = 0$. An efficient estimator based on this construction would be standard GMM using the moments $E\{H(Z, \alpha, \gamma)\} = 0$ where $H(Z, \alpha, \gamma)$ is a vector that consists of the stacked vectors

$$H(Z,\alpha,\gamma) = \left\{ \begin{array}{c} X_c - \gamma_1 \\ X(S - X'\gamma_2) \\ X\left(Y - X'\alpha_x - S\alpha_s\right) \\ (X_c - \gamma_1)(S - X'\gamma_2)\left(Y - X'\alpha_x - S\alpha_s\right) \end{array} \right\}. \tag{19}$$

This estimator will have more moments than parameters if X_c has more than one element. As

shown in Lewbel (2012), one set of conditions under which the instruments P are valid (yielding consistency of this estimator) is if the measurement error in S is classical and if a component of ϵ is homoscedastic. So this estimator does not require finding a covariate from outside the model like income to use an instrument, but still could be inconsistent if the required measurement error assumptions do not hold.

The moments given by $E\{G(Z,\alpha)\} = 0$ or $E\{H(Z,\alpha,\gamma)\} = 0$ correspond to two very different sets of identifying conditions. GDR estimation based on these moments therefore allows for consistent estimation of α if either one of these sets of conditions hold.

4 The GDR Estimator Asymptotics

In this section we first show consistency of our GDR estimator, and then derive its limiting distribution, showing it is root N consistent and asymptotically normal.

4.1 GDR Consistency

To show consistency of $\widehat{\alpha}$, we apply Theorem 2.1 in Newey and McFadden (1994), which provides a set of standard sufficient conditions for identification and consistency of extremum estimators.

Let $Q_0^g(\alpha,\beta) \equiv g_0(\alpha,\beta)'\Omega_g g_0(\alpha,\beta)$ and $Q_0^h(\alpha,\gamma) \equiv h_0(\alpha,\gamma)'\Omega_h h_0(\alpha,\gamma)$ for positive definite matrices Ω_g and Ω_h. We later discuss choices for Ω_g and Ω_h.

Assumption A4: Assume there exists a unique $\{\alpha_g,\beta_g\} \in \Theta_\alpha \times \Theta_\beta$ that minimizes $Q_0^g(\alpha,\beta)$, and there exists a unique $\{\alpha_h,\gamma_h\} \in \Theta_\alpha \times \Theta_\gamma$ that minimizes $Q_0^h(\alpha,\gamma)$.

Given assumptions A1 to A4, if the minimized value $Q_0^g(\alpha_g,\beta_g) = 0$, then G is a correct model and $\{\alpha_g,\beta_g\}$ will equal $\{\alpha_0,\beta_0\}$. Otherwise, if the minimized value $Q_0^g(\alpha_g,\beta_g) > 0$, then G is not a correct model, and in this case we can think of $\{\alpha_g,\beta_g\}$ as unique values that are pseudo-true, in the sense that they are the values that GMM estimation of model G will converge to if model G

is wrong. Assumption A4 requires that these pseudo-true values are unique. The same holds with $\{\alpha_h, \gamma_h\}$ in model H.

To satisfy the continuity and uniform convergence conditions of Theorem 2.1 in Newey and McFadden (1994), we make the following additional assumptions.

Assumption A5: $G(Z, \alpha, \beta)$ and $H(Z, \alpha, \gamma)$ are continuous at $\{\alpha, \beta\} \in \Theta_\alpha \times \Theta_\beta$ and $\{\alpha, \gamma\} \in \Theta_\alpha \times \Theta_\gamma$ respectively with probability one.

Assumption A6: $E[\sup_{\{\alpha,\beta\}\in\Theta_\alpha\times\Theta_\beta} ||G(Z, \alpha, \beta)||] < \infty$ and $E[\sup_{\{\alpha,\gamma\}\in\Theta_\alpha\times\Theta_\gamma} ||H(Z, \alpha, \gamma)||] < \infty$.

Together, Assumptions A1, A2, A3, A5, and A6, are the standard conditions that yield consistency of the GMM estimates $\{\widehat{\alpha}_g, \widehat{\beta}_g\}$ in equation (1) if model G is correctly specified, and similarly yield consistency of $\{\widehat{\alpha}_h, \widehat{\gamma}_h\}$ if model H is correctly specified. With the additional Assumption A4, these conditions ensure that these estimates all have finite probability limits whether the models are correct or not.

Theorem 1 : Suppose that z_i, $i = 1, 2, ...,$ are iid, $\widehat{\Omega}_g \to^P \Omega_g$, $\widehat{\Omega}_h \to^P \Omega_h$, Ω_g and Ω_h are positive definite, and Assumptions A1 to A6 hold. Assume that, if both $Q_0^g(\alpha_0, \beta_0) = 0$ and $Q_0^h(\alpha_0, \gamma_0) = 0$, then $\widehat{W}_g \to^P C$ where C is finite and nonzero. Then for $\widehat{\alpha}$ given by equation (3), $\widehat{\alpha} \to^P \alpha_0$.

Theorem 1 shows consistency of $\widehat{\alpha}$. In the next subsection, where we derive the limiting distribution of $\widehat{\alpha}$, we give a mild condition regarding model moments that suffices to make C be a constant that is finite and nonzero. We discuss choice of $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$ later, but note for now that these can be standard GMM weight matrix estimates.

Proof : By A1-A6, the four conditions that consist of uniqueness, compactness, continuity, and uniform convergence, of Theorem 2.1 of in Newey and McFadden (1994) hold for GMM_g and GMM_h. Therefor when either A2-1) or A2-2), or both hold, the corresponding GMM estimators are consistent by Theorem 2.1 in Newey and McFadden.

For simplicity, let $\hat{Q}^g \equiv \hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$, $\hat{Q}^h \equiv \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$, $Q_0^g \equiv Q_0^g(\alpha_g, \beta_g)$, and $Q_0^h \equiv Q_0^h(\alpha_h, \gamma_h)$ unless we specify it. Now we will consider consistency of $\widehat{\alpha}$ in three possible cases.

Case 1: Suppose A2-1) holds. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^P \{\alpha_0, \beta_0\}$ and $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^P \{\alpha_h, \gamma_h\}$. And $\hat{Q}^g \to^P Q_0^g = Q_0^g(\alpha_0, \beta_0) = 0$ and $\hat{Q}^h \to^P Q_0^h > 0$. By the continuity and uniform convergence of $\hat{Q}^g$ and $\hat{Q}^h$, $\widehat{W}_g$ in (T.1) converges to zero in probability, and thus the consistency of $\widehat{\alpha}$ is followed by that of $\widehat{\alpha}_g$.

Case 2: Suppose A2-2) holds. Then, $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^P \{\alpha_g, \beta_g\}$ and $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^P \{\alpha_0, \gamma_0\}$. And $\hat{Q}^g \to^P Q_0^g > 0$ and $\hat{Q}^h \to^P Q_0^h = Q_0^h(\alpha_0, \gamma_0) = 0$. By the continuity and uniform convergence of $\hat{Q}^g$ and $\hat{Q}^h$, $\widehat{W}_g$ in (T.1) converges to one in probability, and thus the consistency of $\widehat{\alpha}$ is followed by that of $\widehat{\alpha}_h$.

