Over-Identified Doubly Robust Identification and Estimation *

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

Original 2018, Revised February 2021

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 (assuming the models are over identified). Our Over-identified Doubly Robust (ODR) technique is a simple extension of the Generalized Method of Moments. We illustrate our ODR with a variety of models. 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. 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).

We provide a general technique for constructing doubly robust (DR) estimators. The main requirements for applying our method is that models G and H each be characterized by a set of moment conditions, and each is over identified. We therefore call our method Over-identified Doubly Robust (ODR) estimation. Our ODR takes the form of a weighted average of Hansen's (1982) Generalized Method of Moments (GMM) based estimates of α, and has similar root-n asymptotics to GMM.

The main drawback of existing DR estimators is that they are not generic, meaning that for each problem, one needs to find a DR estimator, which can then be used only for that one specific application. No general method exists for finding or constructing DR estimators, and only a few examples of such models are known in the literature. Perhaps the closest thing to a general method is Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018). These authors derive a set of locally robust estimators, provide a characterization result showing when these estimators will also be DR and thereby provide some new examples of constructing DR estimators.[1] In contrast, our ODR provides a simple general method of constructing DR estimators for a very wide class of models.

[1] Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018) also show that their DR estimators possess some additional useful asymptotic properties that the ODR estimators we construct may not possess. Ideally, some different terminology would distinguish between estimators that just have the DR property (including ours and theirs) vs. estimators that have the additional properties, including local robustness, that they document.

Most existing applications of DR methods, like ATE estimation, have models G and H that are exactly identified rather than overidentified. In such cases, it may be possible to add additional overidentifying moments, and thereby apply our ODR (e.g., in a online supplemental appendix, we provide details for doing so in the ATE application). However, we do not advise using our ODR for applications where DR methods already exist, particularly when existing DR methods do not require overidentification. Instead, the main virtue of our ODR is its widespread potential application to situations where there are *not* already existing DR estimators. We provide some examples in section 3 below.

Suppose we have data consisting of n observations of a random vector Z. Assume that the true value of α satisfies either $E\left[G\left(Z,\alpha,\beta\right)\right]=0$ or $E\left[H\left(Z,\alpha,\gamma\right)\right]=0$ (or both) for some known vector valued functions G and H, and some unknown additional parameter vectors β and γ. Our ODR estimator then consistently estimates α, despite not knowing which of these two sets of equalities actually holds, for any G and H that satisfy some regularity and identification conditions.

Consider three different possible estimators for the vector α, called $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$. The estimator $\widehat{\alpha}_g$ is a GMM estimator of α that is asymptotically efficient if just the model G is correctly specified, i.e., if $E\left[G\left(Z,\alpha_0,\beta_0\right)\right]=0$ at the true α_0 and β_0. Similarly, let $\widehat{\alpha}_h$ be an asymptotically efficient GMM estimator if $E\left[H\left(Z,\alpha_0,\gamma_0\right)\right]=0$, and let $\widehat{\alpha}_f$ be a GMM estimator based on both sets of moments, which would be asymptotically efficient if both sets of moments hold at α_0, β_0, and γ_0.

One possible approach to estimation of α would be to engage in some form of model selection. Under our assumptions, model selection would be relatively straightforward. However, model selection has some disadvantages relative to DR methods, e.g., one needs to correct limiting distributions for pretest bias, and tests for which model is superior can be inconclusive. In the context of GMM based models, selection methods like Andrews and Lu (2001), Caner (2009), and Liao (2013) use test-based methods or shrinkage penalties to select moments that are most likely to be valid.

Another alternative would be model averaging, which is generally not consistent unless both G and H happen to be correctly specified. Like DR, our ODR avoids these issues. However, our

ODR estimator does take the form of a weighted average of $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$, 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 papers, we do not use typical model averaging criteria like mean squared error, Bayes weights, or information criteria to choose weights. Instead, we construct weights to yield the DR consistency property and for relative efficiency.

In the next section, we describe our ODR estimator. Section 3 then gives examples of potential applications of our ODR estimator (additional examples, including showing how existing DR applications could have alternatively been estimated using our ODR, are provided in an online supplemental appendix). In section 4 we show consistency and provide limiting distribution theory for our ODR. Section 5 provides Monte Carlo simulations and Section 6 gives an empirical application. In Section 7 we analyze properties of our estimator when the models G and H may be locally misspecified, i.e., where the parameter α_0 in the data generating process is replaced with $\alpha_0 + \delta n^{-s}$ for a constant δ and some $s > 0$. Section 8 considers extensions to more than two competing models, and Section 9 concludes. Proofs and additional results are provided in the Appendices.

2 The ODR Estimator

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 independent, identically distributed (iid) observations z_i of the vector Z.[2] The goal is root-n consistent, asymptotically normal estimation of α. Let α_0 denote the true value of α. Define model G to be 'correct,'or 'true,' if

[2]We assue iid data mainly for convenience. Our ODR is a straightforward generalization of GMM, so it should be applicable under more general conditions. We mainly require that the GMM estimators and associated objective functions satisfy some standard properties.

$E\left[G\left(Z,\alpha_0,\beta_0\right)\right]=0$ for some unique β_0. Similarly, define model H to be true if $E\left[H\left(Z,\alpha_0,\gamma_0\right)\right]=0$ for some unique γ_0. Define model F to consist of both sets of moments, and model F is true if both models G and H are true.

As discussed in the introduction, we begin with three different possible estimators for the vector α, called $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$. The estimator $\widehat{\alpha}_g$ is a GMM estimator of α that would be asymptotically efficient if model G is true and model H is not true. Specifically, $\widehat{\alpha}_g$ (along with $\widehat{\beta}_g$) minimizes the Hansen (1982) two-step quadratic GMM objective function, which we will call $\widetilde{Q}^g(\alpha,\beta)$. This $\widehat{\alpha}_g$ will generally be inconsistent if G is not true. If model G is true, then $n\widetilde{Q}^g(\widehat{\alpha}_g,\widehat{\beta}_g)$ is asymptotically chi-squared. But more importantly for us, if model G is true then $\widetilde{Q}^g(\widehat{\alpha}_g,\widehat{\beta}_g)$ itself will converge to zero in probability, and (under our assumptions) not converge to zero otherwise. We use this property to construct our ODR estimator.

Analogous to $\widehat{\alpha}_g$, let $\widehat{\alpha}_h$ denote the estimator of α based on the moments $E\left[H\left(Z,\alpha_0,\gamma_0\right)\right]=0$, so $\widehat{\alpha}_h$ and $\widehat{\gamma}_h$ minimize a quadratic GMM objective function $\widetilde{Q}^h(\alpha,\gamma)$, and are asymptotically efficient if model H is true and model G is not true. Finally, let $\widehat{\alpha}_f$ be the GMM estimator of α based on assuming both sets of moments $E\left[G\left(Z,\alpha_0,\beta_0\right)\right]=0$ and $E\left[H\left(Z,\alpha_0,\gamma_0\right)\right]=0$ hold. This $\widehat{\alpha}_f$ along with $\widehat{\beta}_f$ and $\widehat{\gamma}_f$ minimizes a GMM objective function $\widetilde{Q}^f(\alpha,\beta,\gamma)$, and is asymptotically efficient (generally more efficient than either $\widetilde{Q}^g$ or $\widetilde{Q}^h$) if both models G and H are true, but will otherwise generally be inconsistent.

Our proposed ODR estimator is a weighted average of $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$, taking the form

$$\widehat{\alpha}=\hat{W}_f\hat{W}_g\widehat{\alpha}_h+\hat{W}_f\left(1-\hat{W}_g\right)\widehat{\alpha}_g+(1-\hat{W}_f)\widehat{\alpha}_f \tag{1}$$

The novelty in our estimator relative to existing model averaging estimators is in the construction of the weights $\hat{W}_g$ and $\hat{W}_f$, given below in equations (3) and (5). In particular, we construct these weights so that, asymptotically, $\widehat{\alpha}$ becomes arbitrarily close to $\widehat{\alpha}_f$ if both models G and H are true, and otherwise becomes arbitrarily close to either $\widehat{\alpha}_g$ or $\widehat{\alpha}_h$, depending on which model is true. So, instead of the typical model averaging criteria such as minimizing mean squared error, we assume at least one of the models is correctly specified, and choose weights for efficiency, while satisfying

the DR criterion.

2.1 Starting Assumptions

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 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, for some unknown true coefficient values α_0, β_0, and γ_0, either model G is true, or model H is true, or both are true. This is a defining feature of DR estimators, and hence of our ODR estimator.

Assumption A3: The vector $G(Z, \alpha, \beta)$ has more elements than the set of elements in α and β. The vector $H(Z, \alpha, \gamma)$ has more elements than the set of elements in α and γ. 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 other DR estimators. Satisfying this assumption essentially implies that models G and H are each over identified. The first part of Assumption A3 is typically necessary to satisfy the second part, since if G contained the same number of elements as the set $\{\alpha, \beta\}$, then the equation $g_0(\alpha, \beta) = 0$ would have as many equations as unknowns, and so typically a pseudo-true solution α_1, β_1 would exist satisfying $g_0(\alpha_1, \beta_1) = 0$ even if G were misspecified.

Define the following functions:

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

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

where $\hat{\Omega}_g$ and $\hat{\Omega}_h$ are estimates of the usual weighting matrices obtained in two step GMM, which under correct specification yields asymptotic efficiency of GMM. In the above definition, $\widetilde{Q}^g(\alpha,\beta)$ is the standard Hansen (1982) and Hansen and Singleton (1982) Generalized Method of Moments (GMM) objective function, which the GMM estimator minimizes to estimate α and β. Similarly, minimizing $\widetilde{Q}^h(\alpha,\gamma)$ is the standard GMM estimator for model H. Define $\widehat{\alpha}_g$, $\widehat{\beta}_g$, $\widehat{\alpha}_h$, and $\widehat{\gamma}_h$ by

$$\{\widehat{\alpha}_g,\widehat{\beta}_g\} = \arg\min_{\{\alpha,\beta\}\in\Theta_\alpha\times\Theta_\beta} \widetilde{Q}^g(\alpha,\beta) \quad \text{and} \quad \{\widehat{\alpha}_h,\widehat{\gamma}_h\} = \arg\min_{\{\alpha,\gamma\}\in\Theta_\alpha\times\Theta_\gamma} \widetilde{Q}^h(\alpha,\gamma). \tag{2}$$

So $\{\widehat{\alpha}_g,\widehat{\beta}_g\}$ is the standard GMM estimate of model G, and $\{\widehat{\alpha}_h,\widehat{\gamma}_h\}$ is the standard GMM estimate of model H. In our applications, we likewise use the standard efficient two step GMM method for estimating the matrices $\hat{\Omega}_g$ and $\hat{\Omega}_h$.

Define $\widetilde{Q}_0^g(\alpha,\beta)$ and $\widetilde{Q}_0^h(\alpha,\gamma)$ by

$$\widetilde{Q}_0^g(\alpha,\beta) \equiv g_0(\alpha,\beta)'\Omega_g g_0(\alpha,\beta) \quad \text{and} \quad \widetilde{Q}_0^h(\alpha,\gamma) \equiv h_0(\alpha,\gamma)'\Omega_h h_0(\alpha,\gamma)$$

for positive definite matrices Ω_g and Ω_h, where $\hat{\Omega}_g \to^p \Omega_g$ and $\hat{\Omega}_h \to^p \Omega_h$.

Assumption A4: Assume there exists $\{\alpha_g,\beta_g\} \in \Theta_\alpha \times \Theta_\beta$ such that $\widetilde{Q}_0^g(\alpha_g,\beta_g) < \widetilde{Q}_0^g(\alpha,\beta)$ for all $\{\alpha,\beta\} \in \Theta_\alpha \times \Theta_\beta \backslash \{\alpha_g,\beta_g\}$ and there exists $\{\alpha_h,\gamma_h\} \in \Theta_\alpha \times \Theta_\gamma$ such that $\widetilde{Q}_0^h(\alpha_h,\gamma_h) < \widetilde{Q}_0^h(\alpha,\gamma)$ for all $\{\alpha,\gamma\} \in \Theta_\alpha \times \Theta_\gamma \backslash \{\alpha_h,\gamma_h\}$.

Given Assumptions A2 and A3, Assumption A4 will automatically be satisfied for model G when G is correctly specified, with $\{\alpha_g,\beta_g\} = \{\alpha_0,\beta_0\}$, and similarly for $\{\alpha_h,\gamma_h\}$ when H is correctly specified, by Lemma 2.3 of Newey and McFadden (1994). Together with Assumptions A1 to A3, Assumption A4 implies that GMM estimators of G or H will also converge to unique values (pseudo-true values) when they are misspecified. Assumption A4 is also imposed by Hall (2000) and Hall and Inoue (2003) for misspecified GMM models.

Our main reason for having Assumption A4 is to ensure that the weights $\hat{W}_f$ and $\hat{W}_g$ are asymptotically well behaved, which simplifies derivation of limiting distributions (and asymptotics under local misspecification). However, some of our results (like consistency of the SODR estimator defined below) will not require Assumption A4.

2.2 The SODR and ODR estimators

Let $c_g \equiv g_0(\alpha_g, \beta_g)$. Under minimal, standard regularity conditions (see details in the next section), we have $\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) \to^p c_g' \Omega_g c_g$. If G is correctly specified, then $\alpha_g = \alpha_0$ and $\beta_g = \beta_0$, which makes $c_g = 0$, so $c_g' \Omega_g c_g = 0$. What is important for our ODR estimator is that the probability limit of $\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ is zero if G is correctly specified, and positive otherwise.

Having G correctly specified also means (again with minimal regularity), that $n^{1/2}\widehat{g}(\widehat{\alpha}_g, \widehat{\beta}_g)\Omega_g^{1/2} \to_d N\left(0, I_{k_g}\right)$ so $n\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) \to_d \chi^2_{k_g}$. However, if G is incorrectly specified, then $c_g \neq 0$, so $c_g' \Omega_g c_g > 0$ and $n\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ does not follow the chi-squared distribution asymptotically. Analogous statements hold for model H.

Let $\hat{Q}^g(\alpha, \beta) \equiv \widetilde{Q}^g(\alpha, \beta)/k_g$ and $\hat{Q}^h(\alpha, \gamma) \equiv \widetilde{Q}^h(\alpha, \gamma)/k_h$, where the integer k_g is the degrees of freedom of the chi-squared statistic that $n\widetilde{Q}^g$ converges to if the G model is true. This is the number of moments in G minus the number of elements in α and β, which is positive as discussed earlier. Similarly, k_h is the degrees of freedom of the chi-squared statistic that $n\widetilde{Q}^h$ equals if the H model is true. This scaling by k_g and k_h is not necessary for our estimator, but improves its finite sample performance (see below for details).

Define $\hat{W}_g$ by

$$\hat{W}_g \equiv \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)}. \tag{3}$$

From the above derivations, we have that, if G is correctly specified and H is not,

$$\hat{W}_g \to^p \frac{0}{0 + c_h' \Omega_h c_h / k_h} = 0,$$

while if H is correctly specified and and G is not,

$$\hat{W}_g \to^p \frac{c_g' \Omega_g c_g / k_g}{c_g' \Omega_g c_g / k_g + 0} = 1.$$

Before getting to our ODR estimator given by equation (1), consider the simpler estimator $\widetilde{\alpha}$ defined by

$$\widetilde{\alpha} = \hat{W}_g \widehat{\alpha}_h + \left(1 - \hat{W}_g\right) \widehat{\alpha}_g. \tag{4}$$

So $\widetilde{\alpha}$ is simply a weighted average of the GMM estimates $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$, where the weights are proportional to $\hat{Q}^g$ and $\hat{Q}^h$. We will call $\widetilde{\alpha}$ the SODR (simpler ODR) estimator.

The intuition behind $\widetilde{\alpha}$ is straightforward (the asymptotic statements in this paragraph are proved formally in the next section). Suppose 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 $\hat{W}_g$, the weight on $\widehat{\alpha}_h$ in equation (4) will go to zero, and $\left(1 - \hat{W}_g\right)$, the weight on $\widehat{\alpha}_g$, will go to one. As a result, $\widetilde{\alpha}$ 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 $\widetilde{\alpha}$ 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, $\widetilde{\alpha} \to^p \alpha_0$.[3]

We could have defined the weight $\hat{W}_g$ without scaling each GMM objective function by its degrees of freedom. Asymptotically, the estimator would still be doubly robust. The reason we scale is because, even when a model is correctly specified, in finite samples the greater is the degrees of freedom of a model, the larger its GMM objective function is likely to be. Asymptotically, the mean of $n\widetilde{Q}^g$ converges to k_g when g is correctly specified, and similarly for h. So, by scaling, when both models are correctly specified, both $n\hat{Q}^g$ and $n\hat{Q}^h$ will asymptotically have mean one. Otherwise,

[3]Notice that when both G and H are correctly specified, $\hat{W}_g$ converges to a ratio of correlated chi-squared distributions, not to a constant. Nevertheless, $\widetilde{\alpha}$ is still consistent because $\widetilde{\alpha} = \widehat{\alpha}_g + (\widehat{\alpha}_h - \widehat{\alpha}_g)\hat{W}_g$, and when both are correctly specified, $\widehat{\alpha}_g \to_p \alpha_0$ and $\widehat{\alpha}_h - \widehat{\alpha}_g \to_p 0$.

if we didn't scale, whichever model has more moments will tend to have a larger GMM objective function, which would then undesirably penalize that model in finite samples.

Although the SODR $\widetilde{\alpha}$ has the desired DR property, it also has two drawbacks. First, when G and H are both correct, the ratio $\hat{W}_g$ converges to a random variable rather than a constant, which complicates the limiting distribution of $\widetilde{\alpha}$. Second, when both G and H are correct, $\widetilde{\alpha}$ may be inefficient, relative to a GMM estimator that efficiently combines the moments from both models.

To address both of these issues, reconsider now the third model F, defined as the union of moments of the models G and H. Specifically, let $F(Z, \alpha, \beta, \gamma)$ be the vector valued function consisting of the union of elements of $G(Z, \alpha, \beta)$ and $H(Z, \alpha, \gamma)$. Then, letting $\widehat{f}(\alpha, \beta, \gamma) \equiv \frac{1}{n}\sum_{i=1}^{n} F(Z_i, \alpha, \beta, \gamma)$, we can define a third GMM estimator

$$\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} = \arg \min_{\{\alpha,\beta,\gamma\} \in \Theta_\alpha \times \Theta_\beta \times \Theta_\gamma} \widetilde{Q}^f(\alpha, \beta, \gamma)$$

where $\widetilde{Q}^f(\alpha, \beta, \gamma) \equiv \widehat{f}(\alpha, \beta, \gamma)' \hat{\Omega}_f \widehat{f}(\alpha, \beta, \gamma)$. This is efficient GMM assuming both specifications are correct, and so uses all the moments from both. If models G and H are correctly specified, then $\widehat{\alpha}_f$ is at least as asymptotically efficient, and generally much more asymptotically efficient, than $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, or $\widetilde{\alpha}$. Let $c_f \equiv f_0(\alpha_f, \beta_f, \gamma_f) \equiv E\{F(Z, \alpha_f, \beta_f, \gamma_f)\}$. Then $\widetilde{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \to^p c_f' \Omega_f c_f$, which equals zero if both models G and H are correctly specified, and is positive otherwise.

