

A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators

Daniel Ackerberg
UCLA

Xiaohong Chen
Yale University

Jinyong Hahn*
UCLA

First Version: March 20, 2009
This Version: September 29, 2009

Abstract

The goal of this paper is to develop techniques to simplify semiparametric inference. We do this by deriving a number of numerical equivalence results. These illustrate that in many cases, one can obtain estimates of semiparametric variances using standard formulas derived in the already-well-known parametric literature. This means that for computational purposes, an empirical researcher can ignore the semiparametric nature of the problem and do all calculations “as if” it were a parametric situation. We hope that this simplicity will promote the use of asymptotic semiparametric variance estimates.

1 Introduction

Many recently introduced empirical methodologies utilize two-step semiparametric estimation approaches. In the first step, certain functions are estimated nonparametrically. In the second step, structural/causal parameters are estimated parametrically, using the nonparametric estimates from the first stage as inputs. Such estimators have been used both in the treatment effect literature to estimate average treatment effects (e.g. Hahn (1998), and Hirano, Imbens, and Ridder (2003)) and in the Labor and IO literatures to estimate rich, often dynamic, structural models (Hotz and Miller (1993, 1994), Olley and Pakes (1995), Aguirregabiria and Mira (2002, 2007), Jofre-Bonet and Pesendorfer (2003), Bajari, Benkard, and Levin (2007), Pakes, Ostrovsky, and Berry (2007), Pesendorfer and Schmidt-Dengler (2007), and Bajari, Hong, Krainer, and Nekipelov (2008)).

These two-step semiparametric estimators are often very computationally convenient to use. This is particularly true in the structural models cited above. As argued by that literature, these two-step methods allow one to estimate the underlying structural parameters without having to repeatedly solve for complicated (and often dynamic) equilibrium. This can tremendously reduce computational burden. As such, these estimators are being used quite regularly in applications (e.g. Ryan (2006), Collard-Wexler (2006), Dunne, Klimek, Roberts, and Xu (2006), Sweeting (2007), Macieira (2008), Ellickson and Misra (2008), Snider (2008), Ryan and Tucker (2008)).

These methods do, however, rely crucially on being nonparametric in the first step. The first step involves estimating reduced form policy functions that arise from the equilibria of the underlying structural model. From a practical perspective, there is a sense in which the nonparametric first step estimation is parametric - since one needs to choose, e.g. the number of terms in a series approximation, the bandwidth of a kernel, or the flexibility of a sieve. But naïve parametric specification of these reduced form policy functions is likely to contradict the

*Thanks to Victor Aguirregabiria, Lanier Benkard, Richard Blundell, Jeremy Fox, Bryan Graham, Phil Haile, Jim Heckman, Guido Imbens, Pat Kline, Pedro Mira, Whitney Newey, Jim Powell, Geert Ridder, and Jeff Wooldridge for helpful comments. All remaining errors are our own.

underlying structural model.¹ So, researchers have to take seriously the “nonparametric promise” of increasing the flexibility of the first-step specification as the number of observations increases.

Taking this promise seriously also requires one to explicitly consider the problem’s semiparametric nature when estimating the variances of the estimated finite-dimensional (structural) parameters. Asymptotic variances in parametric models are based on the assumption that the total number of parameters stays fixed as the number of observations increases. In contrast, asymptotic variances in semiparametric models are based on the assumption that the total number of parameters increases as the number of observations increase. These asymptotic variances will generally be different.

There is a long line of theoretical literature that derives expressions for semiparametric asymptotic variances of two-step estimators (Newey (1994), Andrews (1994), Newey and McFadden (1994), Ai and Chen (2007), Chen, Linton and van Keilegom (2003), Ichimura and Lee (2008), to name a few). Some of these papers also show how to consistently estimate the asymptotic variances. While these theoretical results are useful, their implementation is typically not straightforward in practice. Some (e.g. Newey (1994), Chen, Linton and van Keilegom (2003)) give very high level results on asymptotic variance expressions, requiring an applied researcher to calculate the asymptotic variance in a closed-form in order to apply the theoretical result to a particular situation.² Others (e.g., Ai and Chen (2007)) are very general and do not need to solve the asymptotic variance analytically, but require numeric optimization to compute a consistent estimate of the asymptotic variance. These limitations have often lead applied researchers to use the bootstrap to estimate asymptotic variances (e.g., Ryan (2006), Ellickson and Misra (2008), Macieira (2008)), but this can be computationally demanding and may also be difficult to justify theoretically. Bootstrap validity is typically established for confidence region construction. Even for parametric linear regressions, one needs additional regularity conditions to justify bootstrap validity for standard errors (see, e.g., Gonçalves and White (2005) for a recent discussion).

The purpose of this paper is to show that in a large class of models, one can greatly simplify the estimation of semiparametric asymptotic variances. The core point of our paper is a numerical equivalence result. To describe this, consider researcher A, who estimates the model with a parametric first step. Also consider researcher B, who estimates the model semi-parametrically, using the method of sieves as the nonparametric first step. Since sieves are just “sufficiently flexible” parameterized functions, let us assume that researcher B’s sieve is identical to researcher A’s parameterized function for the first step.

Given this choice of sieve, it is clear that researcher A and researcher B will obtain identical point estimates of the structural parameters. On the other hand, the asymptotic variances of the two estimators will be different, as researcher A is in a parametric world where the total number of unknown parameters is constant (and finite), while researcher B is in a semiparametric world where the total number of unknown parameters is increasing to infinity.

Our results concern the *estimated* asymptotic variance of the structural parameters. We show, somewhat surprisingly, that in a large class of models, the estimate of the *semiparametric* asymptotic variance using the methods of Newey (1994) or Ai and Chen (2007) is *numerically identical* to the estimate of *parametric* asymptotic variance using standard two-step parametric results (described in Section 2, see, e.g. Murphy and Topel (1985), or Newey and McFadden (1994)). In other words, researcher A and researcher B will obtain numerically identical variance estimates (for the structural parameters). This is true even though they are estimating different objects asymptotically – the true asymptotic parametric variance vs. the true asymptotic semiparametric variance of the finite dimensional parameters of interest. To the best of our knowledge, Newey (1994, Section 6) was the first to recognize this equivalence in a simple example involving one infinite-dimensional parameter, which is estimated by least squares using a series approximation in the first step.³ We go one step further and generalize

¹Imposing the structure of the underlying model on the reduced form policy functions would necessitate solving for the equilibrium, which is exactly what these methods are trying to avoid.

²As illustrated in Newey (1994) and Ichimura and Lee (2008), one can typically solve the asymptotic variance analytically when there is only a scalar unknown function estimated in the first stage. However, it is extremely difficult to calculate the asymptotic variance analytically when there are several unknown functions involved. Recognizing this difficulty, Chen, Linton and van Keilegom (2003) provide nonparametric bootstrap justification for the construction of confidence regions.

³Imbens and Wooldridge (2005) conjectured an equivalence in propensity score estimation.

his insight to other classes of two step semiparametric estimators, including models with multiple nonparametric components, models characterized by likelihoods, and models where the second step moments depend on the first step infinite-dimensional parameter in a more complicated way.

The implication of this numerical equivalence result is that it becomes very simple for applied researchers to take the nonparametric portion of the above models seriously. One of course has to make the nonparametric promise that one will increase the flexibility of the nonparametric approximation as the number of observations increases (which probably should be interpreted in practice as making the approximation sufficiently flexible given the number of observations). However, to compute the estimated variances and standard errors, one can simply use regular parametric formulas rather than either understanding Newey (1994) or Ai and Chen (2007), or bootstrapping. These regular parametric variance formulas will generate consistent estimates of the semiparametric asymptotic variances. In some sense, our result says that for computational purposes, a researcher can ignore the nonparametric aspect of the problem and do all calculations “as if” it were a parametric situation. We hope that this simplicity will promote the use of asymptotic semiparametric variance estimates, and alleviate the need for computationally burdensome bootstrapping.

It is well known that the two step semi-parametric estimator is typically \sqrt{n} consistent and asymptotically normal, but the existing literature seems to take the view that (i) it is necessary to characterize the analytic expression of the asymptotic variance matrix; (ii) such characterization is possible but difficult; and therefore (iii) statistical inference can be a challenge in practice. Hotz and Miller’s (1993) original two-step estimator presented in their equation (5.9) is in fact a parametric estimator because they assume that the nonparametric component is a conditional expectation on some variable with *finite* support. They consider an extension to the case where the conditioning variable is continuous, and end up developing a new estimator. Olley and Pakes (1996) devoted another paper (Pakes and Olley (1995)) to characterizing the analytic formula of the asymptotic variance (for the case where the first step nonparametric estimation is done by kernel methods), and noted “We do not currently know of a theorem that insures \sqrt{n} consistency and asymptotic normality when the series estimator is used...” Bajari, Benkard, and Levin (2007) noted the difficulty of characterizing the semiparametric asymptotic variance of their estimators, and “To simplify this second-stage problem, we henceforth assume that the policy function and transition probabilities are parameterized by a finite parameter vector α and that this vector can be consistently estimated at the first stage.” Noting a separate work by Bajari, Chernozhukov, and Hong (2005), they go on to state “We are optimistic that our approach could be shown to work for a nonparametric first stage with continuous actions on a continuous state space, but we leave this for future research.” Their conjecture was later verified by Bajari, Chernozhukov, Hong, and Nekipelov (2008). We do not disagree with the profession that analytic characterization of the asymptotic variance matrix can be challenging. Our contribution is to note that analytic characterization is often unnecessary in practice because a practitioner simply needs a consistent estimator of the asymptotic variance, and that such an estimator is very easy to compute by adopting the convenient view that the first step sieve nonparametric estimation is in practice a parametric step.⁴ This convenience has applicability across a wide range of literatures, including labor and IO applications.