Case 3: Suppose both A2-1) and A2-2) hold. Then, $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^P \{\alpha_0, \beta_0\}$ and $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^P \{\alpha_0, \gamma_0\}$. And $\hat{Q}^g \to^P Q_0^g = Q_0^g(\alpha_0, \beta_0) = 0$ and $\hat{Q}^h \to^P Q_0^h = Q_0^h(\alpha_0, \gamma_0) = 0$. Since

$$\begin{aligned}
\widehat{\alpha} &= \widehat{W}_g \widehat{\alpha}_h + (1 - \widehat{W}_g)\widehat{\alpha}_g \\
&= \widehat{W}_g[\alpha_0 + (\widehat{\alpha}_h - \alpha_0)] + (1 - \widehat{W}_g)[\alpha_0 + (\widehat{\alpha}_g - \alpha_0)] \\
&= \alpha_0 + \widehat{W}_g(\widehat{\alpha}_h - \alpha_0) + (1 - \widehat{W}_g)(\widehat{\alpha}_g - \alpha_0) \qquad \text{(T.1)}
\end{aligned}$$

we get in this case that

$$\begin{aligned}
p\lim \widehat{\alpha} &= \alpha_0 + p\lim\{\widehat{W}_g(\widehat{\alpha}_h - \alpha_0)\} + p\lim\{(1 - \widehat{W}_g)(\widehat{\alpha}_g - \alpha_0)\} \\
&= \alpha_0 + C \times p\lim(\widehat{\alpha}_h - \alpha_0) + (1 - C) \times p\lim(\widehat{\alpha}_g - \alpha_0) \\
&= \alpha_0,
\end{aligned}$$

and the consistency of $\widehat{\alpha}$ is followed by that of $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$. Q.E.D.

4.2 Limiting Distribution

In this section, we present the probability limit and the asymptotic distribution of the proposed estimator $\widehat{\alpha}$.

Let $\theta_0^g \equiv \{\alpha_0, \beta_0\}$, $\theta_0^h \equiv \{\alpha_0, \gamma_0\}$, $\theta^g \equiv \{\alpha_g, \beta_g\}$, and $\theta^h \equiv \{\alpha_h, \gamma_h\}$. Let $\nabla_\theta g_0(\theta^g)$ and $\nabla_\theta h_0(\theta^h)$ denote $E\{\nabla_{\theta^g} G(Z, \alpha_g, \beta_g)\}$ and $E\{\nabla_{\theta^h} H(Z, \alpha_h, \gamma_h)\}$, respectively.

Assumption A7: $\{\alpha_g, \beta_g\}$ and $\{\alpha_h, \gamma_h\}$ lie in the interior of $\Theta_\alpha \times \Theta_\beta$ and $\Theta_\alpha \times \Theta_\gamma$.

Assumption A8: $G(Z, \alpha, \beta)$ and $H(Z, \alpha, \gamma)$ are continuously differentiable in neighborhood $\aleph^g$ of θ^g and $\aleph^h$ of θ^h respectively with probability approaching one.

Assumption A9: $E[||G(Z, \alpha, \beta)||^2] < \infty$ and $E[||H(Z, \alpha, \gamma)||^2] < \infty$.

Assumption A10: $E[\sup_{\{\alpha,\beta\}\in\aleph^g} ||\nabla_{\theta^g} G(Z, \alpha, \beta)||] < \infty$ and $E[\sup_{\{\alpha,\gamma\}\in\aleph^h} ||\nabla_{\theta^h} H(Z, \alpha, \gamma)||] < \infty$.

Assumption A11: $\{\nabla_\theta g_0(\theta_0^g)\}'\Omega_g\{\nabla_\theta g_0(\theta_0^g)\}$ and $\{\nabla_\theta h_0(\theta_0^h)\}'\Omega_h\{\nabla_\theta h_0(\theta_0^h)\}$ are non singular.

In a small abuse of notation, let $\alpha = \left(\alpha_k, \alpha_{(k)}\right)$ where α_k is the k_{th} element of α and $\alpha_{(k)}$ is the remaining elements of α. Also, let $\nabla_{\alpha_k}\nabla_{\alpha_k}$ denote second-partial derivative wrt α_k.

Assumption A12: Let $\aleph^{\alpha_k}$ be a small open interval including α_k with endpoint α_{k0}. For any α_k in a small open neighborhood $\aleph^{\alpha_k}$, the following two terms are non-zero

$$\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k, \alpha_{(k)0}, \beta_0) = \{\nabla_{\alpha_k}\nabla_{\alpha_k} g_0(\alpha_k, \alpha_{(k)0}, \beta_0)\}'\Omega_g g_0(\alpha_k, \alpha_{(k)0}, \beta_0) + A^g(\alpha_k),$$
$$\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^h(\alpha_k, \alpha_{(k)0}, \gamma_0) = \{\nabla_{\alpha_k}\nabla_{\alpha_k} h_0(\alpha_k, \alpha_{(k)0}, \gamma_0)\}'\Omega_h h_0(\alpha_k, \alpha_{(k)0}, \gamma_0) + A^h(\alpha_k),$$

where $A^g(\alpha_k) \equiv \{\nabla_{\alpha_k} g_0(\alpha_k, \alpha_{(k)0}, \beta_0)\}'\Omega_g\{\nabla_{\alpha_k} g_0(\alpha_k, \alpha_{(k)0}, \beta_0)\}$ and $A^h(\alpha_k) \equiv \{\nabla_{\alpha_k} h_0(\alpha_k, \alpha_{(k)0}, \gamma_0)\}'\Omega_h\{\nabla_{\alpha_k} h_0$

Assumptions A7 to A10 are standard regularity conditions of the type used in, e.g., Newey and McFadden (1994). Assumption A11-12, which is roughly analogous to ruling out perfect collinearity issues in linear regression, is used to ensure that our derived influence functions are well behaved. Assumption 12 is new, but not a strong assumption because at $\theta_0^g = \{\alpha_{k0}, \alpha_{(k)0}, \beta_0\}$, $\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\theta_0^g) = A^g(\alpha_{k0})$ and $A^g(\alpha_{k0})$ is an entry of diagonal of $\{\nabla_\theta g_0(\theta_0^g)\}'\Omega_g\{\nabla_\theta g_0(\theta_0^g)\}$. Under Assumption 11, $A^g(\alpha_{k0})$ is non-zero and positive.

Let $\widehat{\eta}_i^g$ and $\widehat{\eta}_i^h$ be consistent estimators of the standard GMM influence functions associated with our first and second stage GMM estimators $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$ (these formulas are given in the Appendix).

Theorem 2: Suppose $\widehat{\Omega}_g \to^P \Omega_g$, $\widehat{\Omega}_h \to^P \Omega_h$, $\widehat{W}_g$ is given by equation (2), and assumptions A1 to A12. Then there exists a matrix $\widetilde{V}$ such that 1)

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}),$$

and 2)

$$\frac{1}{N}\sum_i \widehat{\eta}_i \widehat{\eta}_i' \to^p \widetilde{V} \tag{20}$$

$$\text{where } \widehat{\eta}_i \equiv \widehat{W}_g \widehat{\eta}_i^h + (1 - \widehat{W}_g)\widehat{\eta}_i^g.$$

The first part of Theorem 2 states that the GDR estimator $\widehat{\alpha}$ is root N consistent and asymptotically normal, while the second part gives a consistent estimator for the limiting variance of $\widehat{\alpha}$. In the Appendix we define variance matrices $\widetilde{V}^g$, $\widetilde{V}^h$, and $\widetilde{V}^{gh}$. The matrix $\widetilde{V}$ equals $\widetilde{V}^g$ if model G is correctly specified and model H is not, while $\widetilde{V}$ equals $\widetilde{V}^h$ if H is correctly specified and G is not, and let $\widetilde{V}$ equals $\widetilde{V}^{gh}$ when both G and H are correctly specified. Importantly, the consistent estimator of $\widetilde{V}$ given in equation (20) does not require knowing which of the models G or H is correct.