We again scale by the degrees of freedom (number of moments in F minus number of elements of α, β, and γ), denoted k_f, defining $\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \equiv \widetilde{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)/k_f$. We then define the weight $\hat{W}_f$ by

$$\hat{W}_f \equiv 1 - \frac{1}{n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) + 1} \tag{5}$$

for some τ having $0 < \tau < 1$. Later we discuss selection of the tuning parameter τ, but for consistency we only require that τ lie between zero and one. Our ODR estimator, given by equation (1), can be equivalently written as

$$\widehat{\alpha} = \hat{W}_f \widetilde{\alpha} + \left(1 - \hat{W}_f\right) \widehat{\alpha}_f. \tag{6}$$

The intuition now is, if both G and H are correctly specified, then $\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \to^p 0$ and

$n\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ converges in distribution to a chi-squared statistic (divided by its degrees of freedom), which means that $n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ for $0 < \tau < 1$ converges in probability to zero. Alternatively, if either G or H is incorrectly specified, then $\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ converges in probability to a positive value, so $n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ diverges to infinity. Therefore, if both G and H are correctly specified then $\hat{W}_f \to^p 0$ and so $\widehat{\alpha}$ has the same limiting value as $\widehat{\alpha}_f$, while if either G or H is incorrectly specified, then $\widehat{\alpha}$ has the same limiting value as $\widetilde{\alpha}$, which as shown earlier has the same limiting value as either $\widehat{\alpha}_g$ or $\widehat{\alpha}_h$, depending on which is correctly specified.

The estimator $\widehat{\alpha}$ therefore, like $\widetilde{\alpha}$, has the desired DR property. We show later that $\widehat{\alpha}$ avoids the asymptotic issues $\widetilde{\alpha}$ has when both G and H are correctly specified, and that $\widehat{\alpha}$ generally performs better than $\widetilde{\alpha}$ in finite samples. This is why $\widehat{\alpha}$ is our preferred ODR estimator. However $\widehat{\alpha}$ has the disadvantages of being a little more complicated to estimate (since it requires estimating the third model F), and it requires selection of a tuning parameter τ.

2.3 Tuning Parameters

One tuning parameter is τ, which for consistency can take any value between zero and one. The larger τ is, the less weight is put on $\widehat{\alpha}_f$ in any given sample. So for efficiency, the more likely it is that both models G and H are correct, the smaller one would want τ to be. Based on this observation, a choice of τ that we find works well in Monte Carlo simulations is to let $\tau = 1 - p$, where p is the p-value of the Wald statistic testing the null hypothesis that $\widehat{\alpha}_g = \widehat{\alpha}_h$.[4] [5]

Another potential tuning parameter is as follows. Let Λ be any strictly monotonically increasing function such that $\Lambda(0) = 0$ and $\Lambda(\cdot) \to \infty$ when $\cdot \to \infty$. Then $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$, $n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$, and

[4] Our derivation of the limiting distribution of $\widehat{\alpha}$ assumes $\tau > 1/2$, however, this restriction is only required to handle cases where $\alpha_g \neq \alpha_h$, and we $\tau = 1 - p$ will asymptotically increase to over $1/2$ in those cases.

[5] Under possible local misspecification, which we consider in section 7 below, choice of τ becomes more complicated, for two reasons. First, under local misspecification, having a random τ can affect the limiting distribution of $\widehat{\alpha}$. And second, for some range of rates of local misspecification parameter drift, a relatively large value of τ is needed to avoid complications in limitation distributions.

$n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ can be replaced with $\Lambda\left(n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\right)$, $\Lambda\left(n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\right)$, and $\Lambda\left(n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\right)$ in the definitions of the weights $\hat{W}_g$ and $\hat{W}_f$ in equations (3) and (5). The main asymptotic properties of the ODR estimator are preserved by any such choice of Λ, but finite sample properties of the estimator might be improved by different choices of the function Λ. For example, $\Lambda(z) = \exp(\lambda z) - 1$ for some $\lambda > 0$ resembles exponential tilting. Equation (4) already somewhat resembles Bayesian model averaging, and this choice of Λ would make that resemblance stronger.[6] See e.g., Kim (2002) and Martins and Gabriel (2014).[7] Another choice for Λ would be a simple power transform $\Lambda(z) = z^\lambda$ for $\lambda > 0$. We consider different choices of Λ in our applications. Overall, we found that the exponential Λ works well, though choice of Λ had only modest effects on our monte carlo simulations, and virtually no effect on our empirical estimates.

Finally, we require estimators for the GMM weighting matrices $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$. As discussed later in section 4.3, these are the standard estimated weighting matrices used in two step GMM, but recentered. See in particular equation (11).

3 ODR Examples

Before proceeding to show consistency and deriving the limiting distribution of the ODR estimator, we present two example applications. Both are new applications for which no existing DR estimators are known. One concerns estimation of preference parameters in consumption Euler equations and asset pricing kernels. The second is alternative sets of instruments for linear model estimation.

In an Online Supplemental Appendix, we provide two additional examples, comparing the requirements of our ODR estimator to existing DR applications. The first discusses average treatment effect estimation, while the second concerns additive regression models.

[6] A key difference with Bayesian or information based weighting is that we weight model G based on the model H objective function, and vice versa, instead of weighting each model by its own objective function.

[7] We discuss comparisons of our estimator with Martins and Gabriel (2014) in more detail later, in sections 4.4 and 5.0.

3.1 Preference Parameter Estimates

One of the original applications of GMM estimation, Hansen and Singleton (1982), was the estimation of marginal utility parameters and of pricing kernels. 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 lifetime budget constraint yields Euler equations of the form

$$E\left(bR_{t+1}\frac{U'\left(C_{t+1}, X_{t+1}, \rho\right)}{U'\left(C_t, X_t, \rho\right)} - 1 \mid W_\tau\right) = 0 \tag{7}$$

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 ODR estimator can be immediately applied, replacing the expression in the inner parentheses in equation (7) with $G(Z, \alpha, \beta)$ or $H(Z, \alpha, \gamma)$ to represent the two different specifications. Here α would represent parameters that are the 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 the coefficient of relative risk aversion, and the function $M\left(X_t\right)$ 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 ODR.

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 (7), we could define $G(Z, \alpha, \beta)$ and $H(Z, \alpha, \gamma)$ by

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

and

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

In this example, we would have $\alpha = (b, \rho)$, and so would consistently estimate the discount rate b and the coefficient of relative risk aversion ρ, no matter which habit model is correct. To satisfy the required overidentification (Assumption A3), we would want W_τ to have more elements than (α, β) and more than (α, γ). This would generally be the case, because the potential information set of consumers at time t is large relative the the number of parameters in the model.

3.2 Alternative Sets of Instruments

Consider a parametric model

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

where Y is an outcome, W is a 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. The errors ϵ may be correlated with W, so to estimate the model we wish to find instruments that are uncorrelated with ϵ. Let R and Q denote two different vectors of observed covariates that are candidate sets of

instruments. One may be unsure if either R or Q are valid instrument vectors or 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 definitions we can then immediately apply the ODR 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 overidentification condition, Assumption A3, is generally satisfied when Q and R each have more elements than W.

Next consider a richer example, which we later empirically apply, 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 $E(\epsilon X) = 0$, and S is an endogenous or mismeasured scalar covariate that is correlated with ϵ. The goal is estimation of the set of coefficients $\alpha = \{\alpha_x, \alpha_s\}$.

The 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 resulting GMM (or linear two stage least squares) estimator would be based on the moments $E[G(Z, \alpha)] = 0$ where $G(Z, \alpha)$ is given by the stacked vectors

$$G(Z, \alpha) = \left\{ \begin{array}{c} X\left(Y - X'\alpha_x - S\alpha_s\right) \\ L\left(Y - X'\alpha_x - S\alpha_s\right) \end{array} \right\}. \tag{8}$$

The main difficulty with applying this two stage least squares or GMM estimator is that one must find one or more covariates L to serve as instruments.

Lewbel (2012) proposes an alternative estimator that, rather than requiring that one find instruments L, instead constructs instruments based on assumptions regarding heteroscedasticity. This estimator consists of first linearly regressing S on X, and obtaining the residuals from that regression. Then a vector of instruments P is constructed by setting P equal to demeaned X (excluding

the constant) times these residuals. This constructed vector P is then used instead of L above as instruments.[8] As shown in Lewbel (2012), one set of conditions under which the vector P can be a valid set of instruments is when the endogeneity in S is due to classical measurement error in S.

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

$$H(Z, \alpha, \gamma) = \left\{ \begin{array}{c} X_c - \gamma_1 \\ X(S - X'\gamma_2) \\ X\left(Y - X'\alpha_x - S\alpha_s\right) \\ (X_c - \gamma_1)(S - X'\gamma_2)\left(Y - X'\alpha_x - S\alpha_s\right) \end{array} \right\}. \tag{9}$$

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. ODR estimation based on these moments therefore allows for consistent estimation of α if either one of these sets of conditions hold. To satisfy the over identification Assumption A3, X_c and L must each have two or more elements.

As a motivating example, consider the following application involving 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. To deal with this budget measurement error, a commonly employed set of instruments L consists of functions of the consumer's income. However, validity of functions of income as instruments for total consumption in a food Engel curve assumes separability between the consumer's decisions on savings and their within period food expenditure decision, and this behavioral assumption may or may not be valid. It is therefore useful to consider the alternative

[8] This estimator is implemented in the STATA module IVREG2H by Baum and Schaffer (2012).

set of potential instruments P defined above. Use of P does not require finding covariates from outside the model, like income, to use as instruments, but does require that certain measurement error assumptions hold. Our later empirical application applies ODR to this application, thereby obtaining consistent estimates of α if either L or P are valid instruments.

4 The ODR Estimator Asymptotics

In this section we show consistency of our ODR estimator $\widehat{\alpha}$, and then derive its limiting distribution, which is root n consistent and asymptotically normal. We make the following additional assumptions. What these assumptions mostly do is make GMM estimation of models G, H, and F asymptotically normal around either the true values when correctly specified, or around pseudo-true values when misspecified, and ensure that the models are over identified.

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

Assumption A6: With $||A|| \equiv \{trace(A'A)\}^{1/2}$ for a matrix A, $E[\sup_{\{\alpha,\beta\}\in\Theta_\alpha\times\Theta_\beta} ||G(Z,\alpha,\beta)||] < \infty$, $E[\sup_{\{\alpha,\gamma\}\in\Theta_\alpha\times\Theta_\gamma} ||H(Z,\alpha,\gamma)||] < \infty$, and $E[\sup_{\{\alpha,\beta,\gamma\}\in\Theta_\alpha\times\Theta_\beta\times\Theta_\gamma} ||F(Z,\alpha,\beta,\gamma)||] < \infty$.

Taken together Assumptions A1, A2, A3, A5, and A6, are standard conditions that suffice for consistency of the GMM estimators of models G, H, and F when they are correctly specified. See, e.g., Theorem 2.1 in Newey and McFadden (1994). Let $\nabla_\theta(\cdot) \equiv \partial(\cdot)/\partial\theta$ be arranged such that its row dimension is that of θ and let $\nabla_{\theta'}(\cdot) \equiv \{\nabla_\theta(\cdot)\}'$. Define $\theta_0^g \equiv \{\alpha_0,\beta_0\}$, $\theta_0^h \equiv \{\alpha_0,\gamma_0\}$, $\theta_0^f \equiv \{\alpha_0,\beta_0,\gamma_0\}$, $\theta^g \equiv \{\alpha_g,\beta_g\}$, $\theta^h \equiv \{\alpha_h,\gamma_h\}$, and $\theta^f \equiv \{\alpha_f,\beta_f,\gamma_f\}$.

Assumption A7: With probability one, $G(Z,\alpha,\beta)$, $H(Z,\alpha,\gamma)$, and $F(Z,\alpha,\beta,\gamma)$ are twice continuously differentiable in a neighborhood $\aleph^g$ of θ^g, $\aleph^h$ of θ^h, and $\aleph^f$ of θ^f, respectively.

Assumption A8: $\nabla_\theta g_0(\theta_0^g)\Omega_g\nabla_{\theta'}g_0(\theta_0^g)$, $\nabla_\theta h_0(\theta_0^h)\Omega_h\nabla_{\theta'}h_0(\theta_0^h)$, and $\nabla_\theta f_0(\theta_0^f)\Omega_f\nabla_{\theta'}f_0(\theta_0^f)$ are non-singular.

Assumption A9: $\{\alpha_g, \beta_g\}$, $\{\alpha_h, \gamma_h\}$, and $\{\alpha_f, \beta_f, \gamma_f\}$ lie in the interior of $\Theta_\alpha \times \Theta_\beta$, $\Theta_\alpha \times \Theta_\gamma$, and $\Theta_\alpha \times \Theta_f \times \Theta_\gamma$.

Assumption A10: $E[||G(Z, \alpha, \beta)||^2] < \infty$, $E[||H(Z, \alpha, \gamma)||^2] < \infty$, and $E[||F(Z, \alpha, \beta, \gamma)||^2] < \infty$.

Assumption A11: $E[\sup_{\{\alpha,\beta\}\in\aleph^g} ||\nabla_{\theta^g} G(Z, \alpha, \beta)||] < \infty$, $E[\sup_{\{\alpha,\gamma\}\in\aleph^h} ||\nabla_{\theta^h} H(Z, \alpha, \gamma)||] < \infty$, and $E[\sup_{\{\alpha,\beta,\gamma\}\in\aleph^f} ||\nabla_{\theta^f} F(Z, \alpha, \beta, \gamma)||] < \infty$.

Assumption A7, A9, A10, and A11 are regularity conditions for a uniform weak law of large numbers and the asymptotic normality of GMM. Assumption A8 rules out perfect collinearity in linearized moment conditions. Assumption A11 gives interchangeability of $\nabla(\cdot)$ and $E(\cdot)$ so that

$$\nabla_\theta g_0(\theta^g) = E\{\nabla_{\theta^g} G(Z, \alpha_g, \beta_g)\},\ \nabla_\theta h_0(\theta^h) = E\{\nabla_{\theta^h} H(Z, \alpha_h, \gamma_h)\},\ \nabla_\theta f_0(\theta^f) = E\{\nabla_{\theta^f} F(Z, \alpha_f, \beta_f, \gamma_f)\}.$$

Assumption A12: $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ are $\sqrt{n}$-consistent, asymptotically normal estimators of Ω_g, Ω_h and Ω_f, respectively, where $\Omega_g^{-1} = Var\left[G(Z, \alpha_g, \beta_g)\right]$, $\Omega_h^{-1} = Var\left[H(Z, \alpha_h, \gamma_h)\right]$, and $\Omega_f^{-1} = Var\left[F(Z, \alpha_f, \beta_f, \gamma_f)\right]$.

Assumption A13: $E[||\nabla_{\theta^g} G(Z, \alpha, \beta)||^2] < \infty$, $E[||\nabla_{\theta^h} H(Z, \alpha, \gamma)||^2] < \infty$, and $E[||\nabla_{\theta^f} F(Z, \alpha, \beta, \gamma)||^2] < \infty$.

Assumption A14: Letting $\nabla_{\theta^g\theta^{g\prime}}(\cdot) \equiv \partial(\cdot)/\partial\theta^g\partial\theta^{g\prime}$, $E[\sup_{\{\alpha,\beta\}\in\aleph^g} ||\nabla_{\theta^g\theta^{g\prime}} G(Z, \alpha, \beta)||] < \infty$, $E[\sup_{\{\alpha,\gamma\}\in\aleph^h} ||\nabla_{\theta^h\theta^{h\prime}} H(Z, \alpha, \gamma)||] < \infty$, and $E[\sup_{\{\alpha,\beta,\gamma\}\in\aleph^f} ||\nabla_{\theta^f\theta^{f\prime}} F(Z, \alpha, \beta, \gamma)||] < \infty$.

Assumption A15: $plim Var\left[\frac{1}{\sqrt{n}}\sum_i G(Z_i, \theta^g)\right]$, $plim Var\left[\frac{1}{\sqrt{n}}\sum_i H(Z_i, \theta^h)\right]$, and $plim Var\left[\frac{1}{\sqrt{n}}\sum_i F(Z_i, \theta$ exist and are positive definite.

Assumption A12 strengthens the standard assumption for asymptotically efficient GMM estimation in requiring that the estimated weighting matrices converge at rate $\sqrt{n}$. This assumption is satisfied by the standard two-step GMM estimators for $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$, provided that the sample moments are demeaned, e.g., Ω_g is based on $Var\left[G(Z, \alpha_g, \beta_g)\right]$ rather than $E\left[G(Z, \alpha_g, \beta_g)G(Z, \alpha_g, \beta_g)'\right]$.

The strengthening of Assumption A12 over the standard assumptions for GMM estimation ensures that the probability limits of $\hat{W}_g$ and $\hat{W}_g\hat{W}_f$ remain well behaved when either model G or H is misspecified.

Assumptions A13, A14, and A15 are for the asymptotic normality of the normalized sum of derivatives of G, H, and F. These assumptions are to ensure asymptotic normality of the GMM estimators when model G or H is misspecified. Assumptions A12 to A14 above are adapted from Hall and Inoue (2003), who use them to derive asymptotics for possibly misspecified GMM estimation.

4.1 ODR Consistency

Lemma 1: Suppose Assumptions A1 to A15 hold. Then, for any τ with $0 < \tau < 1$, $\hat{W}_f$ and $\hat{W}_f\hat{W}_g$, defined in equations (5) and (3), have finite probability limits. Specifically,

$$\begin{aligned}
&\text{Case 1) } G \text{ and } H \text{ are correctly specified} && \Longrightarrow \hat{W}_f \to^p 0 \text{ and } \hat{W}_f\hat{W}_g \to^p 0,\\
&\text{Case 2) } G \text{ is correctly specified but } H \text{ is not} && \Longrightarrow \hat{W}_f \to^p 1 \text{ and } \hat{W}_f\hat{W}_g \to^p 0,\\
&\text{Case 3) } H \text{ is correctly specified but } G \text{ is not} && \Longrightarrow \hat{W}_f \to^p 1 \text{ and } \hat{W}_f\hat{W}_g \to^p 1.
\end{aligned}$$

Lemma 1 is proved in Appendix I, but the intuition is as follows. When either G or H is misspecified, we have $\hat{Q}^f \to^p c_f'\Omega_f c_f/k_f > 0$, so $n^\tau\hat{Q}^f$ diverges to infinity and $\hat{W}_f \to^p 1$. If G is correct but H is not, then $\hat{Q}^g \to^p 0$ while the limiting value of $\hat{Q}^h$ is nonzero. Thus, $\hat{W}_g \to^p 0$ and so $\hat{W}_g\hat{W}_f \to^p 0$. If H is correct but G is not, following the same logic but switching the roles of g and h, $\hat{W}_g \to^p 1$ and so $\hat{W}_g\hat{W}_f \to^p 1$. When both G and H are correctly specified, so F is correctly specified, we have $\hat{Q}^f \to^p c_f'\Omega_f c_f/k_f = 0$, so $n^\tau\hat{Q}^f \to^p 0$ and therefore $\hat{W}_f \to^p 0$, and in this case both $n\hat{Q}^g$ and $n\hat{Q}^h$ converge to chi-squared distributions so $\hat{W}_g$ converges to a ratio of possibly dependent chi-squares, which is bounded in probability, making $\hat{W}_g\hat{W}_f \to^p 0$.