Our result is about numerical equivalence between formulae that are derived in a wide array of semiparametric literature,⁵ and formulae from the simple parametric literature. We do not address the question of improving existing procedures for semiparametric models. Our numerical equivalence results may make some readers feel uncomfortable about existing semiparametric procedures. Some readers may feel that the choice of sieves and the number of terms to be used in the approximation, which have been buried in a list of regularity conditions, should be explicitly addressed. Readers may also feel that the existing estimators of variance in semiparametric models may have room for improvement given our equivalence result. These are questions that can be potentially

⁴Our numerical equivalence results are established for the two-step semiparametric estimators only when sieve (or series) methods are used in the first-step. We doubt such a numerical equivalence result might still hold for other nonparametric first-steps such as kernel, local linear regression, or nearest neighbor methods.

⁵The semiparametric formula in principle addresses nonparametric first step sieve estimation with potentially data dependent choice of the number of terms used in approximation.

addressed within the context of higher order analysis, and we leave it to future research.

We present numerical equivalence results for a set of models that cover many important cases in the empirical literature. Although a unifying theory that explains the equivalence could potentially be produced, we suspect that it would unnecessarily complicate the comparison. Section 2 starts with a review of how to compute standard errors in two-step parametric models. Sections 3 and 4 present our main numerical equivalence results for two classes of semiparametric models. For intuition, Section 5 presents a very simple example where our result holds, the partially linear semiparametric model of Robinson (1988). Section 6 presents some extensions and examples applying our result to some of the current models in the empirical literature. Section 7 briefly concludes.

2 Review: Standard Errors in Two-Step Parametric M-Estimators

In this section, we provide a brief review of how to estimate the asymptotic variance of two-step parametric M-estimators. We assume that a researcher estimates a parameter vector θ using a first-step M-estimator (e.g. OLS, NLLS, MLE, method of moments). This estimate is then plugged into a second-step M-estimator which is used to estimate another parameter vector β . The question is whether and how the estimation error of the first-step M-estimator $\hat{\theta}$ affects the asymptotic variance of the second-step M-estimator $\hat{\beta}$. To the best of our knowledge, Pagan (1984), Newey (1984), and Murphy and Topel (1985) were among the first to investigate this issue. These methods of adjusting the asymptotic variance of $\hat{\beta}$ are now so well-understood that they can even be found in standard textbooks such as Wooldridge (2002, Chapter 12.4).

Suppose that in the first step, a researcher estimates θ with the $\hat{\theta}$ that solves

$$\frac{1}{n} \sum_{i=1}^n \varphi(z_i, \hat{\theta}) = 0 \tag{1}$$

In the case where $\hat{\theta}$ solves some optimization problem, such as OLS, NLLS, or MLE, φ is the first order condition of the optimization problem. In the second step, the researcher estimates β by solving

$$\frac{1}{n} \sum_{i=1}^n \psi(z_i, \hat{\beta}, \hat{\theta}) = 0 \tag{2}$$

Note that the second step M-estimator $\hat{\beta}$ will in general be different from the $\tilde{\beta}$ that solves $\frac{1}{n} \sum_{i=1}^n \psi(z_i, \tilde{\beta}, \theta_*) = 0$, where θ_* denotes the true value of θ satisfying $E[\varphi(z_i, \theta_*)] = 0$. Therefore, the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_*)$ is in general different from that of $\sqrt{n}(\tilde{\beta} - \beta_*)$, due to the estimation error in $\hat{\theta}$.

In order to assess the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_*)$ that correctly reflects the estimation error of $\hat{\theta}$, a researcher can consider the two-step estimator as a component of a one-step M-estimator⁶

$$\frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\beta}, \hat{\theta}) = 0 \tag{3}$$

where

$$g(z_i, \beta, \theta) = \begin{bmatrix} \varphi(z_i, \theta) \\ \psi(z_i, \beta, \theta) \end{bmatrix}$$

The $\hat{\theta}$ and $\hat{\beta}$ that solve (3) are numerically identical to $\hat{\theta}$ and $\hat{\beta}$ that solve (1) and (2). Letting $\alpha = (\beta', \theta)'$ and recognizing that $\hat{\alpha} = (\hat{\beta}', \hat{\theta}')'$ is an M-estimator, we can then use standard arguments⁷ to compute the

⁶This formulation assumes exact identification.

⁷See Wooldridge (2002, Chapter 12.3).

asymptotic variance of $\sqrt{n}(\hat{\alpha} - \alpha_*)$ i.e. a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\alpha} - \alpha_*)$ is given by

$$\left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \hat{\alpha})}{\partial \alpha'} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\alpha}) g(z_i, \hat{\alpha})' \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g'(z_i, \hat{\alpha})}{\partial \alpha} \right)^{-1}.$$

The asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_*)$ is simply the upper left block of the asymptotic variance matrix of $\sqrt{n}(\hat{\alpha} - \alpha_*)$. This one-step interpretation is a device that facilitates our theoretical discussion. In practice, two-step estimation techniques are often adopted for computational convenience.

3 Estimator of Asymptotic Variance of Two-Step Semiparametric Estimators

We present our first main result in this section. We consider semiparametric two-step estimation, where a researcher estimates certain functions with a nonparametric estimator in the first-step. In the second-step, she plugs the nonparametric estimators into a parametric moment equation to compute an estimator $\hat{\beta}$ of some finite dimensional parameter vector. We assume that the first-step nonparametric estimation is implemented by the method of sieves, e.g. a series approximation. Note that the first-step requires computation of a finite dimensional parameter in practice. For example, if the first-step involves nonparametric estimation of a conditional expectation implemented with a series approximation, then the first step amounts to OLS in practice.

Now assume that there are two researchers. Researcher A makes an incorrect assumption that the first-step is in fact parametric, therefore believing that the number of terms in the series approximation remains constant as the sample size grows to infinity. Because she believes the first step to be a parametric procedure (and because the second step is truly parametric), Researcher A would estimate the asymptotic variance of $\hat{\beta}$ using the formula discussed in Section 2.

Researcher B, on the other hand, makes the correct nonparametric assumption that the number of terms in the series approximation increases to infinity as an appropriate function of the sample size. Therefore, Researcher B would like to compute a consistent estimator of the asymptotic variance of $\hat{\beta}$ using a formula that correctly reflects $\hat{\beta}$'s semiparametric nature. Because the two researchers are considering different asymptotic sequences, Researcher A's asymptotic variance formula (i.e., the theoretical formula expressed in population expectations) will generally be different from Researcher B's. In other words, Researcher A is trying to estimate a different theoretical variance object than Researcher B.⁸ Despite this difference, this section proves that the *estimator* of the asymptotic variance that Researcher A implements will be *numerically equivalent* to the *estimator* of the asymptotic variance that Researcher B uses.

We use Ai and Chen's (2007) asymptotic variance formula extensively. To the best of our knowledge, Newey (1994) was the first to characterize the asymptotic variance of a two-step semiparametric estimator. In Appendix C, we show that Newey's asymptotic variance formula can also be given this "parametric" interpretation. In Appendix D, we explain how Newey's asymptotic variance formula can be related to Ai and Chen's.

⁸Researcher A is trying to estimate a theoretical object that is not the true asymptotic variance, since she believes that the number of terms in the series will remain constant in her asymptotics. In fact, Researcher A's estimator $\hat{\beta}$ in the second step will be inconsistent in general because her first step estimator will not converge to the true nonparametric object.

Description of the model and estimator More formally, consider a model given by the following moment restrictions

$$\begin{aligned} E[y_{1i} - h_{1*}(x_{1i}) | x_{1i}] &= 0, \\ &\vdots \\ E[y_{Li} - h_{L*}(x_{Li}) | x_{Li}] &= 0, \\ E[m(z_i, \beta_*, h_{1*}(x_{1i}), \dots, h_{L*}(x_{Li}))] &= 0. \end{aligned} \tag{4}$$

The $h_1(x_{1i}), \dots, h_L(x_{Li})$ functions are the nonparametric components in the model. β is the finite-dimensional component of the model. Note that the conditioning variables x_{1i}, \dots, x_{Li} are allowed to differ from each other. We also allow the dimensions of x_{1i}, \dots, x_{Li} to differ. The practitioner nonparametrically estimates $h_{1*}(x_{1i}), \dots, h_{L*}(x_{Li})$ with the estimators $\hat{h}_1(x_{1i}), \dots, \hat{h}_L(x_{Li})$, and then estimates β_* with the $\hat{\beta}$ that solves⁹

$$\frac{1}{n} \sum_{i=1}^n m(z_i, \hat{\beta}, \hat{h}_1(x_{1i}), \dots, \hat{h}_L(x_{Li})) = 0.$$

Ai and Chen’s (2007) modified SMD estimator We now show that this two-step estimator is numerically identical to Ai and Chen’s (2007) modified SMD estimator as long as $\hat{h}_1(x_{1i}), \dots, \hat{h}_L(x_{Li})$ are approximated using the method of sieves.¹⁰ A more applied reader can skip these details in this and the next subsections. The modified SMD estimator solves the minimization problem

$$\frac{1}{n} \sum_{i=1}^n (y_{1i} - h_1(x_{1i}))^2 + \dots + \frac{1}{n} \sum_{i=1}^n (y_{Li} - h_L(x_{Li}))^2 + \left\| \frac{1}{n} \sum_{i=1}^n m(z_i, \beta, h_1(x_{1i}), \dots, h_L(x_{Li})) \right\|^2$$

over $(\beta, h_1, \dots, h_L) \in \mathcal{B} \times \mathcal{H}_{1,n} \times \dots \times \mathcal{H}_{L,n}$, where $\|a\|$ denotes a vector norm such that $\|a\| = a'a$. Assuming that \mathcal{B} is a compact subset of R^d , and for $l = 1, \dots, L$, the sieve spaces $\mathcal{H}_{l,n}$ are given by:

$$\mathcal{H}_{l,n} = \{h_l : h_l(x_l) = p_{l,1}(x_l)\theta_{(l),1} + \dots + p_{l,K_{l,n}}(x_l)\theta_{(l),K_{l,n}} = h_l(x_l, \theta_{(l)})\}, \tag{5}$$

we can see that the modified SMD is numerically equivalent to the following multi-step estimator:

$$\begin{aligned} \hat{\theta}_{(l)} &= \underset{\theta_{(l),1}, \dots, \theta_{(l),K_{l,n}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_{li} - (p_{l,1}(x_l)\theta_{(l),1} + \dots + p_{l,K_{l,n}}(x_l)\theta_{(l),K_{l,n}}))^2, \quad l = 1, \dots, L, \\ 0 &= \frac{1}{n} \sum_{i=1}^n m(z_i, \hat{\beta}, h_1(x_{1i}, \hat{\theta}_{(1)}), \dots, h_L(x_{Li}, \hat{\theta}_{(L)})). \end{aligned}$$