The proof of Theorem 2 is given in the Appendix. The basic structure of the proof follows Newey and McFadden (1994) for multistep parametric estimators, but a few nonstandard complications arise. One issue is that, if model H is wrong, then we cannot consistently estimate the influence function η_i^h for model H. However, in the limiting variance formula for $\widehat{\alpha}$, the function η_i^h is multiplied by $\widehat{W}_g$. If model H is wrong, then model G must be right, which makes $\hat{Q}^g$ asymptotically zero and $\hat{Q}^h$ asymptotically nonzero, meaning that $\widehat{W}_g$ is asymptotically zero. We therefore only need an estimate for η_i^h that is consistent when model H is right, and that estimate is the standard GMM influence function $\widehat{\eta}_i^h$. The same applies to the influence function $\widehat{\eta}_i^g$ for model G if model G is wrong.

Another nonstandard complication arises when models G and H are both correct. When both are correct, both the numerator and denominator of $\widehat{W}_g$ goes to zero. To resolve this issue, define the function $\widehat{W}_g(\alpha)$ by

$$\widehat{W}_g(\alpha) \equiv \frac{\hat{Q}^g(\alpha, \widehat{\beta}_g)}{\hat{Q}^g(\alpha, \widehat{\beta}_g) + \hat{Q}^h(\alpha, \widehat{\gamma}_h)}.$$

When both G and H are right, we can use L'hopital's rule to give conditions ensuring that $\lim_{\alpha \to \alpha_0} \widehat{W}_g(\alpha)$ is finite and non-zero. To show that the estimate $\widehat{W}_g(\alpha) \to^p W_g(\alpha_0)$, instead of taking the usual expansion around α_0, we expand $\widehat{W}_g$ around a value α that is in a neighborhood of α_0, and then take the limit as α goes to α_0. Details are in the Appendix.

4.3 Efficiency and Numerical Issues

For asymptotic efficiency of α, we could consider estimates of the weighting matrices $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$ that minimize the variance given by equation (20). However, the standard GMM estimates of $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$ should be close to efficient, because the GDR objective function is asymptotically dominated by the GMM objective function of the correct model. Also, the scaling or units of moments affect the relative magnitudes of the two GMM objective functions in finite samples. It is therefore numerically desirable in finite samples to have $\hat{Q}^g$ and $\hat{Q}^h$ be comparable. The standard GMM estimates of $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$ make $\hat{Q}^g$ and $\hat{Q}^h$ comparable, specifically, both will asymptotically have central (if right) or noncentral (if wrong) chi-squared distributions. We therefore find it desirable to use the standard GMM estimates of $\widehat{\Omega}_g$ and $\widehat{\Omega}_h$, even if that potentially sacrifices a small amount of asymptotic efficiency.

5 Simulation Results

Here we do some Monte Carlo analyses to investigate small sample properties of our estimator. Our design is two competing sets of instruments as in section 3.4. For each simulation we draw n independent, identically distributed observations of the random vector $(Y, W, R_1, R_2, Q_1, Q_2)$. The

model is assumed to be

$$Y = \alpha_0 + \alpha_1 W + \epsilon.$$

The goal is estimation of $\alpha = (\alpha_0, \alpha_1) = (1, 1)$. The regressor W is endogenous (correlated with ϵ), so estimation will be by instrumental variables. Model G assumes $E(\epsilon) = E(\epsilon R_1) = E(\epsilon R_2) = 0$, meaning that $R = (1, R_1, R_2)'$ is vector of valid instruments for instrumental variables estimation. Model H assumes $E(\epsilon) = E(\epsilon Q_1) = E(\epsilon Q_2) = 0$, meaning that $Q = (1, Q_1, Q_2)'$ is a vector of valid instruments. So $Z = (Y, W, R, Q)$, $G(Z, \alpha) = (Y - \alpha_0 - \alpha_1 W) R$, and $H(Z, \alpha) = (Y - \alpha_0 - \alpha_1 W) Q$. In this application there is no β or γ.

The simulation design is that we let $W = 1 + 2(R_1 + R_2) + Q_1 + Q_2 + \epsilon$. Having the 2 in this equation means that the G model has stronger instruments than the H model. We Let $R_1, R_2, Q_1, Q_2, \epsilon$ be standard normals, with $corr(R_j, \epsilon) = \rho_R$, $corr(Q_j, \epsilon) = \rho_Q$, and all the other correlations among these normals are zero. The correlations ρ_R and ρ_Q are each either .5 or zero.

We consider three different simulation designs. The first takes $\rho_R = 0$ and $\rho_Q = .5$, which makes model G be right (i.e., R are valid instruments) and model H be wrong (i.e., Q are not valid instruments, because they correlate with the model error ϵ). The second takes $\rho_R = .5$ and $\rho_Q = 0$, which makes model H be right and model G wrong. The third takes $\rho_R = \rho_Q = 0$, which makes both models be right (both sets of instruments are valid).

In Table 1 we report four estimates of α_0 and α_1 for each simulation. First is GMM based on the model G moments (which is only consistent if model G is right), second is GMM based on the H moments (which is only consistent if model G is right), third is GMM based on both sets of moments (which is consistent, and more efficient than either the first or second set of estimates, only if both models are right), and fourth is our GDR estimator, which is consistent if either set of moments is valid.

To be completed.

6 Empirical Application: Engel Curve Estimation

Here we consider the example discussed in section 3.4. Y is a consumer's expenditures on food, X is a vector of covariates that affect the consumer's tastes, and S is the consumer's total consumption expenditures (i.e., their total budget which must be allocated between food and non-food expenditures). The budget S is observed with some mismeasurement error. L consists of two or more functions of the consumer's income.

To be completed.

7 Extensions: Multiple Robustness and Alternative GDR's

Our GDR can be readily extended to obtain triply and higher multiply robust estimators. Suppose we have a third model, called model F, with GMM objective function $\hat{Q}^f(\alpha, \delta)$. The GMM estimator of model F is $\{\widehat{\alpha}_f, \widehat{\delta}_f\} = \arg\min_{\{\alpha,\delta\}\in\Theta_\alpha\times\Theta_\delta} \hat{Q}^f(\alpha, \delta)$. The objective function for triply robust estimation of α would then be the weighted average

$$\widehat{\alpha} = \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\widehat{\alpha}_f + \hat{Q}^f(\widehat{\alpha}_f, \widehat{\delta}_f)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\widehat{\alpha}_g + \hat{Q}^f(\widehat{\alpha}_f, \widehat{\delta}_f)\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\widehat{\alpha}_h}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) + \hat{Q}^f(\widehat{\alpha}_f, \widehat{\delta}_f)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) + \hat{Q}^f(\widehat{\alpha}_f, \widehat{\delta}_f)\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)} \tag{21}$$

So the weight on $\widehat{\alpha}_f$ is proportional to the product of objective functions for the other models, $\hat{Q}^g\hat{Q}^h$, and similarly for the weights on $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$.

The logic of this estimator is the same as for our GDR. For example, if model G is right and models F and H are wrong, then only $\widehat{\alpha}_g$ will get a nonzero weight asymptotically. Now suppose two but not all three models are right, e.g., suppose models G and H are right and F is wrong. Then all the weights in both the numerator and denominator of equation (21) go to zero. However, in this case we can divide the numerator and denominator by $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$. Both $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ and $\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ converge to zero, but if the probability limit of $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ is finite and nonzero, then the limiting weights on $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$ will be nonzero while the limiting weight on $\widehat{\alpha}_f$ will be zero, as desired. We required that this same $\hat{Q}^g/\hat{Q}^h$ ratio have a finite and nonzero probability limit to

obtain the limiting distribution of our GDR, and showed that this requirement holds under mild regularity conditions.