The following theorem shows consistency of the ODR estimator $\widehat{\alpha}$ in equation (1). We will further discuss construction of $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ later, but note for now that these are recentered GMM weight matrix estimates using the sample moments in mean deviation form.

Theorem 1: Under Assumptions A1 to A15, for $\widehat{\alpha}$ given by equation (1), $\widehat{\alpha} \to^p \alpha_0$.

Proof of Theorem 1: By A1, A2, A3, A5, and A6, the conditions of Theorem 2.1 of in Newey and McFadden (1994) (uniqueness, compactness, continuity, and uniform convergence) hold for GMM based on model G, model H, or both when these moments are correctly specified. Therefore, if $g_0(\alpha_0, \beta_0) = 0$ then the GMM estimator of model G is consistent, if $h_0(\alpha_0, \gamma_0) = 0$ holds then the GMM estimator of model H is consistent, and if both the equalities hold parts hold then the GMM estimator of F is consistent.

For simplicity, let $\hat{Q}^g \equiv \hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$, $\hat{Q}^h \equiv \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$, $\hat{Q}^f \equiv \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$, $Q_0^g \equiv c_g' \Omega_g c_g / k_g$, $Q_0^h \equiv c_h' \Omega_h c_h / k_h$, and $Q_0^f \equiv c_f' \Omega_f c_f / k_f$. Assumption A2 says that either $g_0(\alpha_0, \beta_0) = 0$, $h_0(\alpha_0, \gamma_0) = 0$, or both. Consider each of these three cases.

Case 1) Suppose both $g_0(\alpha_0, \beta_0) = 0$ and $h_0(\alpha_0, \gamma_0) = 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_0, \beta_0, \gamma_0\}$, so $\hat{Q}^g \to^p 0$, $\hat{Q}^h \to^p 0$, and $\hat{Q}^f \to^p 0$. By Lemma 1, $\hat{W}_f$ and $\hat{W}_f \hat{W}_g$ both converge to zero, and the consistency of $\widehat{\alpha}$ therefore follows from consistency of $\widehat{\alpha}_f$.

Case 2) Suppose that $g_0(\alpha_0, \beta_0) = 0$ and $h_0(\alpha_0, \gamma_0) \neq 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_h, \gamma_h\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. By Lemma 1, $\hat{W}_g$ converges to zero and $\hat{W}_f$ converges to one in probability. The consistency of $\widehat{\alpha}$ then follows from consistency of $\widehat{\alpha}_g$.

Case 3) Suppose that $g_0(\alpha_0, \beta_0) \neq 0$ and $h_0(\alpha_0, \gamma_0) = 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_g, \beta_g\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. By Lemma 1, $\hat{W}_g$ and $\hat{W}_f$ both converge to one in probability, so consistency of $\widehat{\alpha}$ follows from consistency of $\widehat{\alpha}_h$. Q.E.D.

4.2 Limiting Distribution

We now provide the asymptotic distribution of $\widehat{\alpha}$, and a simple consistent estimator of its limiting variance. Let $\widehat{\eta}_i^g$, $\widehat{\eta}_i^h$ and $\widehat{\eta}_i^f$ be consistent estimators of the GMM influence functions for $\widehat{\alpha}_g$, $\widehat{\alpha}_h$ and $\widehat{\alpha}_f$, the details of which are in Appendix III.

Theorem 2: Suppose Assumptions A1 to A15 hold. Then, for $1/2 < \tau < 1$, there exists a

matrix $\widetilde{V}$ such that

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

and

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \widehat{\eta}_i \widehat{\eta}_i' \rightarrow^p \widetilde{V} \tag{10}$$

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

The first part of Theorem 2 states that the ODR estimator $\widehat{\alpha}$ is root n consistent and asymptotically normal, while the second part gives a consistent estimator for the limiting variance of $\widehat{\alpha}$. The proof of Theorem 2 is given in the Appendix I. The basic structure of the proof follows Newey and McFadden (1994) for multistep parametric estimators.

Note that while consistency only requires $0 < \tau < 1$, Theorem 2 assumes $\tau > 1/2$ to ensure $\sqrt{n}$-consistency of $\widehat{\alpha}$. This condition is only required for the case where $\alpha_g \neq \alpha_h$.

The estimator of $\widetilde{V}$ given in equation (10) does not require knowing which of the models G or H is correct. Nevertheless, as shown in Appendix I, $\widetilde{V}$ will either equal a matrix $\widetilde{V}^g$ or $\widetilde{V}^h$ or $\widetilde{V}^f$, depending on whether models G, H, or both are correctly specified.

A complication in the derivation of Theorem 2 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 $\hat{W}_f \hat{W}_g$, so if model H is wrong then $\hat{W}_f \hat{W}_g$ goes to 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 $\widehat{\eta}_i^h$. A similar analysis applies to the influence function $\widehat{\eta}_i^g$ for model G when model G is wrong.

4.3 Efficiency and Numerical Issues

For asymptotic efficiency of α, we could consider estimating the weighting matrices $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ to minimize the variance given by equation (10). However, the standard two step GMM estimators of $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ should be at least close to efficient for $\widehat{\alpha}$. This is because the ODR

objective function is asymptotically dominated by the GMM objective function of the correct model when either G or H is correct, and dominated by the GMM objective function of model F when both models are correct.

The scaling of moments affects the relative magnitudes of $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$ (and hence the estimated weights $\hat{W}_g$ and $\hat{W}_f$). It is therefore numerically desirable in finite samples to have these matrices be comparable in magnitude. The standard two step GMM estimators of $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ help make $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$ comparable. Specifically, standard two step GMM makes $n\hat{Q}^g$ have a mean of one asymptotically when model G is right, and similarly for $n\hat{Q}^h$ and $n\hat{Q}^f$ (this is also the role of scaling each by the degrees of freedom k_g, k_h, and k_f, respectively). We therefore find it desirable to use the standard GMM estimates of $\hat{\Omega}_g$ and $\hat{\Omega}_h$ (as in Assumption A12) even if that possibly sacrifices a small amount of efficiency. In particular, we let

$$\hat{\Omega}_g \equiv \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(G(Z_i,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})-\overline{G}(Z,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})\right)\left(G(Z_i,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})-\overline{G}(Z,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})\right)' \tag{11}$$

where $\widehat{\alpha}_{1g}$ and $\widehat{\beta}_{1g}$ are first step GMM estimates based on a constant weighting matrix such as the identity matrix, and $\overline{G}$ is the sample average of $G(Z_i,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})$. Analogous formulas apply for $\hat{Q}^h$ and $\hat{Q}^f$.

4.4 Comparison to Model Averaging

The weights in our SODR and ODR estimators can be compared to more traditional model averaging methods. An example of GMM model averaging (for instrument selection in linear instrumental variables models) is Martins and Gabriel (2014), who construct weights based on Andrews (1999)'s J-statistic based GMM model selection criteria. To most readily compare their weights to ours, consider the special case of our ODR in which the candidate models G and H are linear regressions with different sets of instruments. This comparison is particularly apt because our simulations and empirical application are choice of instruments in linear models.

Martins and Gabriel (2014) provide a variety of estimators, but the one that is closest to our

model is

$$\widetilde{\alpha}^{MG} \equiv \hat{W}_g^{MG}\widehat{\alpha}_h + \left(1 - \hat{W}_g^{MG}\right)\widehat{\alpha}_g$$
$$\text{where } \hat{W}_g^{MG} \equiv \frac{\exp\left(-\frac{1}{2}(n\widetilde{Q}^h - \kappa_n k_h)\right)}{\exp\left(-\frac{1}{2}(n\widetilde{Q}^h - \kappa_n k_h)\right) + \exp\left(-\frac{1}{2}(n\widetilde{Q}^g - \kappa_n k_g)\right)}$$

and $\kappa_n = o(n)$ is a sequence depending on the selection criteria, e.g. $\kappa_n = \ln(n)$ for a Bayesian Information Criterion. This estimator is similar to our SODR with an exponential tuning function Λ.

One difference between $\widetilde{\alpha}^{MG}$ and SODR (with exponential Λ) is in the degrees of freedom term κ_n. Another important difference is that, in $\widetilde{\alpha}^{MG}$, as in other model averaging methods, the numerator of the weight on each model depends on the criterion for that model, while in our estimator, the numerator of the weight on model H depends on the criterion for model G (i.e., on $\widetilde{Q}^g$) and vice versa. This is because, for the DR property, we asymptotically need to put all weight on model H when model G is wrong, and vice versa. Note that Martins and Gabriel (2014) assume both models are correctly specified, and they do not account for the weight $\hat{W}_g^{MG}$ having a possibly random probability limit.

In contrast to SODR, our preferred ODR estimator differs more substantially from $\widetilde{\alpha}^{MG}$ in its construction of weights. We compare the finite sample performance of both our SODR and ODR estimators to $\widetilde{\alpha}^{MG}$ in the next section.

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 3.2. For each simulation, we draw $n = 100$ or $n = 500$ independent, identically distributed observations of the random vector $(Y, W, R_1, R_2, Q_1, Q_2)$. We generate data from the model

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

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

We let $W = 1 + 4R_1 + R_2 + 2Q_1 + Q_2 + \epsilon$. Having the 4 and 2 in this equation means that model G has stronger instruments (i.e., instruments more highly correlated with the endogenous regressor W) than model H, and that R_1 and Q_1 are stronger instruments than R_2 and Q_2.

We let R_1, R_2, Q_1, Q_2, and ϵ be standard normals, with $corr(R_j, \epsilon) = \rho_{Rj}$, $corr(Q_j, \epsilon) = \rho_{Qj}$, for $j = 1, 2$, and all the other correlations among these normals are zero. We consider three different simulation designs, that vary by correlations ρ_{Rj} and ρ_{Qj}. The first design takes $\rho_{Rj} = \rho_{Qj} = 0$, which makes both models right (both sets of instruments are valid). The second takes $\rho_{R1} = \rho_{R2} = 0$, $\rho_{Q1} = 0.4$, and $\rho_{Q2} = 0.6$, which makes model G right (i.e., R are valid instruments so G is correctly specified) and model H be wrong (i.e., Q are not valid instruments, because they correlate with the model error ϵ). The third takes $\rho_{R1} = 0.4$, $\rho_{R2} = 0.6$ and $\rho_{Q1} = \rho_{Q2} = 0$, which makes model H right and model G wrong.

For the tuning function Λ discussed in sections 2.3 and 4.4, we consider two different choices; $\Lambda_1\left(n\hat{Q}\right) = \exp\left(n\hat{Q}\right) - 1$ and $\Lambda_2(n\hat{Q}) = (n\hat{Q})^2$ so the weighting functions $\hat{W}_g$ and $\hat{W}_f$ are

$$\Lambda_1 : \hat{W}_g = \frac{\exp\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\} - 1}{\exp\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\} + \exp\{n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\} - 2}, \hat{W}_f = 1 - \frac{1}{\exp\{n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\}}, \quad (12)$$

$$\Lambda_2 : \hat{W}_g = \frac{\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\}^2}{\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\}^2 + \{n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\}^2}, \hat{W}_f = 1 - \frac{1}{\{n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\}^2 + 1}. \quad (13)$$

For the tuning parameter τ, we use $\tau = 1 - p$, where p is the p-value of the Wald statistic as discussed in section 2.3.

We report eight estimates of α_1 and α_0 for each simulation. First is GMM based on the model G moments, denoted by GMM_g (which is only consistent if model G is right). Second is GMM

based on the H moments, denoted by GMM_h (which is only consistent if model H is right). Third is GMM based on both sets of moments, denoted by GMM_f (which is consistent, and more efficient than either the first or second set of estimates, only if both models are right). Fourth is the model averaging estimator provided by Martins and Gabriel (2014) and discussed in section 4.4, denoted by MG. Fifth and sixth are our ODR estimators in equation (1) using tuning functions Λ_1 and Λ_2, respectively, denoted by ODR_{Λ_1} and ODR_{Λ_2} (which are consistent for all designs). Seventh and eighth are our simpler estimators in equation (4), denoted by $SODR_{\Lambda_1}$ and $SODR_{\Lambda_2}$ (which are consistent for all designs, but asymptotically less efficient than ODR when both sets of moments are valid).

For each of the eight estimators, Tables 1 and 2 present simulation results of $n = 100$ observations, and Tables 3 and 4 present simulation results of $n = 500$ observations. All tables are based on 2000 Monte Carlo simulations. The reported summary statistics on the estimated parameters are, respectively, the bias (Bias), median error (MdE), root mean-squared error (RMSE), median absolute error (MAE), and the standard deviation (SD). To check the quality of our limiting distribution, we also calculate the estimated t-statistic $\widehat{\alpha}_j - 1$ divided by the estimated standard error of $\widehat{\alpha}_j$ for $j = 0, 1$ in each simulation. We report skewness (Skew) and kurtosis (Kurt) of these t-statistics across simulations, and the frequency (Freq) that these t-statistics are less than 2 in magnitude, corresponding to the frequency with which a ± 2 estimated standard error confidence interval contains the true parameter value. Also, to check the accuracy of the standard error estimates, we report the average of the estimated standard errors (SE), and standard deviation of the estimated standard errors (SD_{SE}), across the simulations. The last five summary statistics are not reported for $SODR$, because we do not consider its limiting distribution due to the random probability limit of $\hat{W}_g$.

When both sets of instruments are valid, ODR estimates are almost as precise as GMM_f, and when either set of instruments is invalid, ODR estimates are more precise than inconsistent GMM estimators. The $SODR$ estimates are found to be less efficient than ODR when both G and H

Table 1. Simulation Results of α_1 ($n = 100$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	0.0008	0.0015	0.0006	0.0170	0.0247	0.0819	3.1153	0.9390	0.0236	0.0038
GMM_h	-0.0010	0.0010	0.0023	0.0302	0.0480	0.2350	2.8924	0.9520	0.0455	0.0138
GMM_f	0.0012	0.0018	0.0005	0.0159	0.0222	0.0833	3.0008	0.9290	0.0202	0.0030
MG	-0.0010	-0.0004	0.0008	0.0184	0.0288	0.1681	2.9945	0.9535	0.0268	0.0084
ODR_{Λ_1}	0.0004	0.0012	0.0006	0.0164	0.0255	0.0829	3.0406	0.9250	0.0214	0.0049
ODR_{Λ_2}	0.0006	0.0011	0.0005	0.0149	0.0232	0.1840	3.3537	0.9285	0.0210	0.0040
$SODR_{\Lambda_1}$	-0.0016	-0.0003	0.0012	0.0200	0.0348					
$SODR_{\Lambda_2}$	-0.0011	-0.0003	0.0012	0.0201	0.0342					
G correct										
GMM_g	0.0007	0.0022	0.0006	0.0169	0.0248	0.2852	3.2493	0.9380	0.0237	0.0046
GMM_h	0.1991	0.1951	0.0413	0.1951	0.0408	0.3284	3.3054	0.0000	0.0348	0.0100
GMM_f	0.0731	0.0725	0.0059	0.0725	0.0244	0.2104	3.1962	0.0540	0.0166	0.0023
MG	0.0372	0.0284	0.0038	0.0317	0.0487	0.8547	3.0249	0.5760	0.0201	0.0054
ODR_{Λ_1}	0.0229	0.0114	0.0038	0.0207	0.0570	1.4912	4.8689	0.7730	0.0230	0.0061
ODR_{Λ_2}	0.0247	0.0130	0.0038	0.0223	0.0563	1.3465	4.3800	0.7560	0.0229	0.0059
$SODR_{\Lambda_1}$	0.0229	0.0114	0.0038	0.0207	0.0570					
$SODR_{\Lambda_2}$	0.0242	0.0122	0.0038	0.0223	0.0569					
H correct										
GMM_g	0.1123	0.1121	0.0130	0.1121	0.0201	0.3521	3.4293	0.0000	0.0163	0.0025
GMM_h	0.0003	0.0069	0.0025	0.0308	0.0498	0.6336	3.3317	0.9220	0.0465	0.0238
GMM_f	0.0938	0.0939	0.0092	0.0939	0.0193	0.3582	3.3534	0.0015	0.0145	0.0021
MG	0.0009	0.0075	0.0025	0.0309	0.0499	0.6559	3.3722	0.9165	0.0462	0.0238
ODR_{Λ_1}	0.0025	0.0080	0.0024	0.0317	0.0494	1.2264	5.4509	0.8925	0.0449	0.0214
ODR_{Λ_2}	0.0047	0.0110	0.0024	0.0320	0.0489	1.5845	7.7317	0.8800	0.0437	0.0204
$SODR_{\Lambda_1}$	0.0003	0.0070	0.0025	0.0308	0.0499					
$SODR_{\Lambda_2}$	0.0001	0.0073	0.0026	0.0311	0.0509					