Ai and Chen’s (2007) estimator of asymptotic variance Ai and Chen (2007) show that $\hat{\beta}$ is \sqrt{n} -consistent and asymptotically normal under certain regularity conditions. They also provide a consistent estimator of the semiparametric asymptotic variance (V) of $\sqrt{n}(\hat{\beta} - \beta_*)$, which we now describe. For simplicity of notation, we will write

$$r(z_i, \alpha_*) = \begin{bmatrix} y_{1i} - h_{1*}(x_{1i}) \\ \vdots \\ y_{Li} - h_{L*}(x_{Li}) \end{bmatrix} \tag{6}$$

where $\alpha_* = (\beta_*, h_*)$, and h is an abbreviation of (h_1, \dots, h_L) . We adopt a similar convention for \hat{h} . Denote $\hat{\alpha} = (\hat{\beta}, \hat{h})$. Assuming the sieve space $\mathcal{H}_n = \mathcal{H}_{1,n} \times \dots \times \mathcal{H}_{L,n}$ with $\mathcal{H}_{l,n}$ given by (5) for $l = 1, \dots, L$, Ai and Chen’s estimator \hat{V} of the asymptotic variance of $\hat{\beta}$ can be computed using the following algorithm:

⁹For simplicity we assume exact identification for β .

¹⁰See their equation (5) or their plug-in estimation equations (6)-(7). In fact, Ai and Chen (2007) consider a much broader class of models, including misspecified semi/nonparametric models. Our discussion here is a “translation” of their procedure for the specific model we consider here.

1. Compute $\widehat{\mathbf{w}}^* = (\widehat{w}_1^*, \dots, \widehat{w}_d^*)$ that solves for $j = 1, \dots, d$,

$$\widehat{w}_j^* = \operatorname{argmin}_{w \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left\{ \left(\frac{\partial r(z_i, \widehat{\alpha})}{\partial \beta_j} - \sum_{l=1}^L \frac{\partial r(z_i, \widehat{\alpha})}{\partial h_l} w_{j,l}(x_{l,i}) \right)' \left(\frac{\partial r(z_i, \widehat{\alpha})}{\partial \beta_j} - \sum_{l=1}^L \frac{\partial r(z_i, \widehat{\alpha})}{\partial h_l} w_{j,l}(x_{l,i}) \right) + \left\| \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta_j} - \sum_{l=1}^L \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_l} w_{j,l}(x_{l,i}) \right) \right\|^2 \right\}.$$

2. Compute

$$\rho(z_i, \widehat{\alpha}) = \begin{bmatrix} r(z_i, \widehat{\alpha}) \\ m(z_i, \widehat{\alpha}) \end{bmatrix},$$

$$\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) = \begin{bmatrix} \sum_{l=1}^L \left(\frac{\partial r(z_i, \widehat{\alpha})}{\partial \beta_1} - \sum_{l=1}^L \frac{\partial r(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_{1,l}^*(x_{l,i}) \right) & \cdots & \sum_{l=1}^L \left(\frac{\partial r(z_i, \widehat{\alpha})}{\partial \beta_d} - \sum_{l=1}^L \frac{\partial r(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_{d,l}^*(x_{l,i}) \right) \\ \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta_1} - \sum_{l=1}^L \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_{1,l}^*(x_{l,i}) \right) & \cdots & \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta_d} - \sum_{l=1}^L \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_{d,l}^*(x_{l,i}) \right) \end{bmatrix}$$

and

$$\widehat{\Omega} = \frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)' \rho(z_i, \widehat{\alpha}) \rho(z_i, \widehat{\alpha})' \widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i)$$

3. Compute

$$\widehat{V} = \left(\frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)' \widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)^{-1} \widehat{\Omega} \left(\frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)' \widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)^{-1}.$$

A Naïve practitioner's estimator We now consider how the semiparametric estimators $\widehat{\beta}$ and \widehat{V} relate to what one obtains if the estimation problem is approached from a purely parametric perspective (i.e. Researcher A). First, note that a parametric estimator based on the parametric specification $h_l(x_l) = p_{l,1}(x_l)\theta_{(l),1} + \dots + p_{l,K_l}(x_l)\theta_{(l),K_l} = h_l(x_l, \theta_{(l)})$ (where $K_l = K_{l,n}$ is a function of n although it is perceived to be fixed for our fictitious Researcher A) will result in an estimate of β that is numerically equivalent to $\widehat{\beta}$. This means that for the purpose of computing $\widehat{\beta}$, it is harmless to “pretend” that the h_l 's are parametrically specified. We now show that the same idea holds for the estimated variance.

Our parametric Researcher A perceives $\widehat{\beta}$ to be a simple M-estimator solving the moment equation $E[g(z_i, \beta_*, \theta_*)] = 0$, where

$$g(z_i, \alpha) = \begin{bmatrix} p_1^{K_1}(x_{1,i})(y_{1i} - h_1(x_{1i}, \theta_{(1)})) \\ \vdots \\ p_L^{K_L}(x_{L,i})(y_{Li} - h_L(x_{Li}, \theta_{(L)})) \\ m(z_i, \beta, h_1(x_{1i}, \theta_{(1)}), \dots, h_L(x_{Li}, \theta_{(L)})) \end{bmatrix},$$

where $\alpha = (\beta', \theta)'$, $\theta = (\theta'_{(1)}, \dots, \theta'_{(L)})'$, and for $l = 1, \dots, L$, $h_l(x_{li}, \theta_{(l)}) = p_l^{K_l}(x_{li})'\theta_{(l)}$ with $p_l^{K_l}(x_{li}) = (p_{l,1}(x_{li}), \dots, p_{l,K_l}(x_{li}))'$. Here both β and θ are finite dimensional parameters such that $\dim(g) = \dim(\beta) + \dim(\theta)$. A consistent estimator of variance matrix of all the parameters is given by the usual formula

$$\left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \widehat{\alpha})}{\partial \alpha'} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n g(z_i, \widehat{\alpha}) g(z_i, \widehat{\alpha})' \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \widehat{\alpha})}{\partial \alpha} \right)^{-1} \quad (7)$$

and like in Section 2 an estimator \widehat{V}_p of the parametric asymptotic variance of $\widehat{\beta}$ can be obtained from the upper left corner of (7).

Numerical equivalence Note that \widehat{V}_p is obtained from a completely different perspective than the one underlying \widehat{V} . In fact, the idea that led to \widehat{V}_p is wrong! However, Appendix B shows that \widehat{V}_p is numerically identical to \widehat{V} . While subtle, this has a profound consequence for semiparametric statistical inference. Researchers wanting (or needing) to do semiparametric inference need not explicitly consider the semiparametric nature of the

problem in estimation. After specifying the flexible series approximation, they can proceed as if the problem was completely parametric for the purpose of inference on β . Obviously, this does not necessarily mean that the same is true for inference on the nonparametric components of the problem.

4 Estimator of Asymptotic Variance of Sieve MLE

In this section, we consider consistent estimation of the asymptotic variances of sieve maximum likelihood estimators (MLE). We assume that an econometric model is characterized by a probability density with two kinds of parameters: finite dimensional parameters β and some unknown functions $h(\cdot)$. We estimate (β, h) by sieve maximum likelihood in which h is approximated by finite dimensional flexible parametric families. This implies that the estimator of (β, h) is in fact identical to the maximizer of a (potentially) misspecified parametric likelihood. As in Section 3, we show that the estimator of the asymptotic variance of the parametric component can be given a parametric interpretation.

Assume that we observe z_i for each individual. We further assume that z_i are independent and identically distributed.¹¹ The log likelihood of the data $\{z_i\}_{i=1}^n$ is given by $\frac{1}{n} \sum_{i=1}^n \ell(z_i, \beta, h(\cdot))$, where $\beta \in \mathcal{B}$ is a vector of finite-dimensional parameter of interest and $h \in \mathcal{H}$ is a vector of L real-valued unknown functions (i.e., $h(\cdot) = (h_1(\cdot), \dots, h_L(\cdot))$ and each $h_l(\cdot)$ could depend on different argument x_l for $l = 1, \dots, L$). We take $h(\cdot)$ to be the nonparametric nuisance functions. Denote $\alpha = (\beta, h) \in \mathcal{B} \times \mathcal{H}$. We assume that the true parameter value $\alpha_* = (\beta_*, h_*) \in \mathcal{B} \times \mathcal{H}$ uniquely solves the population problem

$$\sup_{(\beta, h) \in \mathcal{B} \times \mathcal{H}} E[\ell(z_i, \beta, h(\cdot))].$$

The sieve MLE is a sample counterpart, except that the function parameter space $\mathcal{H} = \mathcal{H}_1 \times \dots \times \mathcal{H}_L$ is replaced by a sieve parameter space $\mathcal{H}_n = \mathcal{H}_{1,n} \times \dots \times \mathcal{H}_{L,n}$. In other words, the sieve MLE $(\hat{\beta}, \hat{h})$ is the solution to

$$\max_{(\beta, h) \in \mathcal{B} \times \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \ell(z_i, \beta, h(\cdot)). \quad (8)$$

Shen's result (1997) implies that $\hat{\beta}$ is \sqrt{n} -consistent, asymptotically normal and semiparametrically efficient (under regularity conditions), i.e., $\sqrt{n}(\hat{\beta} - \beta_*) \rightarrow N(0, \mathcal{I}^{-1})$, with \mathcal{I}^{-1} given in Appendix E.