It is possible to construct alternative GDR's based on the idea of our proposed GDR. For example, one could replace $\hat{Q}^g$ and $\hat{Q}^h$ in equation (3) $\zeta\left(\hat{Q}^g\right)$ and $\zeta\left(\hat{Q}^h\right)$ for any strictly monotonically increasing function ζ having $\zeta(0) = 0$. Another alternative GDR would be the estimator

$$\widehat{\alpha} = \arg\min_{\alpha \in \Theta_\alpha} \hat{Q}^g(\alpha, \widehat{\beta}_g)\hat{Q}^h(\alpha, \widehat{\gamma}_h) \tag{22}$$

Consistency of this alternative GDR follows a similar logic to our original GDR. For ecample, if model G is right and H is not, then minimizing $\hat{Q}^g(\alpha, \widehat{\beta}_g)\hat{Q}^h(\alpha, \widehat{\gamma}_h)$ will be asymptotically equivalent to minimizing $\hat{Q}^g(\alpha, \widehat{\beta}_g)$ because $\hat{Q}^g$ will go to zero while $\hat{Q}^h$ cannot. A disadvantage of this alternative estimator is that it's more numerically complicated than our earlier GDR, because it entails an additional numerical search for $\widehat{\alpha}$ instead of just taking a weighted average of $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$. The extension of this alternative GDR to higher muliply robust estimators is immediate, e.g., the alternative general triply robust estimator would be

$$\widehat{\alpha} = \arg\min_{\alpha \in \Theta_\alpha} \hat{Q}^f(\alpha, \widehat{\delta}_f)\hat{Q}^g(\alpha, \widehat{\beta}_g)\hat{Q}^h(\alpha, \widehat{\gamma}_h) \tag{23}$$

A potential exercise for future work would be to compare these alternative GDR and muliply robust estimators to see if any are systematically more efficient or behave better in finite samples.

8 Conclusions

To be completed.

REFERENCES

Altonji, J., and Segal, L.M. (1996): "Small Sample Bias in GMM Estimation of Covariance Structures", *Journal of Economic and Business Statistics*, 14(3), 353-366.

Anderson, T.W., and Sawa, T. (1979): "Evaluation of the Distribution Function of the Two-Stage Least Squares Estimate", *Econometrica*, 47(1), 163-182.

Andrews, D. W. K. and B. Lu a(2001): "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models", *Journal of Econometrics*, 101(1), 123-164.

Babu, G.J. (1986): "A Note on Bootstrapping the Variance of Sample Quantile", *Annals of the Institute of Statistical Mathematics,* 38(A), 439-443

Bang, H., and Robins, J. (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models", *Biometrics*, 61(4), 962-973.

Baum, C., and Schaffer M. (2012): "IVREG2H: Stata Module to Perform Instrumental Variables Estimation Using Heteroskedasticity-based Instruments", *Statistical Software Components S457555*, Boston College Department of Economics, revised 18 Feb 2018.

Brown, B.W., and Newey, W.K. (1992): "Bootstrapping for GMM", Notes for seminar at Monash University.

Campbell, J., and Cochrane, J. (1999): "By Force of Habit: A Consumption?Based Explanation of Aggregate Stock Market Behavior", *Journal of Political Economy*, 107(2), 205-251.

Chen, X., and Ludvigson, S. (2009): "Land of Addicts? an Empirical Investigation of Habit-based Asset Pricing Models", *Journal of Applied Econometrics*, 24(7), 1057-1093.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. and Robins, J. M. (2018): "Locally Robust Semiparametric Estimation," Unpublished Manuscript.

Cochrane, J. (2001): "Long-Term Debt and Optimal Policy in the Fiscal Theory of the Price Level", *Econometrica*, 69(1), 69-116.

DiTraglia, F. (2016): "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM", *Journal of Econometrics*, 195(2), 187-208.

Funk, M., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M., and Davidian, M. (2011): "Doubly Robust Estimation of Causal Effects", *American Journal of Epidemiology*, 173(7), 761-7.

Hahn, J. (1996): "A Note on Bootstrapping Generalized Method of Moments Estimators", *Econometric Theory,*

12(1), 187-196.

Hall, P., and Horowitz, J. (1996): "Bootstrap Critical Values for Tests Based on Generalized Method of Moments", *Econometrica*, 64(4), 891-916.

Hansen, B. (2997): "Least Squares Model Averaging", *Econometrica*, 75, 1175-1189.

Hansen, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50(4), 1029-1054.

Hansen, L.P., Heaton, J., and Yaron, A. (1996): "Finite-Sample Properties of Some Alternative GMM Estimator", *Journal of Business and Economic Statistics,* 14(3), 262-280.

Hansen, L., and Singleton, K. (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, 50(5), 1269-1286.

Kuersteiner, G. and R. Okui. (2010): "Constructing Optimal Instruments by First-Stage Prediction Averaging.", *Econometrica*, 78(2), 697-718.

Lee, M. J., and S. Lee. (2018): "Double Robustness Without Weighting", *Statistics and Probability Letters, Forthcoming*

Lewbel, A. (2008): "Engel curves", entry for *The New Palgrave Dictionary of Economics*, 2nd Edition, MacMillan Press.

Lewbel, A. (2012): "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models", *Journal of Business and Economic Statistics*, 30(1), 67-80.

Liao, Z. (2013): "Adaptive GMM Shrinkage Estimation With Consistent Moment Selection", *Econometric Theory,* 29(5), 857-904.

Lunceford, J.K., and Davidian, M. (2004): "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: a Comparative Study", *Statistics in Medicine*, 23(19), 2937–2960.

Martins, L. F.., and V. J. Gabriel. (2014): "Linear Instrumental Variables Model Averaging Estimation", *SComputational Statistics and Data Analysis*, 71, 709-724.

Newey, W. K. and Mcfadden. D. (1994): "Chapter 36 Large Sample Estimation and Hypothesis Testing", in *Handbook of Econometrics*, 4, 2111-2245.

Okui, R., Small, D., Tan, Z., and Robins, J. (2012): "Doubly Robust Instrumental Variable Regression", *Statistica Sinica*, 22(1), 173-205.

Robins, J., Rotnitzky, A., and Van Der Laan, M. (2000): "On Profile Likelihood: Comment", *Journal of the American Statistical Association*, 95(450), 477-482.

Robins, J., Rotnitzky, A., and Zhao, L. (1994): "Estimation of Regression Coefficients When Some Regressors are not Always Observed", *Journal of the American Statistical Association*, 89(427), 846-866.

Rose, S., and Van der Laan, M. (2014): "A Double Robust Approach to Causal Effects in Case-Control Studies", *American Journal of Epidemiology*, 179(6), 663-669.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999): "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models", *Journal of the American Statistical Association*, 94(448), 1096-1120.

Słoczyński, T., and Wooldridge, J. (2018): "A General Double Robustness Result for Estimating Average Treatment Effects", *Econometric Theory*, 34(01), 112-133.

Sueishi, M. (2013): "Generalized Empirical Likelihood-Based Focused Information Criterion and Model Averaging", *Econometrics*, 1(2), 141-156.

Wooldridge, J. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems", *Journal of Econometrics*, 141(2), 1281-1301.

Appendix

Proof of Theorem 2:

Recall equation (T.1) and rewrite as follows

$$\widehat{\alpha} = \alpha_0 + \widehat{W}_g(\widehat{\alpha}_h - \alpha_0) + (1 - \widehat{W}_g)(\widehat{\alpha}_g - \alpha_0)$$
$$\Rightarrow \sqrt{N}(\widehat{\alpha} - \alpha_0) = \sqrt{N}(\widehat{\alpha}_g - \alpha_0)(1 - \widehat{W}_g) + \sqrt{N}(\widehat{\alpha}_h - \alpha_0)\widehat{W}_g \qquad (Inf)$$

$\widehat{W}_g$ can be seen as a weight function between $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$.