Table 2. Simulation Results of α_0 ($n = 100$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	-0.0038	-0.0048	0.0112	0.0687	0.1058	0.0005	3.1738	0.9415	0.1009	0.0089
GMM_h	-0.0024	-0.0090	0.0134	0.0757	0.1157	-0.0131	2.9788	0.9490	0.1115	0.0182
GMM_f	-0.0046	-0.0073	0.0113	0.0688	0.1063	0.0212	3.1124	0.9350	0.0981	0.0085
MG	-0.0022	-0.0063	0.0115	0.0697	0.1071	0.0291	3.0642	0.9440	0.1022	0.0110
ODR_{Λ_1}	-0.0039	-0.0062	0.0113	0.0686	0.1063	0.0583	3.0524	0.9370	0.0989	0.0092
ODR_{Λ_2}	0.0001	-0.0017	0.0105	0.0687	0.1025	-0.0532	3.1744	0.9525	0.0990	0.0088
$SODR_{\Lambda_1}$	-0.0016	-0.0067	0.0120	0.0703	0.1097					
$SODR_{\Lambda_2}$	0.0014	0.0023	0.0108	0.0707	0.1041					
G correct										
GMM_g	-0.0038	-0.0060	0.0112	0.0683	0.1060	-0.0390	3.1287	0.9395	0.1009	0.0108
GMM_h	-0.2005	-0.1977	0.0554	0.1977	0.1234	0.1485	3.0509	0.5750	0.1103	0.0179
GMM_f	-0.0744	-0.0737	0.0219	0.0999	0.1280	-0.0354	3.1266	0.7540	0.0867	0.0074
MG	-0.0401	-0.0396	0.0140	0.0774	0.1115	-0.1154	3.1855	0.8885	0.0954	0.0109
ODR_{Λ_1}	-0.0258	-0.0198	0.0147	0.0722	0.1186	-0.2332	3.2476	0.9010	0.0996	0.0120
ODR_{Λ_2}	-0.0245	-0.0198	0.0136	0.0744	0.1139	-0.2004	3.0110	0.9065	0.0995	0.0114
$SODR_{\Lambda_1}$	-0.0258	-0.0198	0.0147	0.0722	0.1186					
$SODR_{\Lambda_2}$	-0.0240	-0.0194	0.0136	0.0745	0.1142					
H correct										
GMM_g	-0.1151	-0.1166	0.0230	0.1198	0.0989	0.0139	2.8983	0.6735	0.0808	0.0069
GMM_h	-0.0028	-0.0088	0.0133	0.0722	0.1153	-0.2405	2.9748	0.9530	0.1123	0.0344
GMM_f	-0.0963	-0.0966	0.0203	0.1039	0.1050	-0.0085	2.9169	0.7095	0.0791	0.0068
MG	-0.0035	-0.0094	0.0133	0.0720	0.1151	-0.2389	2.9660	0.9515	0.1120	0.0343
ODR_{Λ_1}	-0.0051	-0.0105	0.0131	0.0725	0.1146	-0.2535	2.9609	0.9475	0.1109	0.0320
ODR_{Λ_2}	-0.0084	-0.0187	0.0135	0.0753	0.1159	-0.1964	3.0290	0.9380	0.1095	0.0287
$SODR_{\Lambda_1}$	-0.0029	-0.0089	0.0133	0.0722	0.1153					
$SODR_{\Lambda_2}$	-0.0038	-0.0144	0.0138	0.0760	0.1176					

Table 3. Simulation Results of α_1 ($n = 500$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	-0.0001	0.0001	0.0001	0.0074	0.0108	0.0652	2.8494	0.9565	0.0108	0.0008
GMM_h	-0.0005	-0.0004	0.0004	0.0131	0.0199	0.0602	2.9137	0.9565	0.0200	0.0023
GMM_f	0.0000	0.0001	0.0001	0.0066	0.0096	0.0227	2.7919	0.9495	0.0094	0.0006
MG	-0.0004	-0.0007	0.0001	0.0081	0.0120	0.0009	2.7642	0.9525	0.0119	0.0024
ODR_{Λ_1}	-0.0001	0.0001	0.0001	0.0069	0.0106	0.0145	2.7364	0.9390	0.0097	0.0013
ODR_{Λ_2}	-0.0004	-0.0006	0.0001	0.0067	0.0109	0.1468	3.1553	0.9415	0.0097	0.0013
$SODR_{\Lambda_1}$	-0.0005	-0.0006	0.0002	0.0091	0.0142					
$SODR_{\Lambda_2}$	-0.0007	-0.0004	0.0002	0.0089	0.0149					
G correct										
GMM_g	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1479	2.8751	0.9560	0.0108	0.0009
GMM_h	0.1990	0.1986	0.0399	0.1986	0.0177	0.1549	3.0003	0.0000	0.0155	0.0018
GMM_f	0.0729	0.0728	0.0054	0.0728	0.0109	0.1287	3.0088	0.0000	0.0077	0.0005
MG	0.0001	0.0004	0.0001	0.0073	0.0110	0.3379	4.0679	0.9535	0.0108	0.0009
ODR_{Λ_1}	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1480	2.8743	0.9560	0.0108	0.0009
ODR_{Λ_2}	0.0010	0.0010	0.0001	0.0076	0.0115	1.2373	10.9036	0.9425	0.0107	0.0009
$SODR_{\Lambda_1}$	-0.0001	0.0002	0.0001	0.0073	0.0108					
$SODR_{\Lambda_2}$	0.0009	0.0009	0.0001	0.0076	0.0115					
H correct										
GMM_g	0.1124	0.1125	0.0127	0.1125	0.0091	0.1833	2.9687	0.0000	0.0074	0.0005
GMM_h	-0.0004	0.0006	0.0004	0.0132	0.0201	0.3063	3.0314	0.9580	0.0201	0.0034
GMM_f	0.0939	0.0937	0.0089	0.0937	0.0088	0.1908	3.0597	0.0000	0.0067	0.0004
MG	-0.0004	0.0006	0.0004	0.0132	0.0201	0.3063	3.0314	0.9580	0.0201	0.0034
ODR_{Λ_1}	-0.0004	0.0006	0.0004	0.0132	0.0201	0.3063	3.0314	0.9580	0.0201	0.0034
ODR_{Λ_2}	0.0002	0.0018	0.0004	0.0131	0.0203	0.3885	3.1614	0.9475	0.0201	0.0035
$SODR_{\Lambda_1}$	-0.0004	0.0006	0.0004	0.0132	0.0201					
$SODR_{\Lambda_2}$	0.0001	0.0017	0.0004	0.0131	0.0203					

Table 4. Simulation Results of α_0 ($n = 500$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	-0.0010	-0.0002	0.0021	0.0315	0.0458	-0.1391	2.9732	0.9565	0.0459	0.0018
GMM_h	-0.0008	0.0005	0.0024	0.0328	0.0492	-0.1701	3.0631	0.9500	0.0491	0.0030
GMM_f	-0.0011	0.0000	0.0021	0.0311	0.0458	-0.1335	2.9799	0.9550	0.0454	0.0017
MG	-0.0007	0.0004	0.0021	0.0311	0.0463	-0.1527	3.0340	0.9570	0.0462	0.0021
ODR_{Λ_1}	-0.0010	0.0000	0.0021	0.0310	0.0459	-0.1327	3.0063	0.9540	0.0455	0.0018
ODR_{Λ_2}	0.0009	-0.0005	0.0022	0.0315	0.0471	0.0061	2.9664	0.9445	0.0455	0.0019
$SODR_{\Lambda_1}$	-0.0005	0.0003	0.0022	0.0321	0.0468					
$SODR_{\Lambda_2}$	0.0010	0.0003	0.0023	0.0334	0.0483					
G correct										
GMM_g	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1566	2.9735	0.9570	0.0459	0.0021
GMM_h	-0.2000	-0.2000	0.0428	0.2000	0.0529	0.0663	3.1573	0.0225	0.0495	0.0033
GMM_f	-0.0732	-0.0731	0.0084	0.0739	0.0554	0.0501	2.9813	0.5400	0.0402	0.0014
MG	-0.0012	-0.0004	0.0021	0.0314	0.0458	-0.1553	2.9705	0.9570	0.0458	0.0021
ODR_{Λ_1}	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1563	2.9744	0.9570	0.0459	0.0021
ODR_{Λ_2}	-0.0020	-0.0011	0.0021	0.0315	0.0459	-0.1685	2.9918	0.9550	0.0457	0.0021
$SODR_{\Lambda_1}$	-0.0010	-0.0003	0.0021	0.0314	0.0458					
$SODR_{\Lambda_2}$	-0.0020	-0.0011	0.0021	0.0315	0.0459					
H correct										
GMM_g	-0.1122	-0.1121	0.0146	0.1121	0.0448	-0.0037	3.0575	0.1945	0.0367	0.0013
GMM_h	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
GMM_f	-0.0938	-0.0948	0.0111	0.0948	0.0481	-0.0661	2.9792	0.3445	0.0366	0.0013
MG	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
ODR_{Λ_1}	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
ODR_{Λ_2}	-0.0011	-0.0038	0.0025	0.0340	0.0500	-0.1804	2.9318	0.9555	0.0491	0.0049
$SODR_{\Lambda_1}$	-0.0007	-0.0007	0.0024	0.0329	0.0494					
$SODR_{\Lambda_2}$	-0.0011	-0.0037	0.0025	0.0340	0.0500					

models are valid (as expected), but when one model is invalid, $SODR$ is similar to ODR. In this application, the cost in efficiency of choosing the simpler $SODR$ seems small[9]. Presumably the gains to ODR would have been larger in a simulation design where the efficiency of GMM_f more greatly exceeded that of GMM_g.

Despite the fact that MG is specifically designed for instrument selection in linear models, while our ODR is a generic estimator for arbitrary moment based models, the finite sample performance of ODR is close to, and in some cases slightly better than, MG, particularly when both models are correctly specified.

Our simulation results also show that the limiting distributions provide reasonably good approximations to their finite sample counterparts, and these approximations improve substantially when going from the sample size $n = 100$ to $n = 500$. In particular, the quality of ODR estimated standard errors and confidence intervals is similar to that of the corresponding correctly specified GMM standard errors and confidence intervals. This can be seen by comparing the SE and SD columns, and comparing how close Freq is to .95 in the ODR rows, relative to same comparisons in the correctly specified GMM rows. Indeed, at $n = 500$ almost all of the summary statistics of ODR become close to those of the most efficient correctly specified GMM in each block. One exception is ODR_{Λ_2} when the model H is invalid. In this case, there were a few large outlier ODR_{Λ_2} estimates, resulting in substantial nonnormal skewness and kurtosis in the t-statistic distribution. But other summary statistics are still similar to those of ODR_{Λ_1} and correctly specified GMM. This suggests a modest advantage of the exponential tuning function Λ_1.

One should expect correctly specified GMM estimators to be more efficient than ODR, and that is indeed the case. But in many of the simulations, the loss in efficiency from using ODR is very low. In particular, when model G is invalid, so only the weaker instruments are valid, the precision of ODR is almost identical to that of the efficient GMM_h. So, using our ODR, there is little loss in efficiency from not knowing which specification is correct. In summary, we conclude

[9]However, $SODR$ incurs the additional cost of possibly not having a normal limiting distribution when both G and H are correctly specified.

that our proposed ODR works well, even at low sample sizes.

6 Empirical Application: Engel Curve Estimation

Here we empirically estimate the Engel curve example discussed in section 3.2. Y is the food budget share, S is log real total consumption expenditures, and X is a vector of other covariates that serve as controls[10]. The goal is estimation of the coefficient of S in a regression of Y on S and X. Total consumption S is observed with measurement error, so instrumental variables estimation is used to correct for the resulting endogeneity. The vector L consists of two candidate external instrument variables, real total income and real total income squared. Model G assumes these external instruments are valid. Model H instead assumes that constructed instruments based on heteroscedasticity as described by Lewbel (2012) and summarized in section 3.2 above are valid. Model F assumes boths sets of instruments are valid.

The data consist of 854 households collected from the UK Family Expenditure Survey 1980-1982 as studied by Banks, Blundell, and Lewbel (1997), Lewbel (2012), and Baum and Schaffer (2012). The sample means are $\overline{Y} = 0.285$ and $\overline{S} = 0.599$, and the standard deviations are 0.106 for Y and 0.410 for S.

The parameter of interest is the coefficient of log real total expenditure α_s. Table 5 summarizes estimates of α_s and of the constant term α_0. GMM_{g0} is the estimate reported in Lewbel (2012) and Baum and Schaffer (2012). GMM_g is the GMM estimator using the moments in equation (8), which makes use of the external instruments L.[11] GMM_h is the GMM estimator that uses the moments in equation (9), which are heteroscedasticity based constructed instruments. GMM_f is the GMM estimator that uses both sets of instruments, and $SODR$ and ODR are our new estimators given

[10] These covariates are a constant, age, spouse's age, squared ages, seasonal dummies, and dummies for spouse working, gas central heating, ownership of a washing machine, one car, and two cars.

[11] The estimates of GMM_{g0} and GMM_g are not identical because we use the two external instruments income and income squared, instead of just using income. There's a similar small difference between GMM_f and the models based on both sets of moments reported in Lewbel (2012) and Baum and Schaffer (2012), for the same reason.

in equations (4) and (1) with the tuning functions Λ_1 and Λ_2..

The estimated results show that the external instruments of model G are much stronger than the constructed instruments of model H. This is not surprising since the constructed instruments are based on higher moments of the data. This difference in strength can be seen in the standard errors of $\hat{\alpha}_s$, which are much lower in model G than in model H, and also in model GMM_f which gives estimates much closer to GMM_g than GMM_h.

The point estimates of GMM_g and GMM_h are substantially different, which could be due to having one of these sets of instruments be invalid. However, this difference could also just be due to imprecision, particularly of GMM_h. This illustrates the usefulness of our ODR, which does not require resolving which set of instruments is valid, or if both are valid.

Table 5. Engel Curve Estimates

	GMM_{g0}	GMM_g	GMM_h	GMM_f	$SODR_{\Lambda_1}$	ODR_{Λ_1}	$SODR_{\Lambda_2}$	ODR_{Λ_2}
$\hat{\alpha}_s$	-0.0859 (0.0198)	-0.0840 (0.0197)	-0.0521 (0.0546)	-0.0862 (0.0177)	-0.0812	-0.0862 (0.0192)	-0.0831	-0.0862 (0.0192)
$\hat{\alpha}_0$	0.336 (0.0122)	0.335 (0.0120)	0.317 (0.0328)	0.337 (0.0109)	0.333	0.337 (0.0118)	0.335	0.337 (0.0118)
χ^2		0.191	12.91	15.94				
$d.f.$		1	11	13				
p-value		0.662	0.299	0.252				
$\hat{Q}$		0.0002	0.0014	0.0014				
$\hat{W}_g, \hat{W}_f, p$					0.09, 0.004, 0.86		0.03, 0.000, 0.86	

[12]

[12] Table 5 notes: We report coefficient estimates with associated standard errors in parentheses, except SODR. Also reported is χ^2, the Hansen (1982) test statistics for overidentified GMM, along with their degrees of freedom $d.\hat{f}.$ and p-values. $\hat{Q}$ is the normalized minimand of the GMM estimators. The last row reports weights $\hat{W}_g$, $\hat{W}_f$, and gives p, which is the p-value of the Wald statistic testing the null hypothesis that $\widehat{\alpha}_g = \widehat{\alpha}_h$. This p is used to construct $\tau = 1 - p$ in $\hat{W}_f$ in equation (5), as explained in section 2.3.

The estimated weight $\hat{W}_g$ is 0.09 with the tuning function Λ_1 and 0.03 with Λ_2, so $SODR$ puts over ten times as much weight on model G as on model H. However, in ODR the weight on model F, $1 - \hat{W}_f$, is 0.996 with Λ_1 and one to three decimal places with Λ_2. The very small difference in $\hat{W}_f$ between Λ_1 and Λ_2 is why both of the ODR estimates appear the same in Table 5 (they actually differ in the fourth significant digit: -0.08617 vs. -0.08619 for $\hat{\alpha}_s$).

The very high weight on model F strongly suggests that both models are likely to be correctly specified. This therefore implies that the difference between GMM_g and GMM_h is likely due to imprecision of GMM_h rather than misspecification of the constructed instruments in model H. Further evidence that both are correctly specified is given by the chi-squared statistics in Table 5, which test validity of the moments comprising each of the GMM estimates. This situation, where both models appear to be correctly specified, is when we would expect ODR to perform better than $SODR$.

Lewbel (2012) observes that a virtue of the constructed instruments is that they are valid under very different conditions than those required for validity of the external instruments, and suggests that they therefore are useful for testing overidentification. Our proposed ODR estimator makes further use of these instruments, by delivering estimates that are consistent if either (or both) sets of instruments are valid.

7 Local Misspecification

Consider the case where model G or H is locally misspecified with the parameter in the data generating process being $\theta^g = \theta_0^g + \delta_g n^{-s}$ or $\theta^h = \theta_0^h + \delta_h n^{-s}$ for constants δ_g and δ_h, and $s > 0$. Note $s = 0$ is equivalent to global misspecification, while $s = \infty$ is equivalent to correct specification, which are the cases we have already considered in our previous theorems. Pitman (1949) drift corresponds to the case of $s = 1/2$. This model is used by, e.g. Newey and West (1987), Bera and Yoon (1993) and Newey and McFadden (1994) to develop local power analyses. Here we summarize the asymptotic properties of our ODR estimator under local misspecification, with formal results

provided in Appendix II.

The asymptotic distribution of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ depends on the value of s. We show in Appendix II that the influence function of our ODR estimator consists of three terms; the first is the weighted sum of three different well behaved influence functions, the second converges to zero in probability for all $s \geq 0$, and the third either converges to a constant or diverges depending on s (and sometimes τ) as discussed below.[13]

First suppose model G is locally misspecified with $s > 1/2$. Then $n\widetilde{Q}^g\left(\widehat{\alpha}_g, \widehat{\beta}_g\right) \rightarrow^d \chi^2_{k_g}(0)$, which is the same limit as when G is correctly specified, and similarly for H. As a result, in this case the SODR and ODR estimators have the same $\sqrt{n}$ consistent, asymptotically normal limiting distribution as they have when G is correctly specified, and similarly for H. Note this means that instead of requiring that either G or H (or both) be correctly specified, it is sufficient to assume that either G or H (or both) are locally misspecified with $s > 1/2$, noting that correct specification is the special case of $s = \infty$.

If model G is locally misspecified with $s < 1/2$, then $n\widetilde{Q}^g\left(\widehat{\alpha}_g, \widehat{\beta}_g\right)$ diverges, and the SODR has the same $\sqrt{n}$ consistent, asymptotically normal limiting distribution as when G is globally misspecified. The ODR will also have the same limiting distribution as when G is globally misspecified, as long as the tuning parameter τ has $\tau > s + 0.5$. This then guarantees that model G will asymptotically have zero weight. Since these cases are equivalent asymptotically to G being globally misspecified, we need to assume that H is either correctly specified, or locally misspecified with its $s > 1/2$. This generalizes our original theorems that simply assumed either G or H is correctly specified.

Finally, suppose model G is locally misspecified with $s = 1/2$. Then $n\widetilde{Q}^g$ converges to a noncentral chi-squared distribution. Specifically, $n\widetilde{Q}^g\left(\widehat{\alpha}_g, \widehat{\beta}_g\right) \rightarrow^d \chi^2_{k_g}(\omega_g' \Omega_g^{1/2} \Pi_g^* \Omega_g^{1/2} \omega_g)$, where the object in parentheses is the noncentrality parameter and the definitions of Π_g^* and ω_g are given in equation

[13] In Appendix II we also explicitly derive the implications of these results for the limiting distribution of the ODR estimator when one model is correctly specified and the other is locally misspecified for varying values of s. The results summarized in this subsection are all either directly verified in Appendix II, or are immediate extentions.

(17) and at the beginning of Appendix II, respectively. In this case the GMM estimator of model G is consistent but not $\sqrt{n}$ consistent, as established in, e.g., Newey and McFadden (1994). Here $n\widetilde{Q}^g$ is still bounded in probability, so ODR will asymptotically put weight on either model G or, if H is correctly specified (or locally misspecified with its $s > 1/2$) on model F, which then is consistent but may not be $\sqrt{n}$ consistent. As a result, in this knife edge case, ODR will be consistent, but not $\sqrt{n}$ consistent.