In the following for $w = (w_{(1)}, \dots, w_{(L)}) \in \mathcal{H}_n = \mathcal{H}_{1,n} \times \dots \times \mathcal{H}_{L,n}$, we let $\frac{d\ell(z_i, \hat{\alpha})}{dh}[w] = \sum_{l=1}^L \frac{d\ell(z_i, \hat{\alpha})}{dh_l}[w_{(l)}]$, and $\frac{d\ell(z_i, \hat{\alpha})}{dh_l}[w_{(l)}]$ denotes the directional derivative of ℓ with respect to h_l at the direction $w_{(l)}$, loosely speaking. For more precise definition, see Appendix E. We propose the following simple sieve estimator of the asymptotic variance of the sieve MLE $\hat{\beta}$:

1. Compute a consistent estimator \hat{w}_j^* of w_j^* , $j = 1, \dots, d$:

$$\hat{w}_j^* = \operatorname{argmin}_{w \in \mathcal{H}_n} \sum_{i=1}^n \left(\frac{d\ell(z_i, \hat{\alpha})}{d\beta_j} - \frac{d\ell(z_i, \hat{\alpha})}{dh}[w] \right)^2.$$

2. Compute

$$\hat{\Delta}(z) = \begin{bmatrix} \frac{d\ell(z, \hat{\alpha})}{d\beta_1} - \frac{d\ell(z, \hat{\alpha})}{dh}[\hat{w}_1^*] \\ \vdots \\ \frac{d\ell(z, \hat{\alpha})}{d\beta_d} - \frac{d\ell(z, \hat{\alpha})}{dh}[\hat{w}_d^*] \end{bmatrix}.$$

3. Compute

$$\hat{V}_{smle} = \left(\frac{1}{n} \sum_{i=1}^n \hat{\Delta}(z_i) \hat{\Delta}(z_i)' \right)^{-1}. \quad (9)$$

¹¹In other words, we do not need to worry about the dependence as in Chen and Shen (1998).

In Appendix E, we provide a proof that \widehat{V}_{smle} is a consistent estimator of \mathcal{I}^{-1} (the asymptotic variance of the sieve MLE $\widehat{\beta}$).¹²

We now discuss the practical implication of (9). Consider a fictitious practitioner who assumes that \mathcal{H} is in fact \mathcal{H}_n . In other words, the practitioner believes that h can be and should be parametrically specified. In terms of estimating (β, h) , this fictitious practitioner's estimator would be numerically identical to ours. After all, he will solve the same problem (8). Would his standard error for $\widehat{\beta}$ be identical to ours?

We now show that the practitioner's estimator of the asymptotic variance is identical to ours if the sieve space is $\mathcal{H}_n = \mathcal{H}_{1,n} \times \cdots \times \mathcal{H}_{L,n}$ with $\mathcal{H}_{l,n}$ given by (5) for $l = 1, \dots, L$. The practitioner would write

$$h_l(x_l) = p_{l,1}(x_l)\theta_{(l),1} + \cdots + p_{l,K_l}(x_l)\theta_{(l),K_l} = p_l^{K_l}(x_l)'\theta_{(l)} \quad \text{for } \theta_{(l)} = (\theta_{(l),1}, \dots, \theta_{(l),K_l})'$$

with $p_l^{K_l}(x_l) = (p_{l,1}(x_l), \dots, p_{l,K_l}(x_l))'$, where $K_l = K_{l,n}$ is a function of n although it is perceived to be fixed for our fictitious practitioner. Denote $\theta = (\theta'_{(1)}, \dots, \theta'_{(L)})'$ which is a $K \times 1$ -vector with $K = K_1 + \cdots + K_L$. The parametric practitioner would estimate $(\beta_*, \theta_*) = \arg\max_{\beta, \theta} E[\ell(z_i, \beta, \theta)]$ via parametric MLE, and obtain:

$$\sqrt{n} \left(\widehat{\beta} - \beta_*, \widehat{\theta} - \theta_* \right)' \rightarrow N \left(0, \begin{bmatrix} E \left[\frac{d\ell(z_i, \beta_*, \theta_*)}{d\beta} \frac{d\ell(z_i, \beta_*, \theta_*)}{d\beta'} \right] & E \left[\frac{d\ell(z_i, \beta_*, \theta_*)}{d\beta} \frac{d\ell(z_i, \beta_*, \theta_*)}{d\theta'} \right] \\ E \left[\frac{d\ell(z_i, \beta_*, \theta_*)}{d\theta} \frac{d\ell(z_i, \beta_*, \theta_*)}{d\beta'} \right] & E \left[\frac{d\ell(z_i, \beta_*, \theta_*)}{d\theta} \frac{d\ell(z_i, \beta_*, \theta_*)}{d\theta'} \right] \end{bmatrix}^{-1} \right),$$

and the asymptotic variance for $\widehat{\beta}$, V_p , is simply the upper-left block of the above variance and covariance matrix, which can be computed by the partitioned inverse formula. The partitioned inverse formula on the other hand, has another interpretation as the inverse of the variance of the least squares projection residual of $\frac{d\ell(z_i, \beta_*, \theta_*)}{d\beta}$ on $\frac{d\ell(z_i, \beta_*, \theta_*)}{d\theta'}$:

$$V_p = \left(E \left[\Delta_p(z_i) \Delta_p(z_i) \right]' \right)^{-1},$$

where

$$\Delta_p(z) = \begin{bmatrix} \frac{d\ell(z, \beta_*, \theta_*)}{d\beta_1} - \frac{d\ell(z, \beta_*, \theta_*)}{d\theta'} c_1^* \\ \vdots \\ \frac{d\ell(z, \beta_*, \theta_*)}{d\beta_d} - \frac{d\ell(z, \beta_*, \theta_*)}{d\theta'} c_d^* \end{bmatrix}$$

and

$$c_j^* = \operatorname{argmin}_{c_j \in R^K} E \left[\left(\frac{d\ell(z, \beta_*, \theta_*)}{d\beta_j} - \frac{d\ell(z, \beta_*, \theta_*)}{d\theta'} c_j \right)^2 \right] \quad \text{for } j = 1, \dots, d.$$

If the practitioner uses the outer-product based estimator of the information matrix, then the asymptotic variance matrix for $(\widehat{\beta}, \widehat{\theta})'$ can be consistently estimated by the following matrix:

$$\begin{bmatrix} \frac{1}{n} \sum_{i=1}^n \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\beta} \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\beta'} & \frac{1}{n} \sum_{i=1}^n \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\beta} \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta'} \\ \frac{1}{n} \sum_{i=1}^n \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta} \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\beta'} & \frac{1}{n} \sum_{i=1}^n \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta} \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta'} \end{bmatrix}^{-1},$$

and the asymptotic variance for $\widehat{\beta}$ can be consistently estimated by the upper-left block of the above matrix, which can be computed by the partitioned inverse formula, which also has another interpretation that can be characterized by the following algorithm:

¹²We provide a proof of the consistency of \widehat{V}_{smle} along with regularity conditions in Appendix E because we are not aware of any published papers that establish the consistency of \widehat{V}_{smle} , albeit such an estimator has been used in the literature without proofs; see, e.g., Chen (2007, remark 4.2), Chen, Fan and Tsyrennikov (2006). For most other results in this paper, we do not provide any rigorous asymptotic theory, which is already done in the existing literature.

1. Compute the solution \widehat{c}_j^* to

$$\min_{c_j \in R^K} \sum_{i=1}^n \left(\frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\beta_j} - \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta'} c_j \right)^2.$$

2. Compute

$$\widehat{\Delta}_p(z_i) = \begin{bmatrix} \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\beta_1} - \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta'} \widehat{c}_1^* \\ \vdots \\ \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\beta_d} - \frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta'} \widehat{c}_d^* \end{bmatrix}.$$

3. Compute

$$\widehat{V}_p = \left(\frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_p(z_i) \widehat{\Delta}_p(z_i)' \right)^{-1}.$$

We argue that \widehat{V}_p is in fact numerically identical to \widehat{V}_{smle} , since $\widehat{\Delta}_p(z_i)$ is numerically identical to $\widehat{\Delta}(z)$. For this purpose, it suffices to note that with $h_l(x_l) = p_l^{K_l}(x_l)' \theta_{(l)}$, $\theta = (\theta'_{(1)}, \dots, \theta'_{(L)})'$ and $c = (c'_{(1)}, \dots, c'_{(L)})'$, we have:

$$\frac{d\ell(z_i, \widehat{\beta}, \widehat{\theta})}{d\theta'} c = \sum_{l=1}^L \frac{d\ell(z_i, \widehat{\beta}, \widehat{h})}{dh_l} p_l^{K_l}(\cdot)' c_{(l)}$$

Therefore, the minimization problem over $c \in R^K$ is in fact identical to the minimization problem over all linear combinations $w_{(l)} = p_l^{K_l}(\cdot)' c_{(l)}$, which in turn is identical to the minimization over $w = (w_{(1)}, \dots, w_{(L)}) \in \mathcal{H}_n = \mathcal{H}_{1,n} \times \dots \times \mathcal{H}_{L,n}$. It follows that the variance estimator \widehat{V}_p obtained from the pretension that the model is parametrically specified is exactly identical to the sieve variance estimator \widehat{V}_{smle} obtained under the correct assumption that the model is semiparametrically specified.

We conclude that, as long as outer-product is used for calculation of information, the “parametric” inference is numerically identical to semiparametric inference.

5 Intuition: Partially Linear Regression Model

In order to gain intuition of our general result, it is useful to consider the partially linear regression model

$$y_i = x_i \beta_* + h(s_i) + u_i$$

where h is a scalar-valued unknown function. For simplicity of notation, we assume that β is a scalar.