Now we will show the asymptotic normality of $\widehat{\alpha}$ and the form of $\widetilde{V}$ depending on which model is correctly specified.

Case 1) Suppose G is correctly specified, but H is not. If $p\lim \widehat{W}_g = 0$, $\sqrt{N}(\widehat{\alpha} - \alpha_0)$ asymptotically follows the same asymptotic distribution of $\sqrt{N}(\widehat{\alpha}_g - \alpha_0)$ which is nothing but that of GMM_g. By following the same argument as given in Theorem 3.4 of Newey and McFadden, under A7-10 and the consistency of GMM_g, $(\widehat{\alpha}_g, \widehat{\beta}_g) \to^P (\alpha_0, \beta_0)$ and $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) \to^P Q^g(\alpha_0, \beta_0) = 0$, while $p\lim(\widehat{\alpha}_g, \widehat{\gamma}_h)$ is not (α_0, γ_0) but (α_h, γ_h) (the pseudo-true value by A4) so that we have $\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) \to^p Q^h(\alpha_h, \gamma_h) \neq 0$, which is positive. Thus, the $p\lim$ of denominator of $\widehat{W}_g$ is positive while that of numerator is zero, so that $p\lim \widehat{W}_g = 0$. By the central limit theorem, $\frac{1}{\sqrt{N}}\sum_i G(Z_i, \alpha_0, \beta_0) \to^d N(0, \Sigma_g)$ where $\Sigma_g = E\{G(Z, \alpha_0, \beta_0)G(Z, \alpha_0, \beta_0)'\}$. Along with $\widehat{g}(\widehat{\alpha}, \widehat{\beta}_g) \to^p g_0(\theta_0^g) = 0$ and $\nabla_\alpha \widehat{g}(\widehat{\alpha}, \widehat{\beta}_g) \to^p \nabla_\alpha g_0(\theta_0^g)$, we can establish asymptotic normality of $\sqrt{N}(\widehat{\alpha}_g - \alpha_0)$. Therefore, by the asymptotic normality of $\widehat{\alpha}_g$, $\widehat{W}_g \to^p 0$, and the continuous mapping theorem,

$$\sqrt{N}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^g)$$

and

$$\frac{1}{N}\sum_i \widehat{\eta}_i^g \widehat{\eta}_i^{g\prime} \to^p \widetilde{V}^g \equiv E(\eta^g \eta^{g\prime})$$

by A11, where

$$\frac{1}{\sqrt{N}}\sum_i \widehat{\eta}_i^g = \sqrt{N}(\widehat{\alpha}_g - \alpha_0).$$

$\widehat{\eta}_i^g$ is the influence function of the first-stage estimate $\widehat{\alpha}_g$ (The formulae is given in the Appendix), making $\widetilde{V}^g = \widetilde{V}$ be the asymptotic variance of $\widehat{\alpha}_g$.

Case 2) Suppose H is correctly specified, but G is not. Then the same argument as Case 1 applies, replacing $\widehat{W}_g$ with $1 - \widehat{W}_g$, and switching the roles of β and γ, and switching the roles of g and h.

Case 3) Suppose G and H are both correctly specified. If $p\lim \widehat{W}_g$ is nonzero and finite, $\sqrt{N}(\widehat{\alpha} - \alpha_0)$ asymptotically follows a weighted average of the asymptotic distributions of $\sqrt{N}(\widehat{\alpha}_g - \alpha_0)$ and $\sqrt{N}(\widehat{\alpha}_h - \alpha_0)$. Now we will establish nonzero and finite $p\lim \widehat{W}_g$ in several steps when G and H are both correct. Before proceeding, note that we can have

$$p\lim \widehat{W}_g = p\lim \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)} = p\lim \frac{Q_0^g(\widehat{\alpha}_g, \widehat{\beta}_g)}{Q_0^g(\widehat{\alpha}_g, \widehat{\beta}_g) + Q_0^h(\widehat{\alpha}_h, \widehat{\gamma}_h)}$$

by the uniform convergence $\sup_{\{\alpha,\beta\}\in\Theta_\alpha\times\Theta_\beta} |\hat{Q}^g(\alpha,\beta) - Q_0^g(\alpha,\beta)| \to^p 0$ and $\sup_{\{\alpha,\gamma\}\in\Theta_\alpha\times\Theta_\gamma} |\hat{Q}^h(\alpha,\gamma) - Q_0^h(\alpha,\gamma)| \to^p 0$ given in Theorem 2.6 of Newey and McFadden. By A8 and Cauchy's mean value theorem, there exists $\overline{\beta}_g$ between $\widehat{\beta}_g$ and β_0 such that

$$Q_0^g(\widehat{\alpha}_g, \widehat{\beta}_g) = Q_0^g(\widehat{\alpha}_g, \beta_0) + g_0(\widehat{\alpha}_g, \overline{\beta}_g)' \Omega_g \{\nabla_\beta g_0(\widehat{\alpha}_g, \overline{\beta}_g)\}(\widehat{\beta}_g - \beta_0).$$

Because $\widehat{\beta}_g \to^p \beta_0$, the following equality holds,

$$p\lim Q_0^g(\widehat{\alpha}_g, \widehat{\beta}_g) = p\lim Q_0^g(\widehat{\alpha}_g, \beta_0).$$

Analogously, because $\widehat{\gamma}_h \to^p \gamma_0$ we obtain

$$p\lim Q_0^h(\widehat{\alpha}_h, \widehat{\gamma}_h) = p\lim Q_0^h(\widehat{\alpha}_h, \gamma_0).$$

Plugging these into $p\lim \widehat{W}_g$ we get

$$p\lim \widehat{W}_g = p\lim \frac{Q_0^g(\widehat{\alpha}_g, \beta_0)}{Q_0^g(\widehat{\alpha}_g, \beta_0) + Q_0^h(\widehat{\alpha}_h, \gamma_0)}.$$

Next, expand $\widehat{\alpha}_h$ in $Q_0^h(\widehat{\alpha}_h, \gamma_0)$ around $\widehat{\alpha}_g$, and substitute it into $p\lim \widehat{W}_g$ to get

$$p\lim \widehat{W}_g = p\lim \frac{Q_0^g(\widehat{\alpha}_g, \beta_0)}{Q_0^g(\widehat{\alpha}_g, \beta_0) + Q_0^h(\widehat{\alpha}_g, \gamma_0) + \nabla_\alpha Q_0^h(\alpha^+, \gamma_0)(\widehat{\alpha}_h - \widehat{\alpha}_g)}$$

where α^+, from the mean value theorem, ranges between $\widehat{\alpha}_h$ and $\widehat{\alpha}_g$. Note that

$$p\lim \nabla_\alpha Q_0^h(\alpha_h^*, \gamma_0)(\widehat{\alpha}_h - \widehat{\alpha}_g) = p\lim \nabla_\alpha Q_0^h(\alpha_h^*, \gamma_0)\{(\widehat{\alpha}_h - \alpha_0) - (\widehat{\alpha}_g - \alpha_0)\}$$
$$= 0$$

when both G and H are correctly specified.