The main results here can be summarized as follows. If both G and H are locally misspecified, each with $s > 1/2$ (including the special case where one or both is correctly specified, corresponding to $s = \infty$), then ODR will have the same limiting distribution as efficient GMM with both G and H correctly specified. If just G is locally misspecified with $s > 1/2$ (again including as a special case having G be correctly specified by $s = \infty$), and H is either misspecified or locally misspecified with $s < 1/2$, then (assuming $\tau > s + 0.5$) ODR will have the same limiting distribution as efficient GMM based just on model G (and vice versa, exchanging the roles of G and H). Equivalently we can say that our earlier Theorem 2 still holds, replacing "correctly specified model" with "locally misspecified model having any $s > 1/2$, including $s = \infty$" and replacing "incorrectly specified model" with "locally misspecified model having any $s < 1/2$, including $s = 0$."

We conclude this section with some Monte Carlo results (reported in Tables 6 to 7 below), which we find support these conclusions. We use the same simulation designs and estimators as in section 5 but with a drift parameter s for the locally misspecified cases. Since ODR performed better with the tuning function Λ_1 in section 5, to save space we only report ODR_{Λ_1}, along with GMM_g, GMM_h, and GMM_f for comparison. In all these tables, model H is either globally mispecified, or locally misspecified with s equal to 0.25, 0.50, or 0.75. In Tables 6-1 and 6-2 model G is correctly specified, while in Tables 7-1 and 7-2, G is locally misspecified with $s = 0.75$.

The finite sample results in these tables largely accord with asymptotic theory, with one interesting difference. When model H is locally misspecified with $s = 0.5$ (Pitman drift) our ODR should be comparable to GMM_f, but actually performs slightly better than GMM_f. This is due to our

use of the Wald statistic to select τ. With $s = 0.5$, the Wald statistic over-rejects the null, making τ large and therefore pulling the ODR estimator towards to GMM_g, which is better behaved than GMM_f with Pitman drift.

8 Extension: Multiple Robustness

It is possible to construct triply and higher multiply robust estimators that are similar to $SODR$. Suppose we have a third model, called model L, with GMM objective function $\hat{Q}^l(\alpha, \lambda)$. The GMM estimator of model L is $\{\widehat{\alpha}_l, \widehat{\lambda}_l\} = \arg\min_{\{\alpha,\lambda\}\in\Theta_\alpha\times\Theta_\lambda} \hat{Q}^l(\alpha, \lambda)$. A possible formula for triply robust estimation of α would then be the weighted average

$$\widetilde{\alpha} = \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\widehat{\alpha}_l + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\widehat{\alpha}_g + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\widehat{\alpha}_h}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)}. \tag{14}$$

In equation (14), the weight on $\widehat{\alpha}_l$ 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 $SODR$ estimator. For example, if model G is right and models L 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 L is wrong. Then all the weights in both the numerator and denominator of equation (14) 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 $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ is finite and nonzero, so the limiting weights on $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$ will be nonzero while the limiting weight on $\widehat{\alpha}_l$ will be zero, as desired.

As with $SODR$, the limiting distribution of the triply robust estimator $\widetilde{\alpha}$ in equation (14) is complicated by the potential limiting randomness of ratios like $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ in the weights. In the doubly robust case, we avoided this problem in ODR by using the additional weight W_f for when both models are correctly specified. An analogous construction for triply robust estimation would be more complicated, since we would also need to consider the cases where any pair of models is correct, and when all three are correct. This would require at least constructing

Table 6-1. Model G is Correctly Specified and Model H is Misspecified ($n = 500$)

α_1	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	0.0002	0.0006	0.0001	0.0075	0.0111	0.2310	3.1966	0.9465	0.0108	0.0011
GMM_h	0.2374	0.2367	0.0566	0.2367	0.0157	0.1558	3.1392	0.0000	0.0139	0.0016
GMM_f	0.1094	0.1094	0.0121	0.1094	0.0112	0.0817	3.0557	0.0000	0.0068	0.0005
ODR_{Λ_1}	0.0002	0.0006	0.0001	0.0075	0.0111	0.2311	3.1963	0.9460	0.0108	0.0011
s=0.5										
GMM_g	0.0002	0.0006	0.0001	0.0075	0.0110	0.1255	3.0813	0.9535	0.0108	0.0008
GMM_h	0.0827	0.0822	0.0071	0.0822	0.0174	-0.0439	3.1104	0.0045	0.0175	0.0019
GMM_f	0.0220	0.0223	0.0006	0.0223	0.0093	0.0825	3.0819	0.3365	0.0090	0.0006
ODR_{Λ_1}	0.0128	0.0058	0.0008	0.0102	0.0259	0.9210	3.2417	0.7455	0.0109	0.0019
s=0.75										
GMM_g	-0.0001	0.0001	0.0001	0.0074	0.0108	0.0707	2.8505	0.9570	0.0108	0.0008
GMM_h	0.0181	0.0180	0.0007	0.0194	0.0192	0.0233	2.9125	0.8355	0.0193	0.0021
GMM_f	0.0044	0.0045	0.0001	0.0074	0.0095	0.0275	2.7750	0.9270	0.0094	0.0006
ODR_{Λ_1}	0.0058	0.0052	0.0002	0.0081	0.0123	0.0457	2.7501	0.8905	0.0099	0.0016
Global										
GMM_g	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1479	2.8751	0.9560	0.0108	0.0009
GMM_h	0.1990	0.1986	0.0399	0.1986	0.0177	0.1549	3.0003	0.0000	0.0155	0.0018
GMM_f	0.0729	0.0728	0.0054	0.0728	0.0109	0.1287	3.0088	0.0000	0.0077	0.0005
ODR_{Λ_1}	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1480	2.8743	0.9560	0.0108	0.0009

Table 6-2. Model G is Correctly Specified and Model H is Misspecified ($n = 500$)

α_0	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	-0.0005	-0.0008	0.0022	0.0319	0.0467	0.0210	2.8965	0.9530	0.0459	0.0025
GMM_h	-0.2385	-0.2364	0.0598	0.2364	0.0544	-0.0138	2.8922	0.0030	0.0504	0.0033
GMM_f	-0.1108	-0.1099	0.0166	0.1099	0.0654	-0.0509	2.9047	0.2860	0.0369	0.0013
ODR_{Λ_1}	-0.0005	-0.0008	0.0022	0.0319	0.0467	0.0210	2.8964	0.9530	0.0459	0.0025
s=0.5										
GMM_g	-0.0016	-0.0012	0.0022	0.0324	0.0467	-0.0374	2.9635	0.9515	0.0459	0.0019
GMM_h	-0.0838	-0.0827	0.0092	0.0827	0.0470	-0.0797	2.9642	0.5840	0.0465	0.0026
GMM_f	-0.0234	-0.0226	0.0028	0.0359	0.0473	-0.0714	2.9394	0.8995	0.0441	0.0017
ODR_{Λ_1}	-0.0141	-0.0131	0.0027	0.0346	0.0503	-0.1517	3.0821	0.9175	0.0455	0.0020
s=0.75										
GMM_g	-0.0010	-0.0002	0.0021	0.0314	0.0458	-0.1404	2.9735	0.9565	0.0459	0.0018
GMM_h	-0.0193	-0.0186	0.0027	0.0350	0.0483	-0.1429	3.0438	0.9355	0.0482	0.0028
GMM_f	-0.0054	-0.0044	0.0021	0.0313	0.0456	-0.1263	2.9815	0.9540	0.0452	0.0017
ODR_{Λ_1}	-0.0069	-0.0055	0.0022	0.0314	0.0461	-0.1310	3.0076	0.9490	0.0453	0.0017
Global										
GMM_g	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1566	2.9735	0.9570	0.0459	0.0021
GMM_h	-0.2000	-0.2000	0.0428	0.2000	0.0529	0.0663	3.1573	0.0225	0.0495	0.0033
GMM_f	-0.0732	-0.0731	0.0084	0.0739	0.0554	0.0501	2.9813	0.5400	0.0402	0.0014
ODR_{Λ_1}	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1563	2.9744	0.9570	0.0459	0.0021

Table 7-1. Model G is Misspecified with $s = 0.75$ and Model H is Misspecified ($n = 500$)

α_1	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	0.0088	0.0093	0.0002	0.0102	0.0104	0.1033	3.0590	0.8355	0.0102	0.0010
GMM_h	0.2297	0.2292	0.0530	0.2292	0.0148	0.2036	2.9334	0.0000	0.0130	0.0015
GMM_f	0.1112	0.1108	0.0125	0.1108	0.0108	-0.1158	3.3286	0.0000	0.0065	0.0004
ODR_{Λ_1}	0.0088	0.0093	0.0002	0.0102	0.0104	0.1064	3.0616	0.8355	0.0102	0.0010
s=0.5										
GMM_g	0.0086	0.0087	0.0002	0.0098	0.0109	0.1853	3.1186	0.8600	0.0105	0.0008
GMM_h	0.0807	0.0805	0.0068	0.0805	0.0178	-0.0272	2.9853	0.0090	0.0171	0.0017
GMM_f	0.0276	0.0277	0.0009	0.0277	0.0095	0.1450	3.2462	0.1590	0.0088	0.0006
ODR_{Λ_1}	0.0222	0.0155	0.0011	0.0161	0.0254	0.6447	2.7902	0.6020	0.0108	0.0021
s=0.75										
GMM_g	0.0090	0.0090	0.0002	0.0101	0.0105	0.0550	3.0231	0.8565	0.0106	0.0008
GMM_h	0.0181	0.0185	0.0007	0.0198	0.0198	0.0769	2.8749	0.8185	0.0192	0.0022
GMM_f	0.0113	0.0113	0.0002	0.0115	0.0092	0.0496	3.1639	0.7720	0.0092	0.0006
ODR_{Λ_1}	0.0125	0.0120	0.0003	0.0123	0.0115	0.0250	3.0848	0.7445	0.0098	0.0018
Global										
GMM_g	0.0089	0.0092	0.0002	0.0102	0.0106	0.1271	3.0891	0.8520	0.0103	0.0009
GMM_h	0.1939	0.1926	0.0379	0.1926	0.0167	0.1383	3.1092	0.0000	0.0146	0.0017
GMM_f	0.0768	0.0766	0.0060	0.0766	0.0101	0.0503	2.8409	0.0000	0.0075	0.0005
ODR_{Λ_1}	0.0089	0.0092	0.0002	0.0102	0.0106	0.1241	3.0766	0.8500	0.0103	0.0009

Table 7-2. Model G is Misspecified with $s = 0.75$ and Model H is Misspecified ($n = 500$)

α_0	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	-0.0083	-0.0071	0.0022	0.0315	0.0458	-0.1606	2.8672	0.9475	0.0445	0.0023
GMM_h	-0.2309	-0.2292	0.0560	0.2292	0.0524	-0.0060	3.0920	0.0030	0.0485	0.0032
GMM_f	-0.1115	-0.1098	0.0166	0.1098	0.0647	-0.0952	2.9328	0.2735	0.0363	0.0013
ODR_{Λ_1}	-0.0083	-0.0071	0.0022	0.0315	0.0458	-0.1605	2.8666	0.9475	0.0445	0.0023
s=0.5										
GMM_g	-0.0090	-0.0087	0.0022	0.0317	0.0455	-0.0878	2.9419	0.9485	0.0453	0.0018
GMM_h	-0.0811	-0.0807	0.0087	0.0807	0.0457	-0.0785	2.9850	0.5940	0.0459	0.0024
GMM_f	-0.0281	-0.0278	0.0029	0.0369	0.0455	-0.0613	2.9619	0.8930	0.0437	0.0016
ODR_{Λ_1}	-0.0225	-0.0204	0.0030	0.0351	0.0497	-0.2171	3.1523	0.9000	0.0449	0.0019
s=0.75										
GMM_g	-0.0100	-0.0091	0.0021	0.0321	0.0448	-0.0071	2.9514	0.9535	0.0455	0.0018
GMM_h	-0.0189	-0.0199	0.0027	0.0346	0.0481	-0.0238	2.9757	0.9310	0.0481	0.0029
GMM_f	-0.0122	-0.0122	0.0021	0.0318	0.0446	-0.0059	2.9843	0.9450	0.0449	0.0017
ODR_{Λ_1}	-0.0133	-0.0130	0.0022	0.0330	0.0453	0.0016	2.9476	0.9410	0.0450	0.0018
Global										
GMM_g	-0.0106	-0.0117	0.0021	0.0319	0.0450	-0.0511	2.9491	0.9475	0.0448	0.0020
GMM_h	-0.1952	-0.1941	0.0407	0.1941	0.0513	0.1066	3.2575	0.0200	0.0479	0.0031
GMM_f	-0.0785	-0.0784	0.0092	0.0785	0.0549	-0.0661	2.9737	0.5035	0.0396	0.0014
ODR_{Λ_1}	-0.0106	-0.0118	0.0021	0.0319	0.0450	-0.0508	2.9492	0.9475	0.0448	0.0020

an ODR for each of the three possible pairs of models, and for the model that combines all three.

9 Conclusions

In this paper, we provide a general technique for constructing doubly robust estimators. Our Over-identified Doubly Robust (ODR) technique is a simple extension of the Generalized Method of Moments. It takes the form of a weighted average of Hansen's (1982) Generalized Method of Moments (GMM) based estimators, and has similar associated root-n asymptotics. The proposed estimator appears to work well in a small Monte Carlo study and in an empirical application to instrumental variables estimation, where either one of two sets of instrument vectors might be invalid.

Our estimator requires that the candidate models be over-identified, having more moments than parameters. Ideally the number of moments should not greatly exceed the number of parameters, because GMM can suffer from well known finite sample biases when models have many more moments than parameters, and particularly when some moments might be weak. In such cases, it may be desirable to let models G and H equal just a subset of the available moments for each. Existing moment selection methods such as Andrews and Lu (2001), Caner (2009), or Liao (2013) might be used prior to applying ODR, though this then introduces pretest bias that ODR is intended to avoid. A potential subject for future work could be more formally modifying ODR to deal with many moments and/or with weak moments.

Another potential extension for future work is to consider cases where β and γ are infinite dimensional, e.g., where models G and H may contain unknown functions, perhaps replacing unconditional expectations with conditional expectations as in Ai and Chen (2003). One difficulty in such extensions is guaranteeing that the model is still over-identified regarding α, or more precisely, ensuring that no solution to all the moment conditions exists if the model is misspecified. Chen and Santos (2018) might be helpful regarding this point. Another issue would be ensuring that the objective functions used in constructing weights remain comparable and well behaved.

10 References

Ai, C. and Chen, X. (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", *Econometrica*, 71(6), 1795-1843.

Andrews, D.W.K., (1999): "Consistent moment selection procedures for generalized method of moments estimation", *Econometrica,* 67(3), 543-564.

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

Banks, J., Blundell, R., and Lewbel, A. (1997): "Quadratic Engel Curves and Consumer Demand", *Review of Economics and Statistics*, 79(4), 527-539.

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.

Bera, A. and Yoon, M. (1993): "Specification Testing with Locally Misspecified Alternatives", *Econometric Theory*, 9(4), 649-658.

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.

Caner, M. (2009): "Lasso-type GMM Estimator", *Econometric Theory*, 25(1), 270-290.

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.

Chen, X. and Santos, A. (2018): "Overidentification in Regular Models", *Econometrica*, 86(5), 1771-1817.

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

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.

Hall, A.R. (2000): "Covariance Matrix Estimation and the Power of the Overidentifying Restrictions Test", *Econometrica,* 68(6), 1517-1528.

Hall, A.R. and Inoue, A. (2003): "The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models", *Journal of Econometrics,* 114(2), 361-394.

Hansen, B. (2007): "Least Squares Model Averaging", *Econometrica*, 75(4), 1175-1189.

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

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

Kim, J-Y, (2002): "Limited information likelihood and Bayesian analysis", *Journal of Econometrics*, 107(1-2), 175-193.

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

Lee, M.J., and Lee, S. (2019): "Double Robustness Without Weighting", *Statistics and Probability Letters*, 146, 175-180.

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

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

Newey, W. and West, K. (1987): "Hypothesis Testing with Efficient Method of Moments Testing", *International Economic Review*, 28(3), 777-787.

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

Pitman, E.T.G. (1949): ""Notes on Nonparametric Statistical Inference", Manuscript.

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 I

Recall $\widehat{\alpha} = \hat{W}_f \hat{W}_g \widehat{\alpha}_h + \hat{W}_f \left(1 - \hat{W}_g\right) \widehat{\alpha}_g + (1 - \hat{W}_f)\widehat{\alpha}_f$. To avoid confusion, we collect our notation here. The sample and population moments are

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

$$g_0(\alpha,\beta) \equiv E\{G(Z,\alpha,\beta)\}, \qquad h_0(\alpha,\gamma) \equiv E\{H(Z,\alpha,\gamma)\}, \qquad f_0(\alpha,\beta,\gamma) \equiv E\{F(Z,\alpha,\beta,\gamma)\}.$$

The true and pseudo-true parameters are ($\theta^j = \theta_0^j$ if the model is correct, $j = g, h, f$)

$$\theta_0^g \equiv \{\alpha_0, \beta_0\},\ \theta_0^h \equiv \{\alpha_0, \gamma_0\},\ \theta_0^f \equiv \{\alpha_0, \beta_0, \gamma_0\}, \qquad \theta^g \equiv \{\alpha_g, \beta_g\},\ \theta^h \equiv \{\alpha_h, \gamma_h\},\ \theta^f \equiv \{\alpha_f, \beta_f, \gamma_f\},$$

$$c_g \equiv g_0(\theta^g) \neq 0 \text{ if } \theta^g \neq \theta_0^g, \qquad c_h \equiv h_0(\theta^h) \neq 0 \text{ if } \theta^h \neq \theta_0^h, \qquad c_f \equiv f_0(\theta^f) \neq 0 \text{ if } \theta^f \neq \theta_0^f.$$

With $\hat{\Omega}_g \to^p \Omega_g$ and $\hat{\Omega}_h \to^p \Omega_h$,

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

$$\{\alpha_g, \beta_g\} \text{ minimizes } \widetilde{Q}_0^g(\alpha,\beta) \equiv g_0(\alpha,\beta)'\Omega_g g_0(\alpha,\beta), \quad \{\alpha_h, \gamma_h\} \text{ minimizes } \widetilde{Q}_0^h(\alpha,\gamma) \equiv h_0(\alpha,\gamma)'\Omega_h h_0(\alpha,\gamma);$$

$$\hat{Q}^g(\alpha,\beta) \equiv \frac{\widetilde{Q}^g(\alpha,\beta)}{k_g}, \hat{Q}^h(\alpha,\gamma) \equiv \frac{\widetilde{Q}^h(\alpha,\gamma)}{k_h},\ \hat{Q}^g \equiv \hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g), \hat{Q}^h \equiv \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h),\ Q_0^g \equiv \frac{c_g'\Omega_g c_g}{k_g}, Q_0^h \equiv \frac{c_h'\Omega_h c_h}{k_h};$$

$$\hat{W}_g \equiv \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)} \qquad \text{and} \qquad \hat{W}_f \equiv 1 - \frac{1}{n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) + 1}.$$

Proof of Lemma 1.