This model is usually estimated by the following steps:

1. Compute nonparametric estimators $\widehat{E}[y|s]$ and $\widehat{E}[x|s]$ of $E[y|s]$ and $E[x|s]$
2. Compute $\widehat{\beta}$ using an OLS regression of $y_i - \widehat{E}[y|s_i]$ on $x_i - \widehat{E}[x|s_i]$

It has been shown that $\sqrt{n}(\widehat{\beta} - \beta_*)$ is asymptotically normal under certain regularity conditions; see, e.g., Robinson (1988). Assuming that the unobserved error u_i is independent of (x_i, s_i) , the asymptotic variance is equal to

$$\frac{\text{Var}(u_i)}{E[(x_i - E[x|s_i])^2]} \quad (10)$$

This result can be given a “parametric” interpretation. Suppose that a practitioner believes that

$$h(s) = h(s, \theta) = p_1(s) \theta_1 + \dots + p_K(s) \theta_K$$

Then the practitioner’s estimator can be computed by the partitioned regression formula:

1. Run OLS regressions of both y_i and x_i on $p_1(s_i), \dots, p_{K_n}(s_i)$, and obtain the fitted values $\tilde{E}[y|s_i]$ and $\tilde{E}[x|s_i]$
2. Compute $\tilde{\beta}$ using an OLS regression of $y_i - \tilde{E}[y|s_i]$ on $x_i - \tilde{E}[x|s_i]$

$\tilde{\beta}$ will be numerically identical to $\hat{\beta}$ if the nonparametric estimation of $\hat{E}[y|s]$ and $\hat{E}[x|s]$ computed as sieve estimators over $\mathcal{H}_n = \{h : h(s) = p_1(s)\theta_1 + \dots + p_{K_n}(s)\theta_{K_n}\}$. What about the asymptotic variance estimator? Note that the asymptotic variance (10) can be estimated by

$$\frac{\widehat{\text{Var}}(u_i)}{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{E}[x|s_i])^2}$$

where $\widehat{\text{Var}}(u_i)$ is some consistent estimator of $\text{Var}(u_i)$. But since in this case, $\hat{E}[x|s_i]$ will be identical to $\tilde{E}[x|s_i]$, we conclude that it is numerically identical to

$$\frac{\widehat{\text{Var}}(u_i)}{\frac{1}{n} \sum_{i=1}^n (x_i - \tilde{E}[x|s_i])^2} \tag{11}$$

which is the practitioner's estimator of the asymptotic variance of $\sqrt{n}(\tilde{\beta} - \beta_*)$, which is in turn numerically equivalent to the upper-left corner of the estimator of the asymptotic variance $(\tilde{\beta}, \tilde{\theta}')'$, where $\tilde{\theta}$ is obtained from the one-step regression of y_i on x_i and $p_1(s_i), \dots, p_{K_n}(s_i)$.

How is this equivalence related to our results in Sections 3 and 4? We can see that $\hat{\beta}$ can be given two different interpretations. First, we can understand it as a two-step semiparametric estimator with a nonparametric first-step. As such, our result in Section 3 explains the numerical equivalence between the estimated variances. Second, if u is assumed to have a normal distribution, we can treat $\hat{\beta}$ as a component of the sieve MLE using the standard equivalence of OLS to pseudo-MLE with normal errors. Therefore, our result in Section 4 also explains the numerical equivalence.

6 Extensions and Examples

In the first three subsections of this section, we present three simple extensions to cover models that are commonly seen in applied microeconometrics. In the last two subsections, we discuss some specific examples that are commonly seen in labor and IO applications.

6.1 Dependence of Second-Stage on Full Non-Parametric Function

Consider a model where

$$\begin{aligned} E[y_i - h_*(x_i) | x_i] &= 0, \\ E[m(z_i, \beta_*, h_*)] &= 0. \end{aligned}$$

Note the important difference between this model and the model of Section 3. In this model, the moment equation $m(z_i, \beta_*, h_*)$ depends not only on h_* through its value at x_i but through its values at *all* support points of x_i . Does this change our conclusion? For simplicity of notation, we will assume that y_i is a scalar and h_* is a scalar-valued function.

This model still fits into the framework of Ai and Chen (2007). According to their asymptotic variance formula for their modified SMD estimator $\hat{\beta}$, to consider this model we simply have to replace the term

$\frac{\partial m(z_i, \hat{\alpha})}{\partial h} w_j(x_i)$ in Section 3 by $\frac{\partial m(z_i, \hat{\alpha})}{\partial h} [w_j(\cdot)]$, where the pathwise derivatives are defined as

$$\frac{\partial m(z_i, \hat{\alpha})}{\partial h} [h - \hat{h}] = \left. \frac{dm(z_i, \hat{\beta}, (1 - \tau)\hat{h} + \tau h)}{d\tau} \right|_{\tau=0}.$$

Let the sieve space be $\mathcal{H}_n = \{h : h(\cdot) = \theta_1 p_1(\cdot) + \dots + \theta_{K_n} p_{K_n}(\cdot)\}$. Ai and Chen's sieve estimator \hat{V} of the asymptotic variance of $\hat{\beta}$ can then be computed by the following algorithm:

1. Compute $\hat{\mathbf{w}}^* = (\hat{w}_1^*, \dots, \hat{w}_d^*)$ for $j = 1, \dots, d$ as

$$\hat{w}_j^* = \operatorname{argmin}_{w \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left\{ (-w_j(x_i))^2 + \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta_j} - \frac{\partial m(z_i, \hat{\alpha})}{\partial h} [w_j] \right) \right)^2 \right\}.$$

2. Compute

$$\rho(z_i, \hat{\alpha}) = \begin{bmatrix} y_i - \hat{h}(x_i) \\ m(z_i, \hat{\beta}, \hat{h}) \end{bmatrix},$$

$$\hat{\Delta}_{\hat{\mathbf{w}}^*}(z_i) = \begin{bmatrix} -\hat{w}_1^*(x_i) & \dots & -\hat{w}_d^*(x_i) \\ \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta_1} - \frac{\partial m(z_i, \hat{\alpha})}{\partial h} [\hat{w}_1^*] \right) & \dots & \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta_d} - \frac{\partial m(z_i, \hat{\alpha})}{\partial h} [\hat{w}_d^*] \right) \end{bmatrix}$$

and

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{\mathbf{w}}^*}(z_i) \right)' \rho(z_i, \hat{\alpha}) \rho(z_i, \hat{\alpha})' \hat{\Delta}_{\hat{\mathbf{w}}^*}(z_i).$$

3. Compute

$$\hat{V} = \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{\mathbf{w}}^*}(z_i) \right)' \hat{\Delta}_{\hat{\mathbf{w}}^*}(z_i) \right)^{-1} \hat{\Omega} \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{\mathbf{w}}^*}(z_i) \right)' \hat{\Delta}_{\hat{\mathbf{w}}^*}(z_i) \right)^{-1}.$$

Now assume that a practitioner takes a parametric perspective with $h_\theta(\cdot) = p_1(\cdot)\theta_1 + \dots + p_K(\cdot)\theta_K$, where $K = K_n$ is a function of n although it is perceived to be fixed for our fictitious practitioner. His moment equation is then $E[g(z_i, \beta_*, \theta_*)] = 0$ where

$$g(z_i, \beta, \theta) = \begin{bmatrix} p^K(x_i)(y_i - h_\theta(x_i)) \\ m(z_i, \beta, h_\theta) \end{bmatrix}$$

with

$$\frac{\partial g(z_i, \beta, \theta)}{\partial (\beta, \theta)'} = \begin{bmatrix} 0 & -p^K(x_i) p^K(x_i)' \\ \frac{\partial m(z_i, \beta, h_\theta)}{\partial \beta'} & \mathbf{m}(z_i, \beta, \theta)' \end{bmatrix},$$

where $\mathbf{m}(z_i, \alpha)' = [\mathbf{m}_1(z_i, \alpha), \dots, \mathbf{m}_K(z_i, \alpha)]$, and for $k = 1, \dots, K$,

$$\mathbf{m}_k(z_i, \alpha) \equiv \frac{\partial m(z_i, \alpha)}{\partial h} [p_k].$$

With this notation, it is easy to see that

$$\frac{\partial m(z_i, \hat{\alpha})}{\partial h} [\theta_1 p_1 + \dots + \theta_K p_K] = \sum_k \mathbf{m}_k(z_i, \hat{\alpha}) \theta_k = \mathbf{m}(z_i, \hat{\alpha})' \theta,$$

and the argument in Appendix B applies yet again. Thus we can conclude the upper-left $(\dim(\beta) \times \dim(\beta))$ block of the parametric variance estimator

$$\left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \hat{\alpha})}{\partial (\beta, \theta)'} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\alpha}) g(z_i, \hat{\alpha})' \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \hat{\alpha})}{\partial (\beta, \theta)'} \right)^{-1}$$

is *numerically identical* to Ai and Chen's (2007) asymptotic semiparametric variance estimator \hat{V} .

6.2 First Step with Restriction

As another extension, we can consider a model where

$$\begin{aligned} E[y_{1i} - h_*(x_{1,i}) | x_{1,i}] &= 0, \\ &\vdots \\ E[y_{Li} - h_*(x_{L,i}) | x_{L,i}] &= 0, \\ E[m(z_i, \beta_*, h_*(x_{1,i}), \dots, h_*(x_{L,i}))] &= 0, \end{aligned}$$

where the dimensions of x_{1i}, \dots, x_{Li} are restricted to be identical, and for simplicity we assume $h_*(\cdot)$ is a scalar-valued function. The practitioner estimates $h_*(\cdot)$ by the sieve estimator $\widehat{h}(\cdot)$ which minimizes

$$\frac{1}{n} \sum_{i=1}^n (y_{1i} - h(x_{1,i}))^2 + \dots + \frac{1}{n} \sum_{i=1}^n (y_{Li} - h(x_{L,i}))^2$$

over $\mathcal{H}_n = \{h : h(x) = p_1(x)\theta_1 + \dots + p_{K_n}(x)\theta_{K_n}\}$, and then solves

$$\frac{1}{n} \sum_{i=1}^n m(z_i, \widehat{\beta}, \widehat{h}(x_{1,i}), \dots, \widehat{h}(x_{L,i})) = 0.$$

This again fits into the modified SMD framework because minimization of

$$\frac{1}{n} \sum_{i=1}^n (y_{1i} - h(x_{1,i}))^2 + \dots + \frac{1}{n} \sum_{i=1}^n (y_{Li} - h(x_{L,i}))^2 + \left(\frac{1}{n} \sum_{i=1}^n m(z_i, \beta, h(x_{1,i}), \dots, h(x_{L,i})) \right)^2$$

over $(\beta, h) \in \mathcal{B} \times \mathcal{H}_n$ amounts to this two step estimation. Ai and Chen's asymptotic variance estimator for this case is described in Appendix F.