Using the notation $\alpha = \{\alpha_k, \alpha_{(k)}\}$ defined in section 4.2, let

$$W_g(\alpha_k) \equiv \frac{Q_0^g(\alpha_k, \alpha_{(k)0}, \beta_0)}{Q_0^g(\alpha_k, \alpha_{(k)0}, \beta_0) + Q_0^h(\alpha_k, \alpha_{(k)0}, \gamma_0) + \nabla_{\alpha_k} Q_0^h(\alpha_k^+, \alpha_{(k)0}, \gamma_0)(\widehat{\alpha}_{hk} - \alpha_k)}$$

where α_k^+ and $\widehat{\alpha}_{hk}$ are the corresponding k_{th} elements of α^+ and $\widehat{\alpha}_h$, when α^+ is the mean value between α and $\widehat{\alpha}_h$. Without loss of generality we will consider α_k in $p\lim \widehat{W}_g$. If we can show that $p\lim W_g(\widehat{\alpha}_k)$ is nonzero and finite for any element $\widehat{\alpha}_k$, then $p\lim \widehat{W}_g$ is so. For simplicity, we will omit $(\alpha_{(k)0}, \beta_0)$ and $(\alpha_{(k)0}, \gamma_0)$ in $Q_0^g(\alpha_k, \alpha_{(k)0}, \beta_0)$ and $Q_0^h(\alpha_k, \alpha_{(k)0}, \gamma_0)$, respectively, and use the notation $Q_0^g(\alpha_k)$ and $Q_0^h(\alpha_k)$ unless otherwise specified.

To show nonzero and finite $p\lim W_g(\widehat{\alpha}_k)$, we will borrow the idea used in the proof of L'Hopital's rule and the squeeze theorem. Let $\aleph^{\alpha_k}$ be a small open interval including $\widehat{\alpha}_k$ with endpoint α_{k0}. And note that

$$Q_0^g(\alpha_{k0}) = 0, \quad Q_0^h(\alpha_{k0}) = 0,$$
$$\nabla_{\alpha_k} Q_0^g(\alpha_{k0}) = \{\nabla_{\alpha_k} g_0(\alpha_{k0})\}' \Omega_g g_0(\alpha_{k0}) = 0,$$
$$\nabla_{\alpha_k} Q_0^h(\alpha_{k0}) = \{\nabla_{\alpha_k} h_0(\alpha_{k0})\}' \Omega_h h_0(\alpha_{k0}) = 0,$$

but for $\alpha_k \neq \alpha_{k0}$,

$$Q_0^g(\alpha_k) \neq 0, \quad Q_0^h(\alpha_k) \neq 0, \quad \nabla_{\alpha_k} Q_0^g(\alpha_k) \neq 0, \text{ and } \nabla_{\alpha_k} Q_0^h(\alpha_k) \neq 0.$$

By A8 and Cauchy's mean value theorem, there exists $\overline{\alpha}_k$ between $\widehat{\alpha}_k$ and α_{k0} such that

$$Q_0^g(\widehat{\alpha}_k) = Q_0^g(\alpha_{k0}) + \nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k) = \nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k),$$
$$Q_0^h(\widehat{\alpha}_k) = Q_0^h(\alpha_{k0}) + \nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k) = \nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k),$$

$$\Longrightarrow W_g\left(\widehat{\alpha}_k\right) = \frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k) + \nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k) + D(\overline{\alpha}_k)}.$$

where $\overline{D}(\overline{\alpha}_k) \equiv \nabla_{\alpha_k} Q_0^h(\alpha_k^+)\{(\widehat{\alpha}_{hk} - \overline{\alpha}_k) - (\widehat{\alpha}_k - \overline{\alpha}_k)\}$.

Let $\nabla_{\alpha_k}\nabla_{\alpha_k}$ denote second-partial derivative wrt α_k. For $\overline{\alpha}_k$ between $\widehat{\alpha}_k$ and α_{k0}, define

$$l(\overline{\alpha}_k) \equiv \inf \frac{\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k^*)}{\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k^*) + \nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^h(\alpha_k^*) + D^*(\alpha_k^*)},$$
$$L(\overline{\alpha}_k) \equiv \sup \frac{\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k^*)}{\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k^*) + \nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^h(\alpha_k^*) + D^*(\alpha_k^*)},$$
$$\text{where } D^*(\alpha_k^*) \equiv \nabla_{\alpha_k} Q_0^h(\alpha_k^+)\{(\widehat{\alpha}_{hk} - \alpha_k^*) - (\overline{\alpha}_k - \alpha_k^*) - (\widehat{\alpha}_k - \alpha_k^*) + (\overline{\alpha}_k - \alpha_k^*)\}$$

as α_k^* ranges over all values between $\overline{\alpha}_k$ and α_{k0}. Note that

$$\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^h(\alpha_k) = \{\nabla_{\alpha_k}\nabla_{\alpha_k} h_0(\alpha_k)\}'\Omega_h h_0(\alpha_k) + A^h(\alpha_k)$$

where $A^h(\alpha_k) \equiv \{\nabla_{\alpha_k} h_0(\alpha_k)\}'\Omega_h\{\nabla_{\alpha_k} h_0(\alpha_k)\}$. By A12, for any α_k in $\aleph^{\alpha_k}$, $\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^h(\alpha_k)$ is nonzero, and thus the denominator of $l(\overline{\alpha}_k)$ and $L(\overline{\alpha}_k)$ are well defined.

By A8 and Cauchy's mean value theorem, for any two distinct points $\overline{\alpha}_k$ and $\widetilde{\alpha}_k$ in $\aleph^{\alpha_k}$ there exists α_k^* between $\overline{\alpha}_k$ and $\widetilde{\alpha}_k$ such that

$$\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k) - \nabla_{\alpha_k} Q_0^g(\widetilde{\alpha}_k) = \nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k^*),$$
$$\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k) - \nabla_{\alpha_k} Q_0^h(\widetilde{\alpha}_k) = \nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^h(\alpha_k^*).$$

Therefore,

$$l(\overline{\alpha}_k) \leq \frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k) - \nabla_{\alpha_k} Q_0^g(\widetilde{\alpha}_k)}{\{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k) - \nabla_{\alpha_k} Q_0^g(\widetilde{\alpha}_k)\} + \{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k) - \nabla_{\alpha_k} Q_0^h(\widetilde{\alpha}_k)\} + \overline{D}(\overline{\alpha}_k)} \leq L(\overline{\alpha}_k)$$

for all choices of distinct $\overline{\alpha}_k$ and $\widetilde{\alpha}_k$ in the interval $\aleph^{\alpha_k}$. For any $\overline{\alpha}_k$ between $\widehat{\alpha}_k$ and α_{k0}, and $\widetilde{\alpha}_k$ between $\overline{\alpha}_k$ and α_{k0},

$$l(\overline{\alpha}_k) \leq \frac{\frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} - \frac{\nabla_{\alpha_k} Q_0^g(\widetilde{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)}}{\frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} - \frac{\nabla_{\alpha_k} Q_0^g(\widetilde{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} + 1 - \frac{\nabla_{\alpha_k} Q_0^h(\widetilde{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} + \frac{\overline{D}(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)}} \leq L(\overline{\alpha}_k).$$

Because this holds for any $\widetilde{\alpha}_k$ between $\overline{\alpha}_k$ and α_{k0}, also it holds in the limit as $\widetilde{\alpha}_k \to \alpha_{k0}$. Then we have

$$(\overline{\alpha}_k) \leq \frac{\lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} - \lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\nabla_{\alpha_k} Q_0^g(\widetilde{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)}}{\lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} - \lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\nabla_{\alpha_k} Q_0^g(\widetilde{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} + 1 - \lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\nabla_{\alpha_k} Q_0^h(\widetilde{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} + \lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\overline{D}(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)}} \leq L(\overline{\alpha}_k),$$

and because $\lim_{\widetilde{\alpha}_k \to \alpha_{k0}} Q_0^g(\widetilde{\alpha}_k) = \lim_{\widetilde{\alpha}_k \to \alpha_{k0}} Q_0^h(\widetilde{\alpha}_k) = 0$,

$$\begin{aligned}
&\Longrightarrow l(\overline{\alpha}_k) \leq \frac{\lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)}}{\lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)} + 1 + \lim_{\widetilde{\alpha}_k \to \alpha_{k0}} \frac{\overline{D}(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k)}} \leq L(\overline{\alpha}_k) \\
&\Longrightarrow l(\overline{\alpha}_k) \leq \frac{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k)}{\nabla_{\alpha_k} Q_0^g(\overline{\alpha}_k) + \nabla_{\alpha_k} Q_0^h(\overline{\alpha}_k) + \overline{D}(\overline{\alpha}_k)} \leq L(\overline{\alpha}_k).
\end{aligned}$$

Thus, we have

$$l(\overline{\alpha}_k) \leq W_g\left(\widehat{\alpha}_k\right) \leq L(\overline{\alpha}_k),$$

and it holds for any $\overline{\alpha}_k$ in $\aleph^{\alpha_k}$.