To obtain the probability limits of $\hat{W}_g$ and $\hat{W}_f$, first we consider without loss of generality the probability limit of $\hat{Q}^g$ when model G is correctly specified, and when it's misspecified. The asymptotics for $\hat{Q}^h$ and $\hat{Q}^f$ are obtained following the same logic. After these derivations, we then obtain the probability limits of $\hat{W}_g$ and $\hat{W}_f$ based on $\hat{Q}^g$, $\hat{Q}^h$ and $\hat{Q}^f$. First we have

$$n\hat{Q}^g = \{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)\}'\{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)\}\frac{1}{k_g}. \tag{15}$$

From the first order condition for $\hat{\theta}^g$ minimizing $\widetilde{Q}^g(\theta)$, we have

$$\sqrt{n}\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\widehat{g}(\hat{\theta}^g)=0.$$

Taylor-expanding the last term $\widehat{g}(\hat{\theta}^g)$ around θ^g gives

$$\begin{aligned}0&=\sqrt{n}\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\{\widehat{g}(\theta^g)+\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{\theta}^g-\theta^g)\}\\&=\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\sqrt{n}\widehat{g}(\theta^g)+\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g-\theta^g)\end{aligned}$$

where $\overline{\theta}^g$ is a mean value between θ^g and $\hat{\theta}^g$. If the model is correctly specified, $\theta^g=\theta_0^g$. This gives

$$\sqrt{n}(\hat{\theta}^g-\theta^g)=-(\hat{H}^g)^{-1}\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\sqrt{n}\widehat{g}(\theta^g)\quad\text{where}\quad \hat{H}^g\equiv\nabla_\theta\widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\nabla_{\theta'}\widehat{g}(\overline{\theta}^g).\tag{16}$$

Case i). Suppose that G is correctly specified. By Assumption A1, A2, A3, A5, and A6, the conditions of Theorem 2.1 of in Newey and McFadden (1994) (uniqueness, compactness, continuity, and uniform convergence) hold for GMM estimation of model G, so that $\hat{\theta}^g\to^p\theta_0^g$. For $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ in equation (15), expanding $\widehat{g}$ around θ_0^g, we have

$$\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)=\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)+\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g-\theta_0^g)$$

where $\overline{\theta}^g$ is a mean value between θ_0^g and $\hat{\theta}^g$. Plug equation (16) with θ^g replaced by θ_0^g into this equation to get

$$\begin{aligned}&\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)=\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)-\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{H}^g)^{-1}\nabla_\theta\widehat{g}(\hat{\theta}^g)\hat{\Omega}_g\sqrt{n}\widehat{g}(\theta_0^g)\\&=\{I_{\widetilde{k}_g}-\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{H}^g)^{-1}\nabla_\theta\widehat{g}(\hat{\theta}^g)\hat{\Omega}_g^{1/2}\}\cdot\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)=\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\\&\text{where}\quad\hat{\Pi}_g^*\equiv I_{\widetilde{k}_g}-\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{H}^g)^{-1}\nabla_\theta\widehat{g}(\hat{\theta}^g)\hat{\Omega}_g^{1/2}\end{aligned}\tag{17}$$

and $I_{\widetilde{k}_g}$ is the $\widetilde{k}_g\times\widetilde{k}_g$ identity matrix and $\widetilde{k}_g$ is the number of moments in the model G.

Under Assumption A7, A9, A10, A11 and A12, $\sqrt{n}\widehat{g}(\theta_0^g)\to^d N(0,\Sigma_g)$ where $\Sigma_g=E\{G(Z,\theta_0^g)G(Z,\theta_0^g)'\}$, and with $\Omega_g^{-1}=\Sigma_g$, $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\to^d N(0,I_{\widetilde{k}_g})$. By Assumption A11, $\nabla_\theta\widehat{g}(\overline{\theta}^g)\to^p\nabla_\theta g_0(\theta_0^g)$, $\nabla_\theta\widehat{g}(\hat{\theta}^g)\to^p\nabla_\theta g_0(\theta_0^g)$, and $\hat{H}^g\to^p H^g$ which is non-singular by Assumption A8. Then, we have

$$\hat{\Pi}_g^*\to^p\Pi_g^*\equiv I_{\widetilde{k}_g}-\Omega_g^{1/2}\nabla_{\theta'}g_0(\theta_0^g)(H^g)^{-1}\nabla_\theta g_0(\theta_0^g)\Omega_g^{1/2}$$

where $\hat{\Pi}_g^*$ is a $\widetilde{k}_g \times \widetilde{k}_g$ symmetric matrix that is idempotent with $trace(\Pi_g) = k_g$, $k_g \equiv \widetilde{k}_g - k_g^*$ where k_g^* is the number of parameters in the model G. Therefore,

$$n\hat{Q}^g = \{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\}'\hat{\Pi}_g^*\{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\}/k_g \to^d \chi^2_{k_g}/k_g.$$

Case ii). Suppose that G is misspecified. Under Assumption A1, A3, A4, A5, and A6, $\hat{\theta}^g \to^p \theta^g \neq \theta_0^g$ by Lemma 1 of Hall (2000). For $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ in (15), Taylor-expand $\widehat{g}$ around θ^g to get, with $c_g \equiv g_0(\alpha_g, \beta_g)$

$$\begin{aligned}\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g) &= \hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta^g) + \hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g - \theta^g) \\ &= \hat{\Omega}_g^{1/2}\sqrt{n}\{\widehat{g}(\theta^g) - c_g\} + \hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g - \theta^g) + \hat{\Omega}_g^{1/2}\sqrt{n}c_g. \qquad (18)\end{aligned}$$

Under Assumption A7, A9, A10, A11 and $\hat{\Omega}_g^{1/2} \to^p \Omega_g^{1/2}$, the first term is asymptotically normal. Also, by Assumptions A12 to A15, using Theorem 2 of Hall and Inoue (2003), $\sqrt{n}(\hat{\theta}^g - \theta^g)$ is asymptotically normal with mean zero. Thus, the sum of first two terms in (18) are bounded in probability. However, the third term in (18) diverges at the rate $\sqrt{n}$ $(= O_p(n^{1/2}))$, and consequently, $n\hat{Q}^g$ diverges at the rate n as $n \to \infty$.

In short, the asymptotics of $n\hat{Q}^g$ is summarized as follows:

$$\begin{aligned}&\text{Case i) } G \text{ is correctly specified} \quad \Longrightarrow n\hat{Q}^g \to^d \chi^2_{k_g}/k_g \text{ as } n \to \infty \\ &\text{Case ii) } G \text{ is misspecified} \quad \Longrightarrow n\hat{Q}^g \text{ diverges as } n \to \infty.\end{aligned}$$

In the following, we investigate the probability limits of $\hat{W}_g$ and $\hat{W}_f$, using these results.

Case 1). Suppose both $g_0(\alpha_0, \beta_0) = 0$ and $h_0(\alpha_0, \gamma_0) = 0$. Then, $f_0(\alpha_0, \beta_0, \gamma_0) = 0$. By A1, A2, A3, A5, and A6, $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_0, \beta_0, \gamma_0\}$, so $\hat{Q}^g \to^p 0$, $\hat{Q}^h \to^p 0$, and $\hat{Q}^f \to^p 0$. For $n^\tau\hat{Q}^f$, following the same derivation in (17), we have

$$\begin{aligned}n^\tau\hat{Q}^f &= n^\tau\hat{f}(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)'\hat{\Omega}_f\hat{f}(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\frac{1}{k_f} \\ &= n^{\tau-1}\left\{\hat{\Pi}_f\hat{\Omega}_f^{1/2}\sqrt{n}\widehat{f}(\theta_0^f)\right\}'\left\{\hat{\Pi}_f\hat{\Omega}_f^{1/2}\sqrt{n}\widehat{f}(\theta_0^f)\right\}\frac{1}{k_f} = n^{\tau-1}\hat{\chi}^2_{k_f}\frac{1}{k_f}\end{aligned}$$

where $\hat{\Pi}_f \equiv I_{\tilde{k}_f} - \hat{\Omega}_f^{1/2} \nabla_{\theta'} \widehat{f}(\overline{\theta}^f)(\hat{H}^f)^{-1} \nabla_\theta \widehat{f}(\hat{\theta}^f) \hat{\Omega}_f^{1/2}$ and $\hat{\chi}^2_{k_f} \equiv \{\hat{\Pi}_f \hat{\Omega}_f^{1/2} \sqrt{n} \widehat{f}(\theta_0^f)\}' \{\hat{\Pi}_f \hat{\Omega}_f^{1/2} \sqrt{n} \widehat{f}(\theta_0^f)\}$. Following the same steps as in Case i) of Lemma 1, we have $\hat{\chi}^2_{k_f} \to^d \chi^2_{k_f}$ which is bounded in probability, and consequently, $n^\tau \hat{Q}^f \to^p 0$ for $\tau < 1$, and

$$\hat{W}_f = 1 - \frac{1}{n^\tau \hat{Q}^f + 1} = 1 - \frac{1}{n^{\tau-1} n \hat{Q}^f + 1} \to^p 0.$$

As for $\hat{W}_g$, due to $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) \to_d \chi^2_{k_g}/k_g$ and $n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) \to_d \chi^2_{k_h}/k_h$, $\hat{W}_g = n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\}$ converges to a ratio of possibly dependent random variables, which lies between zero and one with probability one. We do not need the limiting distribution of $\hat{W}_g$ [14], as it is enough to have $\hat{W}_g$ bounded in probability to ensure $\hat{W}_f \hat{W}_g \to^p 0$ when $\hat{W}_f \to^p 0$.

Case 2). Suppose $g_0(\alpha_0, \beta_0) = 0$ but $h_0(\alpha_0, \gamma_0) \neq 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_h, \gamma_h\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. By the continuous mapping theorem and uniform convergence of $\hat{Q}^g$ and $\hat{Q}^h$, we have $\hat{Q}^g \to^p Q_0^g = c_g' \Omega_g c_g / k_g = 0$, $\hat{Q}^h \to^p Q_0^h = c_h' \Omega_h c_h / k_h > 0$, and $\hat{Q}^f \to^p Q_0^f = c_f' \Omega_f c_f / k_f > 0$. From Case ii), $n\hat{Q}^h$ diverges as $n \to \infty$ while $n\hat{Q}^g$ is bounded in probability, and thus $\hat{W}_g = n\hat{Q}^g/(n\hat{Q}^g + n\hat{Q}^h) \to^p 0$. As for $\hat{W}_f$, due to $\hat{Q}^f \to^p Q_0^f = c_f' \Omega_f c_f / k_f > 0$, we have

$$\hat{W}_f = 1 - \frac{1}{n^\tau \hat{Q}^f + 1} = 1 - \frac{1}{n^{\tau-1} n \hat{Q}^f + 1} \to^p 1$$

and $\hat{W}_f \hat{W}_g \to^p 0$.

Case 3). Suppose now $g_0(\alpha_0, \beta_0) \neq 0$ but $h_0(\alpha_0, \gamma_0) = 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_g, \beta_g\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. So $\hat{Q}^g \to^p Q_0^g = c_g' \Omega_g c_g / k_g > 0$, $\hat{Q}^h \to^p Q_0^h = c_h' \Omega_h c_h / k_h = 0$, and $\hat{Q}^f \to^p Q_0^f = c_f' \Omega_f c_f / k_f > 0$. Following the same argument as in Case 2),

[14] If $\hat{Q}^g$ and $\hat{Q}^f$ happen to be independent, then $\hat{W}_g$ would be a ratio of independent Chi-squareds and so converges to a beta distribution with shape parameters $k_g/2$ and $k_k/2$. But there is no reason to impose that these distributions be independent.

$\hat{W}_g \to^p 1$ and $\hat{W}_f \to^p 1$. In short, the probability limits of $\hat{W}_f$ and $\hat{W}_g\hat{W}_f$ are categorized as follows:

Case 1) Both G and H are correctly specified $\implies \hat{W}_f \to^p 0$ and $\hat{W}_f\hat{W}_g \to^p 0$

Case 2) G is correctly specified, but H is not $\implies \hat{W}_f \to^p 1$ and $\hat{W}_f\hat{W}_g \to^p 0$

Case 3) H is correctly specified, but G is not $\implies \hat{W}_f \to^p 1$ and $\hat{W}_f\hat{W}_{fg} \to^p 1$.

Q.E.D.

Proof of Theorem 2.

Recall equation (1) and rewrite it as

$$\widehat{\alpha} = \alpha_0 + \hat{W}_f\hat{W}_g(\widehat{\alpha}_h - \alpha_0) + \hat{W}_f\left(1 - \hat{W}_g\right)(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)(\widehat{\alpha}_f - \alpha_0)$$

$$\begin{aligned}
\implies \sqrt{n}(\widehat{\alpha} - \alpha_0) &= \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h - \alpha_0) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_0) \\
&= \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g - \alpha_g) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \qquad (19) \\
&\quad + \hat{W}_f\hat{W}_g\sqrt{n}(\alpha_h - \alpha_0) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\alpha_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\alpha_f - \alpha_0)
\end{aligned}$$

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

Case 1). Suppose G and H are both correct. Then, because of $\alpha_g = \alpha_h = \alpha_f = \alpha_0$, α_g, α_h, α_f in the first line of (19) are replaced by α_0, and the second line disappears. Following the same argument as in Theorem 3.4 of Newey and McFadden, under Assumption A7, A9, A10 and A11, the central limit theorem yields $n^{-1/2}\sum_i G(Z_i, \alpha_0, \beta_0) \to^d N(0, \Sigma_g)$ 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 $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$. Following the same argument, along with the consistency of $(\widehat{\alpha}_h, \widehat{\beta}_h)$ and $(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_h - \alpha_0)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_0)$ are established. Therefore, by Lemma 1 on $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$, and boundedness of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and $\sqrt{n}(\widehat{\alpha}_h - \alpha_0)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_f - \alpha_0)$, and the continuous mapping theorem, we have

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

By Assumption A8

$$\frac{1}{n}\sum_i \widehat{\eta}_i^f \widehat{\eta}_i^{f\prime} \to^p \widetilde{V}^f \equiv E(\eta^f \eta^{f\prime}) \quad \text{where} \quad \sqrt{n}(\widehat{\alpha}_f - \alpha_0) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}_i^f;$$

$\widehat{\eta}_i^f$ is the influence function of $\widehat{\alpha}_f$ (the details of $\widehat{\eta}_i^f$ are given in (28) of the Appendix III), making $\widetilde{V}^f = \widetilde{V}$.

Case 2). Suppose G is correct, but H is not ($\alpha_h - \alpha_0 \equiv \delta_h \neq 0$). In this case, F is also misspecified ($\alpha_f - \alpha_0 \equiv \delta_f \neq 0$). Then, (19) can be rewritten as

$$\begin{aligned}\sqrt{n}(\widehat{\alpha} - \alpha_0) = \hat{W}_f \hat{W}_g \sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f \left(1 - \hat{W}_g\right) \sqrt{n}(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \\ + \hat{W}_f \hat{W}_g \sqrt{n}\delta_h + (1 - \hat{W}_f)\sqrt{n}\delta_f \qquad (20)\end{aligned}$$

By Theorem 1, $(\widehat{\alpha}_g, \widehat{\beta}_g) \to^p (\alpha_0, \beta_0)$, while $(\widehat{\alpha}_h, \widehat{\gamma}_h) \to^p (\alpha_h, \gamma_h)$ and $(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \to^p (\alpha_f, \beta_f, \gamma_f)$. Following the same argument as above, we have the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$. Under Assumption A7, A9, A10, A11, A12, A13, A14, and A15, by Theorem 2 of Hall and Inoue (2003), $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ is asymptotically normal with mean zero and a complex form of the variance. The same argument holds for $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ too. In the second line of (20), $\hat{W}_f \hat{W}_g \sqrt{n} = \left(1 - \frac{1}{O_p(n^\tau)+1}\right) \frac{O_p(1)}{O_p(1)+O_p(n)} O(\sqrt{n})$ and $(1 - \hat{W}_f)\sqrt{n} = \frac{1}{O_p(n^\tau)+1} O(\sqrt{n})$, and thus for $\tau > 1/2$, the second line disappears as $n \to \infty$. By Lemma 1 on $\hat{W}_f \to^p 1$ and $\hat{W}_g \hat{W}_f \to^p 0$, boundedness of $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and the continuous mapping theorem, we have $\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^g)$. By Assumption A8, we get

$$\frac{1}{n}\sum_i \widehat{\eta}_i^g \widehat{\eta}_i^{g\prime} \to^p \widetilde{V}^g \equiv E(\eta^g \eta^{g\prime}) \quad \text{where} \quad \sqrt{n}(\widehat{\alpha}_g - \alpha_0) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}_i^g;$$

$\widehat{\eta}_i^g$ is the influence function of $\widehat{\alpha}_g$ (the details of $\widehat{\eta}_i^g$ are given in (30) of the Appendix III), making $\widetilde{V}^g = \widetilde{V}$.

Case 3). Suppose H is correct, but G is not ($\Longrightarrow \alpha_g - \alpha_0 \equiv \delta_g \neq 0$). Then the same argument as in Case 2) applies, replacing $\hat{W}_g$ with $1 - \hat{W}_g$, and switching the roles of β and γ and the roles of g and h.

Appendix II

Let the model G be "locally misspecified" when the parameter in the data generating process takes the form $\theta^g = \theta_0^g + \delta_g n^{-s}$ for a constant δ_g and $s > 0$, while θ_0^g satisfies $E\{G(Z, \theta_0^g)\} = 0$ due to Assumption A3. Analogously, let the model H be "locally misspecified" when the parameter in the data generating process is $\theta^h = \theta_0^h + \delta_h n^{-s}$ with $E\{H(Z, \theta_0^h)\} = 0$. When $s = 1/2$, $\delta_g n^{-s}$ is 'Pitman drift' as in Pitman (1949), Newey and West (1987), Bera and Yoon (1993) and Newey and McFadden (1994). When model G or H is locally misspecified, we have, respectively,

$$g_0(\theta^g) \equiv g_0(\theta_0^g) + \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g n^{-s} = \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g n^{-s} \qquad \text{with} \quad \omega_g \equiv \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g,$$

$$h_0(\theta^h) \equiv h_0(\theta_0^h) + \nabla_{\theta'} h_0(\widetilde{\theta}^h)\delta_h n^{-s} = \nabla_{\theta'} h_0(\widetilde{\theta}^h)\delta_h n^{-s} \qquad \text{with} \quad \omega_h \equiv \nabla_{\theta'} h_0(\widetilde{\theta}^g)\delta_h,$$

$\widetilde{\theta}^g$ is a mean value between θ^g and θ_0^g, and $\widetilde{\theta}^h$ is a mean value between θ^h and θ_0^h.