We now assume that a practitioner adopts a parametric specification $h(x) = p^K(x)' \theta$, where $K = K_n$ is a function of n although it is perceived to be fixed for our fictitious practitioner. The practitioner proceeds with the following two-step algorithm: first minimizes

$$\frac{1}{n} \sum_{i=1}^n (y_{1i} - p^K(x_{1,i})' \theta)^2 + \dots + \frac{1}{n} \sum_{i=1}^n (y_{Li} - p^K(x_{L,i})' \theta)^2,$$

and then solves

$$\frac{1}{n} \sum_{i=1}^n m(z_i, \widehat{\beta}, p^K(x_{1,i})' \widehat{\theta}, \dots, p^K(x_{L,i})' \widehat{\theta}) = 0.$$

Note that the practitioner's estimator is identical to the modified SMD estimator. Also, the practitioner's moment condition is then $E[g(z_i, \beta_*, \theta_*)] = 0$ where

$$g(z_i, \beta, \theta) = \begin{bmatrix} p^K(x_{1,i})(y_{1i} - h(x_{1,i}, \theta)) + \dots + p^K(x_{L,i})(y_{Li} - h(x_{L,i}, \theta)) \\ m(z_i, \beta, h(x_{1,i}, \theta), \dots, h(x_{L,i}, \theta)) \end{bmatrix}$$

where $h(x_{li}, \theta) = p^K(x_{li})' \theta$. It follows that the practitioner's estimator of asymptotic variance is (7).

Again, it turns out that the practitioner's asymptotic variance is numerically identical to Ai and Chen's asymptotic variance estimator; see Appendix F for a proof. As before, we obtain the practical conclusion that researchers wanting to do semiparametric inference need not explicitly consider the semiparametric nature of the problem in estimation.

6.3 Nonparametric Sieve M-Estimation As First Step

Next consider semiparametric two-step estimation where the first-step involves nonparametric sieve, maximum-likelihood-like, M-estimation in the first step. Again, these nonparametric estimators are plugged into a parametric moment equation to compute an estimator $\hat{\beta}$ of some finite dimensional parameter in the second step. Note that the first step sieve M-estimation requires computation of a finite dimensional parameter in practice.

Suppose that the true structural parameters β_* and the unknown functions $h_*(\cdot)$ are identified by the following model:

$$h_* = \operatorname{argmax}_{h \in \mathcal{H}} E[\ell(z_i, h(\cdot))], \quad E[m(z_i, \beta_*, h_*(\cdot))] = 0,$$

where $\ell(z_i, h)$ is any criterion function and $h = (h_1, \dots, h_L)$ could be a vector of L unknown real-valued functions, each $h_l(\cdot)$ could depend on different argument for different $l = 1, \dots, L$. Note that this problem does not fit into the framework of Ai and Chen (2007).¹³ We propose the following sieve estimator:

$$(\hat{\beta}, \hat{h}) = \operatorname{argmax}_{(\beta, h) \in \mathcal{B} \times \mathcal{H}_n} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(z_i, h) + \frac{1}{2} \left\| \frac{1}{n} \sum_{i=1}^n m(z_i, \beta, h(\cdot)) \right\|^2 \right\},$$

which is equivalent to the following two-step semiparametric estimator:

$$\hat{h} = \operatorname{argmax}_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \ell(z_i, h(\cdot)), \quad 0 = \frac{1}{n} \sum_{i=1}^n m(z_i, \beta, \hat{h}(\cdot)).$$

It can be shown that $\hat{\beta}$ is \sqrt{n} -consistent and asymptotically normal under certain regularity conditions. See Appendix G for the characterization of the asymptotic variance V of $\sqrt{n}(\hat{\beta} - \beta_*)$.

We now suggest a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_*)$. In the following to simplify presentation we assume that β and h are scalars. Let the sieve space be $\mathcal{H}_n = \{h : h(\cdot) = p_1(\cdot)\theta_1 + \dots + p_{K_n}(\cdot)\theta_{K_n}\}$, a sieve estimator \hat{V} of the asymptotic variance V can be computed by the following algorithm:

1. Compute a consistent estimator \hat{w}^* :

$$\hat{w}^* = \operatorname{argmin}_{w \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left\{ \left(-\frac{\partial^2 \ell(z_i, \hat{h})}{\partial h \partial h} [w(\cdot), w(\cdot)] \right) + \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta} - \frac{\partial m(z_i, \hat{\alpha})}{\partial h} [w(\cdot)] \right) \right)^2 \right\}.$$

and

$$\hat{v}_{\hat{\beta}}^* = \left(\frac{1}{n} \sum_{i=1}^n \left\{ \left(-\frac{\partial^2 \ell(z_i, \hat{h})}{\partial h \partial h} [\hat{w}^*(\cdot), \hat{w}^*(\cdot)] \right) + \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta} - \frac{\partial m(z_i, \hat{\alpha})}{\partial h} [\hat{w}^*(\cdot)] \right) \right)^2 \right\} \right)^{-1}$$

2. Compute

$$\rho(z_i, \hat{\alpha}) = \begin{bmatrix} \frac{\partial \ell(z_i, \hat{\alpha})}{\partial h} [\hat{w}^*(\cdot)] \\ m(z_i, \hat{\alpha}) \end{bmatrix},$$

$$\hat{\Delta}_{\hat{w}^*}(z_i) = \begin{bmatrix} 1 \\ \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta} - \frac{\partial m(z_i, \hat{\alpha})}{\partial h} [\hat{w}^*(\cdot)] \right) \end{bmatrix},$$

and

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{w}^*}(z_i) \right)' \rho(z_i, \hat{\alpha}) \rho(z_i, \hat{\alpha})' \hat{\Delta}_{\hat{w}^*}(z_i).$$

¹³To our knowledge, the result below is new to the literature.

3. Compute

$$\widehat{V} = \widehat{v}_\beta^* \widehat{\Omega} \widehat{v}_\beta^*$$

As before, we note that the $\widehat{\beta}$ is numerically equivalent to the parametric estimator based on the parametric specification $h(\cdot) = p_1(\cdot)\theta_1 + \dots + p_K(\cdot)\theta_K$, where $K = K_n$ is a function of n although it is perceived to be fixed for our fictitious practitioner. For the purpose of computing $\widehat{\beta}$, it is harmless to pretend that h is parametrically specified. We now argue that the above sieve estimator \widehat{V} of the asymptotic variance of $\widehat{\beta}$ is again numerically identical to the well-known Murphy and Topel's (1985) formula.

Suppose that a researcher perceives the first-step sieve nonparametric estimation to be a parametric estimation. The researcher would perceive $\widehat{\beta}$ to be a simple parametric M-estimator solving the moment equation $E[g(z_i, \beta_*, \theta_*)] = 0$, where

$$g(z_i, \beta_*, \theta_*) = \begin{bmatrix} -\frac{\partial \ell(z_i, h(x_i, \theta))}{\partial h} p^K(\cdot) \\ m(z_i, \beta, h(\cdot, \theta)) \end{bmatrix}$$

and $h(\cdot, \theta) = p^K(\cdot)' \theta$. Here, both β and θ are finite dimensional parameters such that $\dim(g) = \dim(\beta) + \dim(\theta)$. A consistent estimator of $\widehat{\alpha} = (\widehat{\beta}, \widehat{\theta})'$ is given by the usual formula (which is (7):

$$\left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \widehat{\alpha})}{\partial \alpha'} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n g(z_i, \widehat{\alpha}) g(z_i, \widehat{\alpha})' \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \widehat{\alpha})'}{\partial \alpha} \right)^{-1}.$$

The estimator \widehat{V}_p of the asymptotic variance of $\widehat{\beta}$ is then obtained from the upper left corner of the above formula.

In Appendix G, it is shown that $\widehat{V} = \widehat{V}_p$. As before, we obtain the practical conclusion that researchers wanting to do semiparametric inference need not explicitly consider the semiparametric nature of the problem in estimation.

6.4 Example: Estimation of Average Treatment Effects

There is a large body of literature on estimation of average treatment effects. We discuss two estimators that fit into our framework. Consider the effect of a treatment on some outcome variable of interest. Let d_i denote the dummy variable such that $d_i = 1$ when treatment is given to the i th individual, and $d_i = 0$ otherwise. Let y_{0i} and y_{1i} denote the potential outcomes when $d_i = 0$ and $d_i = 1$, respectively. We can then say that the treatment *causes* the outcome variable of the i th individual to increase by $y_{1i} - y_{0i}$. Thus, $y_{1i} - y_{0i}$ can be called the treatment effect for the i th individual. See, e.g., Rubin (1974). Individual treatment effect cannot be observed, though, because the econometrician only observes d_i and $y_i \equiv d_i y_{1i} + (1 - d_i) y_{0i}$. On the other hand, the *average treatment effect* $\beta \equiv E[y_{1i} - y_{0i}]$ can be identified and consistently estimated when d_i is assigned independent of (y_{0i}, y_{1i}) . Extending this idea, Hahn (1998) and Hirano, Imbens, and Ridder (2003) proposed estimators of the average treatment effect when the treatment d_i is assigned independent of (y_{0i}, y_{1i}) given the observed covariates x_i .

Hahn's (1998) estimator is

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^n \frac{\widehat{h}_1(x_i)}{\widehat{p}(x_i)} - \frac{\widehat{h}_2(x_i)}{1 - \widehat{p}(x_i)}$$

where $\widehat{h}_1(x_i)$, $\widehat{h}_2(x_i)$, and $\widehat{p}(x_i)$ are nonparametric estimators of $E[d_i y_i | x_i]$, $E[(1 - d_i) y_i | x_i]$, and $E[d_i | x_i]$.