Under A7-10, A12, and the consistency of $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$,

$$\begin{aligned}
p\lim l(\overline{\alpha}_k) &= p\lim\inf \frac{\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k^*)}{\nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^g(\alpha_k^*) + \nabla_{\alpha_k}\nabla_{\alpha_k} Q_0^h(\alpha_k^*) + D^*(\alpha_k^*)} \\
&= p\lim\inf \frac{\{\nabla_{\alpha_k}\nabla_{\alpha_k} g_0(\alpha_k^*)\}'\Omega_g g_0(\alpha_k^*) + A^g(\alpha_k^*)}{\{\nabla_{\alpha_k}\nabla_{\alpha_k} g_0(\alpha_k^*)\}'\Omega_g g_0(\alpha_k^*) + \{\nabla_{\alpha_k}\nabla_{\alpha_k} h_0(\alpha_k^*)\}'\Omega_h h_0(\alpha_k^*) + A^g(\alpha_k^*) + A^h(\alpha_k^*) + D^*(\alpha_k^*)} \\
&= \frac{A^g(\alpha_{k0})}{A^g(\alpha_{k0}) + A^h(\alpha_{k0})},
\end{aligned}$$

because

$$\begin{aligned}
p\lim D^*(\alpha_k^*) &= p\lim \nabla_{\alpha_k} Q_0^h(\alpha_k^+)\{(\widehat{\alpha}_{hk} - \alpha_k^*) - (\overline{\alpha}_k - \alpha_k^*) - (\widehat{\alpha}_k - \alpha_k^*) + (\overline{\alpha}_k - \alpha_k^*)\} \\
&= 0.
\end{aligned}$$

By A11, $\{\nabla_{\theta^g} g_0(\theta_0^g)\}'\Omega_g\{\nabla_{\theta^g} g_0(\theta_0^g)\}$ is positive definite, and $A^g(\alpha_{k0})$ is an entry of diagonal of $\{\nabla_{\theta^g} g_0(\theta_0^g)\}'\Omega_g\{\nabla_{\theta^g} g_0(\theta_0^g)\}$. Thus, $A^g(\alpha_{k0})$ is real and nonzero positive, and the same argument holds for $A^h(\alpha_{k0})$. Analogously, the same argument holds for $p\lim L(\overline{\alpha}_k)$.

Let C denote $\frac{A^g(\alpha_{k0})}{A^g(\alpha_{k0})+A^h(\alpha_{k0})}$. $C = p\lim l(\overline{\alpha}_k) = p\lim L(\overline{\alpha}_k)$ can be written as

$$\lim_{n\to\infty} P(|l(\overline{\alpha}_k) - C| > \varepsilon) = 0 \quad \text{and} \quad \lim_{n\to\infty} P(|L(\overline{\alpha}_k) - C| > \varepsilon) = 0.$$

Since $l(\overline{\alpha}_k) \leq W_g\left(\widehat{\alpha}_k\right) \leq L(\overline{\alpha}_k)$ holds for any $\overline{\alpha}_k$ in $\aleph^{\alpha_k}$,

$$P(|W_g\left(\widehat{\alpha}_k\right) - C| > \varepsilon) \leq \left\{P(|l(\overline{\alpha}_k) - C| > \varepsilon) + P(|L(\overline{\alpha}_k) - C| > \varepsilon)\right\}.$$

Hence,

$$\lim_{n\to\infty} P(|W_g(\widehat{\alpha}_k) - C| > \varepsilon) \leq \left\{ \lim_{n\to\infty} P(|l(\overline{\alpha}_k) - C| > \varepsilon) + \lim_{n\to\infty} P(|L(\overline{\alpha}_k) - C| > \varepsilon) \right\} = 0,$$
$$\Longrightarrow p\lim W_g(\widehat{\alpha}_k) = C.$$

Therefore, $p\lim W_g(\widehat{\alpha}_k)$ is nonzero and finite, so $p\lim \widehat{W}_g$ is.

By following the same argument as given in Theorem 3.4 of Newey and McFadden, under A7-10 and the consistency of $\{\widehat{\alpha}_g, \widehat{\beta}_g\}$ along with $\{\widehat{\alpha}_h, \widehat{\gamma}_h\}$, we can establish the asymptotic normality of $\sqrt{N}(\widehat{\alpha}_g - \alpha_0)$ and $\sqrt{N}(\widehat{\alpha}_h - \alpha_0)$. Therefore, by the nonzero and finite $p\lim \widehat{W}_g$ and the asymptotic normality of $\sqrt{N}(\widehat{\alpha}_g - \alpha_0)$ and $\sqrt{N}(\widehat{\alpha}_h - \alpha_0)$

$$\begin{aligned}
\sqrt{N}(\widehat{\alpha} - \alpha_0) &= (1 - \widehat{W}_g)\sqrt{N}(\widehat{\alpha}_g - \alpha_0) + \widehat{W}_g\sqrt{N}(\widehat{\alpha}_h - \alpha_0) \\
&= (1 - \widehat{W}_g)\frac{1}{\sqrt{N}}\sum_i \widehat{\eta}_i^g + \widehat{W}_g\frac{1}{\sqrt{N}}\sum_i \widehat{\eta}_i^h \\
&= \frac{1}{\sqrt{N}}\sum_i \{(1 - \widehat{W}_g)\widehat{\eta}_i^g + \widehat{W}_g\widehat{\eta}_i^h\} = \frac{1}{\sqrt{N}}\sum_i \widehat{\eta}_i \\
&\to^d N(0, \widetilde{V} = \widetilde{V}^{gh}),
\end{aligned}$$

and

$$\frac{1}{N}\sum_i \widehat{\eta}_i\widehat{\eta}_i' \to^p \widetilde{V}^{gh} \equiv E(\eta\eta'),$$

where $\eta_i = (1 - C)\eta_i^g + C\eta_i^h$,

η_i^g and η_i^h are $p\lim \widehat{\eta}_i^g$ and $p\lim \widehat{\eta}_i^h$, respectively,

by A11-12. $\widehat{\eta}_i$ can be seen as a weighted sum of the influence function of the first-stage estimate $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$. Q.E.D.