Before presenting the detailed proofs, we summarize here our main findings when one of the models is locally misspecified but another is correctly specified. Suppose that model H is correctly specified and model G is locally misspecified, with $\theta^g = \theta_0^g + \delta_g n^{-s}$. This local misspecification does not affect the consistency of our estimator $\hat{\alpha}$, because the local misspecification reduces to the correct specification as $n \to \infty$ and the weights $\hat{W}_g$ and $\hat{W}_g\hat{W}_f$ still have finite probability limits under the local misspecification. As for asymptotic distribution, when $s > 0.5$, the limiting distribution of $\sqrt{n}(\hat{\alpha}-\alpha_0)$ is the same as when both models are correct, because the drift approaches 0 sufficiently quickly. Second, when $s = 0.5$, $\hat{\alpha}$ is consistent but not $\sqrt{n}$-consistent. Third, when $s < 0.5$, if $s + 0.5 < \tau$, then the asymptotic distribution of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ is the same as if model G was globally misspecified (and is still $\sqrt{n}$-consistent, because asymptotically all weight goes on model H) .

Assumption A16: Either 1) model G is correct but model H is locally misspecified, or 2) model H is correct but model G is locally misspecified.

Lemma App.1: Let Assumption A1 and Assumptions A3 to A16 hold. For any τ with $0 < \tau < 1$, $\hat{W}_f$ and $\hat{W}_g\hat{W}_f$ have finite probability limits.

Proof for Lemma App.1.

Analogously to the proof for Lemma 1, first we consider without loss of generality the probability limit of $\hat{Q}^g$ when the model is locally misspecified. Then, the probability limits of $\hat{Q}^h$ and $\hat{Q}^f$ can be found following the same logic. Next, we find the probability limits of $\hat{W}_g$ and $\hat{W}_f$, based on those of $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$.

Case iii). Suppose that G is locally misspecified ($\theta^g = \theta_0^g + \delta_g n^{-s}$). Replacing θ_0^g with θ^g in (17) gives

$$\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g) = \hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta^g) = \hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\sqrt{n}\{\widehat{g}(\theta^g) - \omega_g n^{-s}\} + \hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s} \tag{21}$$

note $E\{\widehat{g}(\theta^g)\} = g_0(\theta^g) = \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g n^{-s} = \omega_g n^{-s}$. Under Assumption A1, A2, A3, A5, and A6, the corresponding GMM estimator is still consistent $\hat{\theta}^g \to^p \theta_0^g$ by Theorem 9.1 of Newey and McFadden (1994). By Assumption A8, A11 and A12, $\nabla_\theta \widehat{g}(\bar{\theta}^g) \to^p \nabla_\theta g(\theta_0^g)$ for $\bar{\theta}^g$ in $\hat{\Pi}_g^*$ and $\hat{\Omega}_g^{-1} \to^p \Omega_g^{-1} = E\{G(Z,\theta_0^g)G(Z,\theta_0^g)'\}$, and thus, $\hat{\Pi}_g^* \to^p \Pi_g^*$, which is a $\widetilde{k}_g \times \widetilde{k}_g$ symmetric and idempotent matrix with $trace(\Pi_g) = k_g$. Therefore, applying the same the argument in Case i) of Lemma 1, along with the consistency of $\hat{\theta}^g$, the first term in the right-hand side of (21) is asymptotically standard normal, and thus bounded in probability. Consequently, we can characterize the asymptotics of $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ depending on s using the last term $\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s}$ in (21).

If $s = 1/2$, then $\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s} \to^p \Pi_g^*\Omega_g^{1/2}\omega_g$ and $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ is asymptotically normal with mean $\Pi_g^*\Omega_g^{1/2}\omega_g$ and unit variance. Hence,

$$n\hat{Q}^g = \{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)\}'\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g) \to^d \chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)/k_g;$$

$\chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)$ is the noncentral chi-squared distribution with noncentrality parameter $\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g$. If $s > 1/2$ in (21), the noncentrality parameter shrinks to zero, so that $n\hat{Q}^g \to^d \chi^2_{k_g}(0)/k_g$ as $n \to \infty$, analogously to Case i) of Lemma 1. If $s < 1/2$ in (21), then $\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s} = O_p(n^{1/2-s})$ diverges

as $n \to \infty$, analogously to Case ii) of Lemma 1. In short,

$$\begin{aligned}
&\text{Case iii) with } s < 1/2 \quad \Longrightarrow n\hat{Q}^g \text{ diverges as } n \to \infty; \\
&\text{Case iii) with } s = 1/2 \quad \Longrightarrow n\hat{Q}^g \to^d \chi^2_{k_g}(\omega_g' \Omega_g^{1/2} \Pi_g^* \Omega_g^{1/2} \omega_g)/k_g \text{ as } n \to \infty; \\
&\text{Case iii) with } s > 1/2 \quad \Longrightarrow n\hat{Q}^g \to^d \chi^2_{k_g}(0)/k_g \text{ as } n \to \infty.
\end{aligned}$$

Next, we investigate the probability limits of $\hat{W}_g$ and $\hat{W}_f$ based on those of $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$, doing analogously to what was done for Case iii).

Case 4). Suppose that model G is correct, but H is locally misspecified with $\theta^h = \theta_0^h + \delta_h n^{-s}$. In this case, F is also locally misspecified with $\theta^f = \theta_0^f + \delta_f n^{-s}$ for some δ_f.

Case 4-1). If $s = 1/2$, as shown in Case iii), $n\hat{Q}^h \to_d \chi^2_{k_h}(\omega_h' \Omega_h^{1/2} \Pi_h^* \Omega_h^{1/2} \omega_h)/k_h$ and $n\hat{Q}^f \to_d \chi^2_{k_f}(\omega_f' \Omega_f^{1/2} \Pi_f^* \Omega_f^{1/2} \omega_f)/k_f$ for some ω_f as $n \to \infty$. Thus $\hat{W}_g = n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\}$ converges to a distribution on $(0, 1)$. For $\hat{W}_f$, we have

$$\hat{W}_f = 1 - \frac{1}{n^\tau \hat{Q}^f + 1} = 1 - \frac{1}{n^{\tau-1} n\hat{Q}^f + 1} \to^p 0,$$

because $n\hat{Q}^f$ is bounded in probability, and $n^{\tau-1} \to^p 0$. Thus, $\hat{W}_g \hat{W}_f \to^p 0$.

Case 4-2). If $s > 1/2$, as shown in Case iii), $n\hat{Q}^h \to_d \chi^2_{k_h}/k_h$, and $n\hat{Q}^f \to_d \chi^2_{k_f}/k_f$. Therefore, it is asymptotically the same as Case 1) of Lemma 1.

Case 4-3). If $s < 1/2$, as shown in Case iii), $n\hat{Q}^h$ and $n\hat{Q}^f$ are $O_p(n^{2(1/2-s)})$, as each is a squared version of a term analogous to (21). In this case, whereas $\hat{W}_g \to^p 0$, convergence of $\hat{W}_f$ depends on the relationship between τ and s. Because $n^\tau \hat{Q}^f = O(n^{\tau-1}) O_p(n^{2(1/2-s)}) = O_p(n^{\tau-2s})$, when $\tau > 2s$, $n^\tau \hat{Q}^f$ diverges to result in $\hat{W}_f \to^p 1$ and $\hat{W}_g \hat{W}_f \to^p 0$. When $\tau < 2s$, $n^\tau \hat{Q}^f \to^p 0$, and consequently $\hat{W}_f \to^p 0$ and $\hat{W}_f \hat{W}_g \to^p 0$. When $\tau = 2s$, however, (21) shows that $n^\tau \hat{Q}^f \to^p \omega_f' \Omega_f^{1/2} \Pi_f^* \Omega_f^{1/2} \omega_f$ because only the last term of (21) matters, so that $\hat{W}_f \to^p W_f^* \equiv 1 - (\omega_f' \Omega_f^{1/2} \Pi_f^* \Omega_f^{1/2} \omega_f + 1)^{-1}$ and $\hat{W}_g \hat{W}_f \to^p 0$.

Case 5). Suppose that model G is locally misspecified with $\theta^g = \theta_0^g + \delta_g n^{-s}$, but model H is correct. Then essentially the same arguments as in Case 4) apply.

Case 5-1). If $s = 1/2$, then $n\hat{Q}^g \to_d \chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)/k_g$ and $n\hat{Q}^f \to_d \chi^2_{k_f}(\omega_f'\Omega_f^{1/2}\Pi_f^*\Omega_f^{1/2}\omega_f)/k_f$. Thus, $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$.

Case 5-2). If $s > 1/2$, then $n\hat{Q}^g \to_d \chi^2_{k_g}/k_g$, and $n\hat{Q}^f \to_d \chi^2_{k_f}/k_f$ as $n \to \infty$, which is asymptotically the same as Case 1) of Lemma 1.

Case 5-3). If $s < 1/2$, then since $n\hat{Q}^g$ and $n\hat{Q}^f$ diverge, $\hat{W}_g \to^p 1$ but the asymptotics of $\hat{W}_f$ depends on the relationship between τ and s. For $\tau > 2s$, $n^{\tau-1}n\hat{Q}^f$ diverges, and thus, $\hat{W}_f \to^p 1$ and $\hat{W}_g\hat{W}_f \to^p 1$; for $\tau < 2s$, $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$. When $\tau = 2s$, $\hat{W}_f \to^p W_f^*$ and $\hat{W}_g\hat{W}_f \to^p W_f^*$ because $\hat{W}_g \to^p 1$.

In sum, the probability limits of $\hat{W}_f$ and $\hat{W}_g\hat{W}_f$ are categorized as follows:

$$
\begin{aligned}
&\text{Case 4-1) and 4-2) with } s \geq 1/2, && \Longrightarrow \hat{W}_f \to^p 0 \text{ and } \hat{W}_g\hat{W}_f \to^p 0 \\
&\text{Case 4-3) with } s < 1/2 \text{ and } 2s < \tau && \Longrightarrow \hat{W}_f \to^p 1 \text{ and } \hat{W}_g\hat{W}_f \to^p 0 \\
&\text{Case 4-3) with } s < 1/2 \text{ and } \tau = 2s && \Longrightarrow \hat{W}_f \to^p W_f^* \text{ and } \hat{W}_g\hat{W}_f \to^p 0 \\
&\text{Case 4-3) with } s < 1/2 \text{ and } \tau < 2s && \Longrightarrow \hat{W}_f \to^p 0 \text{ and } \hat{W}_g\hat{W}_f \to^p 0 \\
&\text{Case 5-1) and 5-2) with } s \geq 1/2, && \Longrightarrow \hat{W}_f \to^p 0 \text{ and } \hat{W}_g\hat{W}_f \to^p 0 \\
&\text{Case 5-3) with } s < 1/2 \text{ and } 2s < \tau && \Longrightarrow \hat{W}_f \to^p 1 \text{ and } \hat{W}_g\hat{W}_f \to^p 1 \\
&\text{Case 5-3) with } s < 1/2 \text{ and } \tau = 2s && \Longrightarrow \hat{W}_f \to^p W_f^* \text{ and } \hat{W}_g\hat{W}_f \to^p W_f^* \\
&\text{Case 5-3) with } s < 1/2 \text{ and } \tau < 2s && \Longrightarrow \hat{W}_f \to^p 0 \text{ and } \hat{W}_g\hat{W}_f \to^p 0.
\end{aligned}
$$

Q.E.D.

Theorem App.1: Under Assumptions A1 and A3 to A16, for $\widehat{\alpha}$ given by equation (1), $\widehat{\alpha} \to^p \alpha_0$.

Proof for Theorem App.1.

Case 4). Suppose that G is correct, but H is the locally misspecified with $\theta^h = \theta_0^h + \delta_h n^{-s}$. By Theorem 9.1 of in Newey and McFadden (1994), still $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$ and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_0, \beta_0, \gamma_0\}$. By Lemma App.1, if $s \geq 1/2$, then $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$, and the consistency of $\widehat{\alpha}$ in (1) follows from consistency of $\widehat{\alpha}_f$. If $s < 1/2$, the probability limits of $\hat{W}_f$ and $\hat{W}_g\hat{W}_f$ depend on the relationship between τ and s. If $s < 1/2$ and $\tau < 2s$, the limits

are the same as in the case with $s \geq 1/2$ by Lemma App.1, and thus, the same argument holds for $\widehat{\alpha}$. If $s < 1/2$ and $\tau > 2s$, by Lemma App.1 $\hat{W}_f \to^p 1$ and $\hat{W}_g \hat{W}_f \to^p 0$ and the consistency of $\widehat{\alpha}$ follows from the consistency of $\widehat{\alpha}_g$. If $s < 1/2$ and $\tau = 2s$, then by Lemma App.1 $\hat{W}_f \to^p W_f^*$ and $\hat{W}_g \hat{W}_f \to^p 0$ and the consistency of $\widehat{\alpha}$ follows from the consistency of $\widehat{\alpha}_g$ and $\widehat{\alpha}_f$, and $\widehat{\alpha}_g - \widehat{\alpha}_f \to^p 0$.

Case 5). Suppose that H is correct, but G is locally misspecified. Then, essentially the same arguments as in Case 4 apply. Q.E.D.

Theorem App.2: Under Assumptions A1 and A3 to A16, for $1/2 < \tau < 1$, when $s > 1/2$ or $s + 1/2 < \tau$, there exists a matrix $\widetilde{V}$ such that

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

and

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

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

Proof of Theorem App.2.

To ease referencing, recall (19):

$$\begin{aligned}\sqrt{n}(\widehat{\alpha} - \alpha_0) = {} & \hat{W}_f \hat{W}_g \sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f \left(1 - \hat{W}_g\right) \sqrt{n}(\widehat{\alpha}_g - \alpha_g) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \\ & + \hat{W}_f \hat{W}_g \sqrt{n}(\alpha_h - \alpha_0) + \hat{W}_f \left(1 - \hat{W}_g\right) \sqrt{n}(\alpha_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\alpha_f - \alpha_0).\end{aligned}$$

Case 4). Suppose model G is correct ($\alpha_g = \alpha_0$), but H is locally misspecified with $\alpha_h = \alpha_0 + \delta_h n^{-s}$; then F is also locally misspecified with $\alpha_f = \alpha_0 + \delta_f n^{-s}$. Rewrite (19) as

$$\begin{aligned}\sqrt{n}(\widehat{\alpha} - \alpha_0) = {} & \hat{W}_f \hat{W}_g \sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f \left(1 - \hat{W}_g\right) \sqrt{n}(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \\ & + \hat{W}_f \hat{W}_g \delta_h n^{1/2-s} + (1 - \hat{W}_f)\delta_f n^{1/2-s}. \qquad (22)\end{aligned}$$

Following the same argument in Case 1) of Theorem 2, we can establish asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$. Call (29) in the Appendix III below replacing g with h to have

$$\sqrt{n}(\widehat{\alpha}_h - \alpha_h) = \widehat{A}_h^{-1}\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^* \sqrt{n}\widehat{h}(\theta^h)$$

where

$$\widehat{A}_h \equiv \nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^* \nabla_{\alpha'} \widehat{h}(\overline{\theta}^h), \quad \hat{\Omega}_h^* \equiv \hat{\Omega}_h^{1/2\prime}\hat{\Pi}_h \hat{\Omega}_h^{1/2},$$
$$\hat{\Pi}_h \equiv [I_{\widetilde{k}_h} - \hat{\Omega}_h^{1/2}\nabla_{\gamma'}\widehat{h}(\overline{\theta}^h)\{\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h \nabla_{\gamma'}\widehat{h}(\overline{\theta}^h)\}^{-1}\nabla_\gamma \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^{1/2\prime}].$$

With $h_0(\theta^h) \equiv \omega_h n^{-s}$, add and subtract $E\{\widehat{h}(\theta^h)\} = \omega_h n^{-s}$ to get

$$\sqrt{n}(\widehat{\alpha}_h - \alpha_h) = \widehat{A}_h^{-1}\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^* \sqrt{n}\{\widehat{h}(\theta^h) - \omega_h n^{-s}\} + \widehat{A}_h^{-1}\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^* \sqrt{n}\omega_h n^{-s}. \tag{23}$$

Applying the vectorization part in Hall and Inoue (2003, p.367) and using the population first-order condition $\nabla_\alpha h_0(\theta^h)\Omega_h^* h_0(\theta^h) = 0$, rewrite $\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^* \sqrt{n}\omega_h n^{-s}$ in the last term other than $\widehat{A}_h^{-1}$ as

$$\begin{aligned}
&\sqrt{n}\nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^* \omega_h n^{-s} = \\
&\sqrt{n}\{\nabla_\alpha \widehat{h}(\hat{\theta}^h) - \nabla_\alpha \widehat{h}(\theta^h)\}\hat{\Omega}_h^* \omega_h n^{-s} + \sqrt{n}\{\nabla_\alpha \widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}\hat{\Omega}_h^* \omega_h n^{-s} + \nabla_\alpha h_0(\theta^h)\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)\omega_h n^{-s} \\
&= \omega_h n^{-s}\hat{M}^h \sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \sqrt{n}\{\nabla_\alpha \widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}\hat{\Omega}_h^* \omega_h n^{-s} + \nabla_\alpha h_0(\theta^h)\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)\omega_h n^{-s}
\end{aligned} \tag{24}$$

for some symmetric $k_h^* \times k_h^*$ matrix $\hat{M}^h$ involving the second-order derivative of $h(\cdot)$ that is bounded in probability. Plugging (24) into (23) and solving for $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ gives

$$\sqrt{n}(\widehat{\alpha}_h - \alpha_h) = [I_{k_h^*} - \widehat{A}_h^{-1}\omega_h n^{-s}\hat{M}^h]^{-1}\widehat{A}_h^{-1}\widehat{\Gamma}_h, \tag{25}$$
$$\begin{aligned}
\widehat{\Gamma}_h \equiv\ & \nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^* \sqrt{n}\{\widehat{h}(\theta^h) - \omega_h n^{-s}\} \\
& + \sqrt{n}\{\nabla_\alpha \widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}\hat{\Omega}_h^* \omega_h n^{-s} + \nabla_\alpha h_0(\theta^h)\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)\omega_h n^{-s}.
\end{aligned}$$

Under Assumptions A12 to A16, $\sqrt{n}\{\nabla_\alpha \widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}$ and $\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)$ are bounded in probability, so that the last two terms of $\widehat{\Gamma}_h$ converge to zero in probability. Under Assumption A7, A9, A10 and A11, $\sqrt{n}\{\widehat{h}(\theta^h) - \omega_h n^{-s}\}$ is asymptotically normal with mean zero. Therefore,

due to $\widehat{\theta}_h \to^p \theta_0$ (from Theorem 9.1 of Newey and McFadden, 1994), $\omega_h n^{-s}\hat{M}^h \to^p 0$, $\nabla_\theta \widehat{h}(\widehat{\theta}^h) \to^p \nabla_\theta h_0(\theta_0^h)$, $\hat{\Omega}_h^* \to^p \Omega_h^*$, $\hat{\Omega}_h^{-1} \to^p \Omega_h^{-1} = E\{H(Z,\alpha_0,\gamma_0)H(Z,\alpha_0,\gamma_0)'\}$, and the continuous mapping theorem, we get

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

where $\widetilde{V}^h$ is the same asymptotic variance as in Case 3) as if model H were correct. Analogously, the same argument holds for $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$, so that we have $\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \to^d N(0, \widetilde{V}^f)$ as if model F were correct. Hence, all of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$, $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in the first line of (22) are asymptotically normal with mean zero and variance being that of the corresponding GMM estimator under correct specification.