Hahn (1998) proposed series estimation, a special case of sieve estimation, for $\widehat{h}_1(x_i)$, $\widehat{h}_2(x_i)$, and $\widehat{p}(x_i)$. Writing

$$\begin{aligned} E[d_i y_i - h_1(x_i) | x_i] &= 0 \\ E[(1 - d_i) y_i - h_2(x_i) | x_i] &= 0 \\ E[d_i - p(x_i) | x_i] &= 0 \\ E\left[\frac{h_1(x_i)}{p(x_i)} - \frac{h_2(x_i)}{1 - p(x_i)} - \beta\right] &= 0 \end{aligned}$$

we can easily recognize that this fits into our framework discussed in Section 3.

Hirano, Imbens, and Ridder’s (2003) estimator is

$$\widetilde{\beta} = \frac{1}{n} \sum_{i=1}^n \frac{d_i y_i}{\widehat{p}(x_i)} - \frac{(1 - d_i) y_i}{1 - \widehat{p}(x_i)}.$$

Writing

$$\begin{aligned} E[d_i - p(x_i) | x_i] &= 0 \\ E\left[\frac{d_i y_i}{p(x_i)} - \frac{(1 - d_i) y_i}{1 - p(x_i)} - \beta\right] &= 0 \end{aligned}$$

this also fits into our framework discussed in Section 3.

They also consider an estimator where the propensity score $p(x_i) = E[d_i | x_i]$ is estimated by nonparametric maximum likelihood estimation with a Logit specification, i.e., $p(x_i) = \exp(h(x_i)) / [1 + \exp(h(x_i))]$. This alternative estimator fits into our framework in Section 6.3. We note that our result there can in principle accommodate the case where the propensity score is specified as a Probit model, which has some minor theoretical significance because the proof in Hirano, Imbens and Ridder (2003) can address only a Logit specification.¹⁴ We also note that our result in Section 6.3 can in principle accommodate cases where the propensity score takes a more restrictive form. An example would be $\exp(h_1(x_1) + \dots + h_L(x_L)) / [1 + \exp(h_1(x_1) + \dots + h_L(x_L))]$, which imposes some additive separability restrictions on how the different x ’s affect the propensity score.

6.5 Example: 2-Step Estimation of Dynamic Models

There is a large recent literature on two step semiparametric estimation of single agent dynamic programming problems and dynamic games, including Hotz and Miller (1993 (HM1), 1994 (HM2)), Aguirregabiria and Mira (2002 (AM1), 2007(AM2)), Jofre-Bonet and Pesendorfer (2003), Bajari, Benkard, and Levin (2007 (BBL)), Pakes, Ostrovsky, and Berry (2007 (POB)), and Pesendorfer and Schmidt-Dengler (2008 (PS)).

Examining how our results relate to this large literature is challenging, especially since each of the above papers develops multiple models and multiple estimators. To simplify, we consider a stylized but general model that is representative of many of these models and estimators. Suppose there are a discrete set of agents i (firms, consumers, etc.) interacting in market j across discrete time t . The relevant variables are:

- y_{ijt} = finite dimensional vector of actions taken by firm i in market j at time t
- x_{ijt} = finite dimensional vector of “public” state variables for firm i in market j at time t
- u_{ijt} = finite dimensional vector of “private” state variables for firm i in market j at time t
- v_{ijt} = finite dimensional vector of shocks to the state evolution of firm i in market j at time t

In each period, agents simultaneous choose y_{ijt} given their respective information sets. Agent i ’s information set at t includes x_{ijt} , u_{ijt} , and all past values of these variables. It does not include u_{jt}^{-i} , i.e. the private state variables of other firms, or v_{ijt} .¹⁵

¹⁴Given the flexibility of $h(x_i)$, it is not clear from a practical perspective why one would prefer a Probit over a Logit specification.

¹⁵In our example, we restrict v_{ijt} from entering information sets at t to keep things simple. Some examples in the literature do allow it to enter, and our conclusions should also apply there.

We assume that u_{ijt} and v_{ijt} are independent of each other, that u_{ijt} is iid across i , j , and t , and that v_{ijt} is iid across j and t . The econometrician is assumed to observe (y_{ijt}, x_{ijt}) , but not (u_{ijt}, v_{ijt}) . Asymptotics will be considered as the number of markets (J) goes to infinity, holding the number of agents per market (N) and the number of periods each market is observed (T) fixed. For simplicity of notation, we consider a stationary, infinite horizon environment - the model can trivially be extended to a finite horizon case by including the variable t in x_{ijt} .

The primitives of the model are given by the following two structural equations

$$\begin{aligned} \text{Single Period Payoffs: } & \Pi(x_{ijt}, y_{ijt}, y_{jt}^{-i}, u_{ijt}; \beta_1) \\ \text{State Evolution Equation: } & x_{ijt+1} = f(x_{ijt}, y_{ijt}, y_{jt}^{-i}, v_{ijt}; \beta_2) \end{aligned} \quad (12)$$

The second equation describes how next period's state x_{ijt+1} evolves depending on the current state x_{ijt} and current actions y_{ijt} . Since these are both structural equations, they are typically parametrically specified - i.e. Π and f are assumed known up to the finite dimensional parameter vectors β_1 and β_2 . Note that given this formulation, we can without loss of generality assume that the individual elements of the vector u_{ijt} (and v_{ijt}) are distributed independently with marginal distributions $U(0, 1)$.¹⁶ Hence, β_1 and β_2 contain all the structural parameters of the model.

The equilibrium of this dynamic game can be characterized by the following "reduced form" distribution of equilibrium actions y_{ijt} conditional on the state x_{ijt} :

$$\text{Equilibrium Action Distribution: } p(y_{ijt}|x_{ijt})$$

Note that the randomness in y_{ijt} given x_{ijt} is induced by variation in the underlying "private" state variable u_{ijt} .

In the first step of these estimation procedures, the econometrician estimates two parts of the model:

1. The state evolution equation $x_{ijt+1} = f(x_{ijt}, y_{ijt}, y_{jt}^{-i}, v_{ijt}; \beta_2)$ is used to estimate the parameter vector $\hat{\beta}_2$. Since f is a primitive of the model, this is often done using parametric methods.¹⁷
2. The equilibrium action distribution $\hat{p}(y_{ijt}|x_{ijt})$ is non-parametrically estimated.

We want to stress that in all these methods, it is very important to estimate $\hat{p}(y_{ijt}|x_{ijt})$ non-parametrically. The equilibrium action distribution $p(y_{ijt}|x_{ijt})$ depends in a very complicated way on the underlying primitives $\Pi(x_{ijt}, y_{ijt}, y_{jt}^{-i}, u_{ijt}; \beta_1)$ and $f(x_{ijt}, y_{ijt}, y_{jt}^{-i}, v_{ijt}; \beta_2)$.¹⁸ If one places arbitrary parametric structure on $p(y_{ijt}|x_{ijt})$, this parametric structure is likely to contradict these underlying primitives, making the resulting estimates inconsistent.

In the second step of these estimation procedures, the estimates $\hat{\beta}_2$ and \hat{p} are used to construct an estimating equation that can often be represented in the following form:

$$E \left[m \left(y_{ijt}, x_{ijt}, \beta_1, \hat{\beta}_2, \hat{p}(\cdot) \right) \right] = 0 \quad (13)$$

where m is a known function. This moment condition is used to estimate the parameter vector $\hat{\beta}_1$. The different papers in the literature (and the different models in each paper) vary considerably in how they construct (13)¹⁹.

¹⁶Arbitrary marginal distributions and correlations can be generated using the functions Π and f , and the parameters β_1 and β_2 (note that we have not restricted the dimension of u_{ijt} or v_{ijt})

¹⁷Note that potential endogeneity problems in this equation have already been assumed away with the assumptions that v_{ijt} is iid, independent of u_{ijt} , and not observed by the agents before making their decisions y_{ijt} .

¹⁸Assessing the precise way in which $p(y_{ijt}|s_{ijt})$ relates to $\Pi(s_{ijt}, y_{ijt}, y_{jt}^{-i}, u_{ijt}; \beta_1)$ and $f(s_{ijt}, y_{ijt}, y_{jt}^{-i}, v_{ijt}; \beta_2)$ would require solving the dynamic programming problem, which is exactly what these methods are trying to avoid (for computational reasons).

¹⁹Some of the estimators in the literature cannot be represented in this form. For example, one of the estimators in BBL, the inequality based estimator, cannot be written in this form. Nor can the iterated version of the pseudo-maximum likelihood estimators of AM1 and AM2.

In some cases, m is specified directly as a moment condition (e.g. HM1, BBL, POB, PS); in others it can be derived from the derivative of a pseudo-likelihood (e.g. AM1, AM2, POB). However, in all cases, the intuition is the same: Once $\widehat{\beta}_2$ and \widehat{p} have been estimated, the econometrician knows agent i 's perceived distribution over all possible paths of future states in the dynamic model (conditional on his action y_{ijt}). Hence, the expected discounted sum of future profits resulting from any choice y_{ijt} can be computed (or simulated²⁰) up to the parameter vector β_1 .²¹ (13) can then be derived from the condition that the observed choices y_{ijt} given x_{ijt} maximize this expected discounted sum of future profits. This condition can then be used to estimate $\widehat{\beta}_1$. Note that this moment condition depends on the entire $\widehat{p}(\cdot)$, not just $\widehat{p}(x_{ijt})$ at the current state x_{ijt} . This is because to calculate the expected discounted sum of future profits (which is implicitly part of m), one needs to consider the distribution of competitor (and own) actions at *all* possible future states.

We now consider the conditions under which this model fits into our framework and our numerical equivalence result applies. We go through the estimating equations one by one. Since $x_{ijt+1} = f(x_{ijt}, y_{ijt}, y_{jt}^{-i}, v_{ijt}; \beta_2)$ is estimated parametrically, it easily fits into our framework where part of the first step includes a parametric M-estimator. This can be done whether the equation is estimated using maximum likelihood (in which case, the M-estimator would involve a moment condition using the score), or using a finite set of moment conditions. This results in a consistent estimator $\widehat{\beta}_2$ of this set of structural parameters.