Derivation of the form of $\widehat{\eta}_i^g$ and $\widehat{\eta}_i^h$:

To find the influence functions $\widehat{\eta}_i^g$, let $\widehat{\theta}^g$ denote the first-stage estimator

$$\widehat{\theta}^g \equiv (\widehat{\alpha}_g, \widehat{\beta}_g) = \arg \min_{\{\alpha,\beta\}\in\Theta_\alpha\times\Theta_\beta} \hat{Q}^g(\alpha, \beta) = \widehat{g}(\alpha, \beta)\widehat{\Omega}_g\widehat{g}(\alpha, \beta).$$

By A7-A9 with probability approaching one the following first-order conditions in the first-stage for $\widehat{\theta}^g$ are satisfied

$$FD^g_\alpha = \frac{\partial \hat{Q}^g(\widehat{\theta}^g)}{\partial \alpha} = \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \widehat{g}(\widehat{\theta}^g) = 0,$$
$$FD^g_\beta = \frac{\partial \hat{Q}^g(\widehat{\theta}^g)}{\partial \beta} = \{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \widehat{g}(\widehat{\theta}^g) = 0.$$

Expending $\widehat{g}$ around the unique minimizer $\theta^g \equiv \{\alpha_g, \beta_g\}$ to get

$$\widehat{g}(\widehat{\theta}^g) = \widehat{g}(\theta^g) + \{\nabla_\alpha \widehat{g}(\overline{\theta}^g)\}(\widehat{\alpha}_g - \alpha_g) + \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\}(\widehat{\beta} - \beta_g)$$

where $\overline{\theta}^g$ is a value from the mean value theorem. Substitute these into each FD^g to get

$$FD^g_\alpha = \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g [\widehat{g}(\theta^g) + \{\nabla_\alpha \widehat{g}(\overline{\theta}^g)\}(\widehat{\alpha}_g - \alpha_g) + \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\}(\widehat{\beta}_g - \beta_g)]$$
$$FD^g_\beta = \{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g [\widehat{g}(\theta^g) + \{\nabla_\alpha \widehat{g}(\overline{\theta}^g)\}(\widehat{\alpha}_g - \alpha_g) + \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\}(\widehat{\beta}_g - \beta_g)]$$
$$FD^g = \{FD^g_\alpha, FD^g_\beta\} = \widehat{I}^g + \widehat{H}^g(\widehat{\theta}^g - \theta^g)$$
$$\Longrightarrow \sqrt{N}(\widehat{\theta}^g - \theta^g) = \widehat{H}^{g-1}\sqrt{N}\widehat{I}^g$$

where $\widehat{I}^g \equiv \begin{bmatrix} \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \widehat{g}(\theta^g) \\ \{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \widehat{g}(\theta^g) \end{bmatrix}$,

$$\widehat{H}^g \equiv \begin{bmatrix} \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\alpha \widehat{g}(\overline{\theta}^g)\} & \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\} \\ \{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\alpha \widehat{g}(\overline{\theta}^g)\} & \{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\} \end{bmatrix}.$$

In this expression for $\sqrt{n}(\widehat{\theta}^g - \theta^g)$, examine the part for $\sqrt{n}(\widehat{\alpha}_g - \alpha_g)$, i.e., the first $k_\alpha \times 1$ components:

$$\sqrt{n}(\widehat{\alpha}_g - \alpha_g) = \widehat{A}_g^{-1} \widehat{\Gamma}_g \sqrt{n} \widehat{g}(\theta^g)$$

where

$$\widehat{A}_g \equiv \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\alpha \widehat{g}(\overline{\theta}^g)\}$$
$$+ \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\} [\{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\beta \widehat{g}_n(\overline{\theta}^g)\}]^{-1} \{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\alpha \widehat{g}(\overline{\theta}^g)\}$$

and

$$\widehat{\Gamma}_g \equiv [\{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g - \{\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\} [\{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g \{\nabla_\beta \widehat{g}(\overline{\theta}^g)\}]^{-1} \{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\}' \widehat{\Omega}_g].$$

Then, we have

$$\sqrt{N}(\widehat{\alpha}_g - \alpha_g) = \frac{1}{\sqrt{N}}\sum_i \widehat{\eta}_i^g,$$

where

$$\widehat{\eta}_i^g \equiv \widehat{A}_g^{-1}\widehat{\Gamma}_g G(Z_i, \theta^g),$$

and $\widehat{\eta}_i^g$ is the influence function of the first-stage estimate $\widehat{\alpha}_g$. If G is correct, θ^g is replaced by θ_0^g.

Analogously for the influence functions $\widehat{\eta}_i^h$, let $\widehat{\theta}^h$ denote the second-stage estimator

$$\widehat{\theta}^h \equiv (\widehat{\alpha}_h, \widehat{\gamma}_h) = \arg\min \hat{Q}^h(\alpha, \gamma) = h(\alpha, \gamma)\widehat{\Omega}_h h(\alpha, \gamma).$$

By A7-A9 with probability approaching one the following first-order conditions in the second-stage for $\widehat{\theta}^h$ are satisfied

$$FD_\alpha^h = \frac{\partial \hat{Q}^h(\widehat{\theta}^h)}{\partial \alpha} = \{\nabla_\alpha \widehat{h}(\widehat{\theta}^g)\}'\widehat{\Omega}_h \widehat{h}(\widehat{\theta}^g) = 0,$$
$$FD_\gamma^h = \frac{\partial \hat{Q}^h(\widehat{\theta}^h)}{\partial \gamma} = \{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h \widehat{h}(\widehat{\theta}^h) = 0.$$

Expending $\widehat{h}$ around the unique minimizer $\theta^h \equiv \{\alpha_h, \gamma_h\}$ and substitute it into each FD^h to get

$$FD_\alpha^h = \{\nabla_\alpha \widehat{h}(\widehat{\theta}^g)\}'\widehat{\Omega}_h[\widehat{h}(\theta^h) + \{\nabla_\alpha \widehat{h}(\overline{\theta}^h)\}(\widehat{\alpha}_h - \alpha_h) + \{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\}(\widehat{\gamma} - \gamma_h)]$$
$$FD_\gamma^h = \{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h[\widehat{h}(\theta^h) + \{\nabla_\alpha \widehat{h}(\overline{\theta}^h)\}(\widehat{\alpha}_h - \alpha_h) + \{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\}(\widehat{\gamma} - \gamma_h)]$$
$$FD^h = \{FD_\alpha^h, FD_\gamma^h\} = I_n^h + \widehat{H}_n^h(\widehat{\theta}^h - \theta^h)$$
$$\Longrightarrow \sqrt{N}(\widehat{\theta}^h - \theta^h) = \widehat{H}_n^{h-1}\sqrt{N}\widehat{I}^h$$

where $\widehat{I}^h \equiv \begin{bmatrix} \{\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h \widehat{h}(\theta^h) \\ \{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h \widehat{h}(\theta^h) \end{bmatrix}$,

$$\widehat{H}^h \equiv \begin{bmatrix} \{\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\alpha \widehat{h}(\overline{\theta}^h)\} & \{\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\} \\ \{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\alpha \widehat{h}(\overline{\theta}^h)\} & \{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\} \end{bmatrix}.$$

In this expression for$\sqrt{n}(\widehat{\theta}^h - \theta^h)$, examine the part for $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$, i.e., the first $k_\alpha \times 1$ components:

$$\sqrt{n}(\widehat{\alpha}_h - \alpha_h) = \widehat{A}_h^{-1}\widehat{\Gamma}_h \sqrt{n}\widehat{h}(\theta^h)$$

where

$$\widehat{A}_h^{-1} \equiv \{\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\alpha \widehat{h}(\overline{\theta}^h)\}$$
$$+ \{\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\}[\{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\}]^{-1}\{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\alpha \widehat{h}(\overline{\theta}^h)\}$$

and

$$\widehat{\Gamma}_h \equiv [\{\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h - \{\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\}[\{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h\{\nabla_\gamma \widehat{h}(\overline{\theta}^h)\}]^{-1}\{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\}'\widehat{\Omega}_h].$$

Then, we have

$$\sqrt{n}(\widehat{\alpha}_h - \alpha_h) = \frac{1}{\sqrt{N}}\sum_i \widehat{\eta}_i^h,$$

where

$$\widehat{\eta}_i^h \equiv \widehat{A}_h^{-1}\widehat{\Gamma}_h H(Z_i, \theta^h),$$

and $\widehat{\eta}_i^h$ is the influence function of the first-stage estimate $\widehat{\alpha}_h$. If H is correct, θ^h is replaced by θ_0^h.