Recall (22):

$$\begin{aligned}\sqrt{n}(\widehat{\alpha} - \alpha_0) = \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \\ + \hat{W}_f\hat{W}_g\delta_h n^{1/2-s} + (1 - \hat{W}_f)\delta_f n^{1/2-s}.\end{aligned}$$

Recalling (21) and its "squared version", we have

$$n\hat{Q}^h = O_p(n^{2(1/2-s)}) \quad \text{and} \quad n\hat{Q}^f = O_p(n^{2(1/2-s)}) \Longrightarrow n^\tau\hat{Q}^f = n^{\tau-1}n\hat{Q}^f = O_p(n^{\tau-1+2(1/2-s)}) = O_p(n^{\tau-2s}).$$

Consequently, for the last two terms in (22), we get

$$\begin{aligned}\hat{W}_f\hat{W}_g\delta_h n^{1/2-s} + (1 - \hat{W}_f)\delta_f n^{1/2-s} &= \left(1 - \frac{1}{n^\tau\hat{Q}^f + 1}\right)\left(\frac{n\hat{Q}^g \cdot \delta_h n^{1/2-s}}{n\hat{Q}^g + n\hat{Q}^h}\right) + \left(\frac{1}{n^\tau\hat{Q}^f + 1}\right)\delta_f n^{1/2-s} \\ &= \left(1 - \frac{1}{O_p(n^{\tau-2s}) + 1}\right)\left(\frac{O_p(1)O(n^{1/2-s})}{O_p(1) + O_p(n^{2(1/2-s)})}\right) + \left(\frac{1}{O_p(n^{\tau-2s}) + 1}\right)O(n^{1/2-s}). \qquad (26)\end{aligned}$$

For the first term in the left-hand side in (26), if $s = 1/2$, its probability limit is zero because $\hat{W}_f \to^p 0$ and $\hat{W}_g n^{1/2-s}$ is bounded in probability. If $s > 1/2$, the probability limit is zero because $\hat{W}_f \to^p 0$ and $\hat{W}_g n^{1/2-s} \to^p 0$. If $s < 1/2$, the probability limit is zero because $\hat{W}_f$ is bounded between zero and one in probability and $\hat{W}_g n^{1/2-s} \to^p 0$. Therefore, the first term in the left-hand side in (26) disappears as $n \to \infty$, regardless of s. However, the probability limit of the second term $(1 - \hat{W}_f)\delta_f n^{1/2-s}$ in the left-hand side in (26) varies, depending on the relationship between s and τ. So, the asymptotic behavior of $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ depends on the values of s and τ as follows.

Case 4-1). If $s = 1/2$, $\hat{W}_f \to^p 0$ and $(1-\hat{W}_f)\delta_f n^{1/2-s} \to^p \delta_f$ as $n \to \infty$. By Lemma App.1, $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$, boundedness of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ and the continuous mapping theorem, only $(1-\hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ survives in (22) and we get

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

Case 4-2). If $s > 1/2$, $\hat{W}_f \to^p 0$ and $(1-\hat{W}_f)\delta_f n^{1/2-s} \to^p 0$ as $n \to \infty$. Therefore, we get $\hat{W}_g\hat{W}_f \to^p 0$ by Lemma App.1. Hence,

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

which is asymptotically equivalent to Case 1).

Case 4-3). If $s < 1/2$, the probability limit of the second term in (26) depends on s and τ:

$$(1-\hat{W}_f)\delta_f n^{1/2-s} = \left(\frac{1}{O_p(n^{\tau-2s})+1}\right) O(n^{1/2-s}) = O_p(n^{s+1/2-\tau}).$$

When $s + 1/2 < \tau$, $(1-\hat{W}_f)\delta_f n^{1/2-s}$ disappears as $n \to \infty$, which implies that the second line of (22) disappears. Due to $s < 1/2$,

$$s + 1/2 < \tau \Longrightarrow 2s < 1/2 + s < \tau \Longrightarrow \hat{W}_f \to^p 1 \text{ because of } 2s < \tau, \text{ and } \hat{W}_g\hat{W}_f \to^p 0$$

by Lemma App.1. Therefore, by the boundedness of $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and the continuous mapping theorem, we have

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

which is asymptotically equivalent to Case 2) of Theorem 2.

When $s + 1/2 > \tau$, $(1-\hat{W}_f)\delta_f n^{1/2-s}$ diverges as $n \to \infty$. Therefore, $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ is not bounded in probability.

When $s + 1/2 = \tau$, $(1-\hat{W}_f)\delta_f n^{1/2-s}$ converges to a constant, say ν, times δ_f, as $n \to \infty$. Also, we have

$$s + 1/2 = \tau \Longrightarrow 2s < 1/2 + s = \tau \Longrightarrow \hat{W}_f \to^p 1 \text{ because of } 2s < \tau, \text{ and } \hat{W}_g\hat{W}_f \to^p 0.$$

Therefore, by the boundedness of $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and the continuous mapping theorem, we get

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(\nu\delta_f, \widetilde{V}^g).$$

In sum, when G is correct but H is locally misspecified, $\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^f)$ if $s > 1/2$, or $\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^g)$ if $s + 1/2 < \tau$.

Case 5). Suppose H is correct specified, but G is locally misspecified with $\alpha^g = \alpha_0^g + \delta_g n^{-s}$. Then essentially the same arguments as in Case 4) apply, replacing $\hat{W}_g$ with $1 - \hat{W}_g$, and switching the roles of β and γ and the roles of g and h. Q.E.D.

Appendix III

Derivation of $\widehat{\eta}_i^f$, $\widehat{\eta}_i^g$, and $\widehat{\eta}_i^h$.

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

$$\widehat{\theta}^f \equiv (\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}^f) = \arg \min_{\{\alpha,\beta,\gamma\} \in \Theta_\alpha \times \Theta_\beta \times \Theta_\gamma} \widetilde{Q}^f(\alpha, \beta, \gamma) = \widehat{f}(\alpha, \beta, \gamma)\hat{\Omega}_f \widehat{f}(\alpha, \beta, \gamma).$$

Under Assumption A7 and A10-12, the following first-order conditions in the first-stage for $\widehat{\theta}^f$ hold:

$$FD_\alpha^f = \frac{\partial \widetilde{Q}^f(\widehat{\theta}^f)}{\partial \alpha} = \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\widehat{\theta}^f) = 0, \qquad FD_\beta^f = \frac{\partial \widetilde{Q}^f(\widehat{\theta}^f)}{\partial \beta} = \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\widehat{\theta}^f) = 0,$$
$$FD_\gamma^f = \frac{\partial \widetilde{Q}^f(\widehat{\theta}^f)}{\partial \gamma} = \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\widehat{\theta}^f) = 0.$$

Expend $\widehat{f}$ around the unique minimizer $\theta^f \equiv \{\alpha_f, \beta_f, \gamma_f\}$ to get

$$\widehat{f}(\widehat{\theta}^f) = \widehat{f}(\theta^f) + \nabla_{\alpha'} \widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'} \widehat{f}(\overline{\theta}^f)(\widehat{\beta} - \beta_f) + \nabla_{\gamma'} \widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)$$

where $\overline{\theta}^f$ is the mean value to apply the mean value theorem. Substitute these into each FD^f to

get

$$FD^f_\alpha = \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\{\widehat{f}(\theta^f) + \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'}\widehat{f}(\overline{\theta}^f)(\widehat{\beta}_f - \beta_f) + \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)\},$$

$$FD^f_\beta = \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\{\widehat{f}(\theta^f) + \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'}\widehat{f}(\overline{\theta}^f)(\widehat{\beta}_f - \beta_f) + \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)\}.$$

$$FD^f_\gamma = \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\{\widehat{f}(\theta^f) + \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'}\widehat{f}(\overline{\theta}^f)(\widehat{\beta}_f - \beta_f) + \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)\},$$

$$FD^f = \{FD^f_\alpha, FD^f_\beta, FD^f_\gamma\} = \widehat{I}^f + \widehat{H}^f(\widehat{\theta}^f - \theta^f), \text{ and from these, } \sqrt{n}(\widehat{\theta}^f - \theta^f) = \widehat{H}^{f-1}\sqrt{n}\widehat{I}^f,$$

$$\widehat{I}^f \equiv \begin{bmatrix} \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\theta^f) \\ \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\theta^f) \\ \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\theta^f) \end{bmatrix}, \widehat{H}^f \equiv \begin{bmatrix} \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f) & \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\beta'}\widehat{f}(\overline{\theta}^f) & \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f) \\ \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f) & \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\beta'}\widehat{f}(\overline{\theta}^f) & \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f) \\ \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f) & \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\beta'}\widehat{f}(\overline{\theta}^f) & \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f) \end{bmatrix}.$$

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

$$\sqrt{n}(\widehat{\alpha}_f - \alpha_f) = \widehat{A}_f^{-1}\nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}^*_f\sqrt{n}\widehat{f}(\theta^f), \qquad \widehat{A}_f \equiv \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}^*_f\nabla_{\alpha'}\widehat{f}(\overline{\theta}^f), \quad \hat{\Omega}^*_f \equiv \hat{\Omega}_f^{1/2}\hat{\Pi}_f\hat{\Omega}_f^{1/2}, \tag{27}$$

$$\begin{aligned}\hat{\Pi}_f \equiv I_{\widetilde{k}_f} &- \hat{\Omega}_f^{1/2}\nabla_\beta \widehat{f}(\overline{\theta}^f)\{\nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\beta'}\widehat{f}(\overline{\theta}^f)\}^{-1}\nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f^{1/2} \\ &- \hat{\Omega}_f^{1/2}\nabla_\gamma \widehat{f}(\overline{\theta}^f)\{\nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)\}^{-1}\nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f^{1/2}.\end{aligned}$$

Then we have

$$\sqrt{n}(\widehat{\alpha}_f - \alpha_f) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}^f_i, \qquad \widehat{\eta}^f_i \equiv \widehat{A}_f^{-1}\nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}^*_f F(Z_i, \theta^f), \tag{28}$$

and $\widehat{\eta}^f_i$ is the influence function of the first-stage estimate $\widehat{\alpha}_f$. If F is correct, θ^f is replaced by θ^f_0.

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

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

Under Assumption A7 and A10-12, with probability approaching one, the following first-order conditions in the first-stage for $\widehat{\theta}^g$ hold:

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

Expend $\widehat{g}$ 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 the value for the mean value theorem. Substitute these into each FD^g to get

$$FD^g_\alpha = \nabla_\alpha \widehat{g}(\widehat{\theta}^g)\hat{\Omega}_g[\widehat{g}(\theta^g) + \nabla_{\alpha'}\widehat{g}(\overline{\theta}^g)(\widehat{\alpha}_g - \alpha_g) + \nabla_{\beta'}\widehat{g}(\overline{\theta}^g)(\widehat{\beta}_g - \beta_g)]$$

$$FD^g_\beta = \nabla_\beta \widehat{g}(\widehat{\theta}^g)\hat{\Omega}_g[\widehat{g}(\theta^g) + \nabla_{\alpha'}\widehat{g}(\overline{\theta}^g)(\widehat{\alpha}_g - \alpha_g) + \nabla_{\beta'}\widehat{g}(\overline{\theta}^g)(\widehat{\beta}_g - \beta_g)]$$

$$FD^g = \{FD^g_\alpha, FD^g_\beta\} = \widehat{I}^g + \widehat{H}^g(\widehat{\theta}^g - \theta^g), \quad \text{and from these, } \sqrt{n}(\widehat{\theta}^g - \theta^g) = \widehat{H}^{g-1}\sqrt{n}\widehat{I}^g,$$

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

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:

$$\sqrt{n}(\widehat{\alpha}_g - \alpha_g) = \widehat{A}_g^{-1}\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\hat{\Omega}^*_g\sqrt{n}\widehat{g}(\theta^g), \qquad \widehat{A}_g \equiv \nabla_\alpha \widehat{g}(\widehat{\theta}^g)\hat{\Omega}^*_g\nabla_\alpha \widehat{g}(\overline{\theta}^g), \quad \hat{\Omega}^*_g \equiv \hat{\Omega}^{1/2}_g\hat{\Pi}_g\hat{\Omega}^{1/2}_g, \quad (29)$$

$$\hat{\Pi}_g \equiv I_{\widetilde{k}_g} - \hat{\Omega}^{1/2}_g\nabla_{g'}\widehat{g}(\overline{\theta}^g)\{\nabla_\beta \widehat{g}(\widehat{\theta}^g)\hat{\Omega}_g\nabla_{\beta'}\widehat{g}(\overline{\theta}^g)\}^{-1}\nabla_\beta \widehat{g}(\widehat{\theta}^g)\hat{\Omega}^{1/2}_g.$$

Then, we have

$$\sqrt{n}(\widehat{\alpha}_g - \alpha_g) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}^g_i, \qquad \widehat{\eta}^g_i \equiv \widehat{A}_g^{-1}\nabla_\alpha \widehat{g}(\widehat{\theta}^g)\hat{\Omega}^*_g G(Z_i, \theta^g), \quad (30)$$

and $\widehat{\eta}^g_i$ is the influence function of the first-stage estimate $\widehat{\alpha}_g$. If G is correct, θ^g is replaced by θ^g_0.

Analogously, we can obtain the influence function $\widehat{\eta}^h_i$ switching the roles of β and γ, and switching the roles of g and h. If H is correct, θ^h is replaced by θ^h_0.

Over-Identified Doubly Robust Identification and Estimation
by Arthur Lewbel, Jin-Young Choi, and Zhuzhu Zhou

Original 2018, Revised February 2021

Supplemental Online Appendix

In this Supplemental Appendix, we provide two additional examples of applying our ODR estimator. For both examples, DR estimators already exist, so we can comparing the requirements of our ODR estimator to existing DR applications. The first example is average treatment effect estimation, while the second concerns additive regression models.

Average Treatment Effect Estimation

Going 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 ODR 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

$$\alpha = E\{E(Y|T=1,X) - E(Y|T=0,X)\}. \tag{31}$$

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

$$\alpha = E\left\{\frac{YT}{E(T|X)} - \frac{Y(1-T)}{1-E(T|X)}\right\}. \tag{32}$$

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 (31), 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

$$E\left[\begin{array}{c} \{Y - \widetilde{G}(T,X,\beta)\} r_1(T,X) \\ \alpha - \{\widetilde{G}(1,X,\beta) - \widetilde{G}(0,X,\beta)\} \end{array}\right] = 0 \tag{33}$$

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 (32), 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

$$E\left[\begin{array}{c} \{T - \widetilde{H}(X,\gamma)\} r_2(X) \\ \alpha - \left\{\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)}\right\} \end{array}\right] = 0 \tag{34}$$

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 other authors assumes α can be expressed as

$$\alpha = E\left\{\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)} + \frac{T-\widetilde{H}(X,\gamma)}{\widetilde{H}(X,\gamma)}\widetilde{G}(1,X,\beta) - \frac{T-\widetilde{H}(X,\gamma)}{1-\widetilde{H}(X,\gamma)}\widetilde{G}(0,X,\beta)\right\}. \quad (35)$$

Observe that if $\widetilde{H}(X,\gamma) = E(T|X)$, then the first two terms in the above expectation equal equation (32) and the second two terms have mean zero. By rearranging terms, equation (35) can be rewritten as

$$\alpha = E\left[\widetilde{G}(1,X,\beta) - \widetilde{G}(0,X,\beta) + \frac{T}{\widetilde{H}(X,\gamma)}\{Y-\widetilde{G}(1,X,\beta)\} - \frac{1-T}{1-\widetilde{H}(X,\gamma)}\{Y-\widetilde{G}(0,X,\beta)\}\right]. \quad (36)$$

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 (36) equal equation (31), and the second two terms have mean zero. This shows that equation (35) or equivalently (36) 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

$$E\begin{bmatrix} \{Y-\widetilde{G}(T,X,\beta)\}r_1(T,X) \\ \{T-\widetilde{H}(X,\gamma)\}r_2(X) \\ \alpha - \left\{\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)} + \frac{T-\widetilde{H}(X,\gamma)}{\widetilde{H}(X,\gamma)}\widetilde{G}(1,X,\beta) - \frac{T-\widetilde{H}(X,\gamma)}{1-\widetilde{H}(X,\gamma)}\widetilde{G}(0,X,\beta)\right\} \end{bmatrix} = 0. \quad (37)$$

Construction of this doubly robust estimator required finding equation (35) which is special to the problem at hand and possesses the DR property. In general, finding such expressions for any particular problem may be difficult or impossible.

In contrast, our proposed ODR estimator does not require any such creativity. All that is required for constructing our ODR for this problem is to know the two alternative standard estimators, based on equations (31) and (32), expressed in GMM form, i.e., equation (33) and equation (34). Just define $G(Z,\alpha,\beta)$ to be the vector of functions given in equation (33) and define $H(Z,\alpha,\gamma)$ to

be the vector of functions given in equation (34). That is,

$$G(Z,\alpha,\beta)=\begin{bmatrix}\{Y-\widetilde{G}\,(T,X,\beta)\}r_1\,(T,X)\\ \alpha-\{\widetilde{G}\,(1,X,\beta)-\widetilde{G}\,(0,X,\beta)\}\end{bmatrix} \tag{38}$$

and

$$H(Z,\alpha,\gamma)=\begin{bmatrix}\{T-\widetilde{H}\,(X,\gamma)\}r_2\,(X)\\ \alpha-\left\{\frac{YT}{\widetilde{H}(X,\gamma)}-\frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)}\right\}\end{bmatrix}. \tag{39}$$

These functions can then be plugged into the expressions in the previous section to obtain our ODR estimator, equation (1), without having to find an expression like equation (35) 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 the required overidentification (discussed earlier, and formally given later in 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.

An Instrumental Variables Additive Regression Model

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

$$Y=M(W,\alpha)+\widetilde{G}(X)+U, \tag{40}$$

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

$$E\,(U\mid X,Q)=0, \tag{41}$$

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 $K\geq S$ vector of observed instruments. Note

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

Okui, Small, Tan, and Robins (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

$$G(Z, \alpha, \beta) = \begin{bmatrix} \{Y - M(W, \alpha) - \widetilde{G}(X, \beta)\} r_1(X) \\ \{Y - M(W, \alpha) - \widetilde{G}(X, \beta)\} Q \end{bmatrix} \tag{42}$$

and

$$H(Z, \alpha, \gamma) = \begin{bmatrix} \{Q - \widetilde{H}(X, \gamma)\} r_2(X) \\ \{Y - M(W, \alpha)\}\{Q - \widetilde{H}(X, \gamma)\} \end{bmatrix}. \tag{43}$$

Okui, Small, Tan, and Robins (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 Robins (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 Robins (2012) DR construction, we could estimate this model using the ODR estimator, equation (1), with G and H given by equations (42) and (43). To satisfy the required overidentification (Assumption A3), $r_1(X)$ and $r_2(X)$ need 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 Robins (2012).