The second estimating equation, i.e. non-parametric estimation of $p(y_{ijt}|x_{ijt})$ can be treated in two ways, depending on whether the actions y_{ijt} take on a discrete set of possible values, or whether there is a continuum of possible actions.²² When y_{ijt} only takes on L discrete values, one could estimate the conditional probabilities via the following finite set of conditional moment conditions:

$$\begin{aligned} E[y_{1ijt} - h_{1*}(x_{ijt})|x_{ijt}] &= 0, \\ &\vdots \\ E[y_{Lijt} - h_{L*}(x_{ijt})|x_{ijt}] &= 0, \end{aligned}$$

where y_{lijt} is an indicator function representing whether action l was taken, and the $h(\cdot) = (h_1(\cdot), \dots, h_L(\cdot))$ functions belong to $\mathcal{H}_{l,J} = \{h_l : h_l(x_{ijt}) = p_{l,1}(x_{ijt})\theta_{(l),1} + \dots + p_{l,K_J}(x_{ijt})\theta_{(l),K_J}\}$ for $l = 1, \dots, L$. Essentially, this is like estimating L linear probability models. Note that the flexibility of these non-parametric functions is indexed by K_J , which is assumed to increase as $J \rightarrow \infty$. This model falls under the extension in Section 6.1.²³ Alternatively, our extension in Section 6.3 (to general first stage non-parametric sieve M-estimation) suggests that our equivalence result would also hold if one modelled $p(y_{ijt}|x_{ijt})$ in other (sufficiently flexible) ways, e.g. a multinomial sieve Logit or sieve Probit. This is common in applied work, e.g. Ryan (2006).

On the other hand, if there are a continuum of possible actions (i.e., if y_{ijt} is a continuous random variable), one could estimate the conditional density $p(y_{ijt}|x_{ijt})$ non-parametrically using either orthogonal series method including Fourier series and wavelets (see, e.g. Efromovich (1999)), or sieve MLE using Hermite polynomial series (Gallant and Nychka (1987)), spline series (Stone (1990)), exponential families (Barron and Sheu (1991)), mixture of normals (Genovese and Wasserman (2000)) and many other flexible sieves. Our equivalence result in Section 6.3 would again apply.

Lastly, note that the third estimating equation, $E\left[m\left(y_{ijt}, x_{ijt}, \beta_1, \widehat{\beta}_2, \widehat{p}(\cdot)\right)\right] = 0$, does fit within our framework. The question is more whether the second stage estimation equation can actually be written in this form.

²⁰Our equivalence results do not formally cover the case where there is simulation error in the moment (13), which would be the case if expected profits were simulated and there was non-negligible simulation error. With a strong enough rate condition on how the number of simulation draws increases in sample size, we suspect our equivalence results would go through, but we leave this for future work.

²¹Again, this is done in the various papers in many different ways. Sometimes additional assumptions are necessary to do this, e.g. monotonicity conditions, or restrictions on the dimensionality of u_{ijt} and v_{ijt} .

²²Our theoretical results allow for x_{ijt} to be either continuous or discrete or mixed. Note that when x_{ijt} and y_{ijt} can only take on a finite set of values, one is in a parametric world and no longer needs first stage nonparametrics.

²³Note that the Murphy and Topel (1985) formula, if appropriately computed, would address within-market correlation due to v_{ijt} possibly being correlated across firms in a given market/time period.

As noted above, most of the second-stage estimating equations in this literature can be written in this form, though there are some notable exceptions (e.g. the inequality version of the BBL estimator, and the iterated version of the AM1 and AM2 estimators).

In summary, our numerical equivalence result applies to many of the models and estimators in this literature, regardless whether the action space is discrete or continuous. This means that when one uses the method of sieves to estimate the non-parametric components of the model, semi-parametric standard errors of the structural parameters $\widehat{\beta}_1$ can be computed by simply treating the chosen sieves as parametric functions and applying the well-known parametric methodology of Section 2.

7 Concluding Remarks

In this paper, we established the numerical equivalence between two estimators of asymptotic variance for two-step semiparametric estimators when the first-step nonparametric estimation is implemented by the method of sieves. Because the method of sieves is equivalent to a parametric model in a given finite sample, it is useful to examine the properties of the “parametric” estimator of the asymptotic variance. We show that this “parametric” estimator is numerically equivalent to a consistent sieve estimator of the semiparametric asymptotic variance. This numerical equivalence is significant because it means that practitioners can simply implement the well-known parametric formulas of Newey (1984) or Murphy and Topel (1985) without the need to understand and apply results in the semiparametric literature.

We derived the numerical equivalence for two classes of semiparametric two-step estimators: the first class involves first-stage sieve nonparametric estimation based on conditional moment restrictions;²⁴ the second class involves first-stage sieve nonparametric estimation based on a maximum-likelihood like criterion.²⁵ For both classes, we for simplicity have assumed that the second stage estimation of finite dimensional structural parameter is based on an exactly identified unconditional moment model. All of these could be relaxed. One could extend the numerical equivalence results to more general semiparametric models, including the misspecified semiparametric models considered in Ai and Chen (2007) and Ichimura and Lee (2008). Nevertheless, we believe that the numerical equivalence results in our current paper already cover a very wide range of practical applications of two-step semiparametric estimation.

Lastly, note that our result is predicated on the assumption that the asymptotic variance of the semiparametric estimator is finite. Practitioners should be careful not to implement the procedure for models where the asymptotic variance is infinite, which happens if the finite dimensional parameter is unidentified or if the semiparametric information bound is zero, as was discussed in Chamberlain (1985) or Hahn (1994). In practice, the latter may be more important because two-step semiparametric estimation tends to be employed only when the finite dimensional parameter of interest is identified. It is not clear whether it would be easy to establish information bound in complicated structural models.

²⁴The first class of semiparametric estimators is a special case of Ai and Chen (2007).

²⁵The second class does not fit into Ai and Chen (2007). To our knowledge, this result is new to the literature.

Appendix

A A Useful Lemma

Our proofs of numerical equivalence are based on the following auxiliary result:

Lemma 1 *Suppose that \mathbb{A} and \mathbb{B} are $(d_1 + d_2) \times d_1$ and $(d_1 + d_2) \times d_2$ matrices such that $[\mathbb{A}, \mathbb{B}]$ is nonsingular. Also suppose that \mathbb{F} is a $(d_1 + d_2) \times (d_1 + d_2)$ symmetric positive semidefinite matrix. Then the upper-left $d_1 \times d_1$ block of the matrix*

$$[\mathbb{A}, \mathbb{B}]^{-1} \mathbb{F} \begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix}^{-1},$$

where \mathbb{A} and \mathbb{B} are $(d_1 + d_2) \times d_1$ and $(d_1 + d_2) \times d_2$ matrix and , can be computed by the following algorithm:

Step 1: For the j th column of \mathbb{A} , solve

$$\min_c (\mathbb{A}_j - \mathbb{B}c)' \Upsilon^{-1} (\mathbb{A}_j - \mathbb{B}c)$$

for some symmetric positive definite matrix Υ . Let c_j^* denote the solution, and let $c^* = [c_1^*, \dots, c_{d_1}^*]$.

Step2: Compute

$$[(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*)]^{-1} [(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} \mathbb{F} \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*)] [(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*)]^{-1}$$

Proof. The first step is a least squares problem, and the solution is given by

$$c_j^* = (\mathbb{B}' \Upsilon^{-1} \mathbb{B})^{-1} \mathbb{B}' \Upsilon^{-1} \mathbb{A}_j$$

Now note that $[\mathbb{A} - \mathbb{B}c^*, \mathbb{B}]$ is such that $\mathbb{B}' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) = 0$ by construction, which implies that

$$\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} [\mathbb{A} - \mathbb{B}c^*, \mathbb{B}] = \begin{bmatrix} \mathbb{A}' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) & \mathbb{A}' \Upsilon^{-1} \mathbb{B} \\ \mathbb{B}' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) & \mathbb{B}' \Upsilon^{-1} \mathbb{B} \end{bmatrix} = \begin{bmatrix} (\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) & (c^*)' \mathbb{B}' \Upsilon^{-1} \mathbb{B} \\ 0 & \mathbb{B}' \Upsilon^{-1} \mathbb{B} \end{bmatrix}$$

and

$$\begin{aligned} & \left(\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} [\mathbb{A} - \mathbb{B}c^*, \mathbb{B}] \right)^{-1} \\ &= \begin{bmatrix} ((\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*))^{-1} & -((\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*))^{-1} (c^*)' \\ 0 & (\mathbb{B}' \Upsilon^{-1} \mathbb{B})^{-1} \end{bmatrix} \end{aligned} \quad (14)$$

Now, we have

$$\begin{aligned} & [\mathbb{A}, \mathbb{B}]^{-1} \mathbb{F} \begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix}^{-1} \\ &= (\Upsilon^{-1} [\mathbb{A}, \mathbb{B}])^{-1} (\Upsilon^{-1} \mathbb{F} \Upsilon^{-1}) \left(\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} \right)^{-1} \\ &= \left(\begin{bmatrix} (\mathbb{A} - \mathbb{B}c^*)' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} [\mathbb{A}, \mathbb{B}] \right)^{-1} \left(\begin{bmatrix} (\mathbb{A} - \mathbb{B}c^*)' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} \mathbb{F} \Upsilon^{-1} [\mathbb{A} - \mathbb{B}c^*, \mathbb{B}] \right) \left(\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} [\mathbb{A} - \mathbb{B}c^*, \mathbb{B}] \right)^{-1} \end{aligned} \quad (15)$$

Using (14), it can be shown that the upper left block of (15) is equal to

$$((\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*))^{-1} [(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} \mathbb{F} \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*)] ((\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*))^{-1},$$

which proves the validity of the algorithm. ■

B Proof of Numerical Equivalence Result in Section 3

We now prove the first main numerical equivalence result stated in Section 3. We assume that the practitioner adopts the parametric specification $h_l(x_{l,i}, \theta_{(l)}) = p_l^{K_l}(x_{l,i})' \theta_{(l)}$, for $l = 1, \dots, L$, and hence, $\widehat{h}_l(x_{l,i}) = p_l^{K_l}(x_{l,i})' \widehat{\theta}_{(l)}$, where $K_l = K_{l,n}$ is a function of n although it is perceived to be fixed from the practitioner's view. The practitioner's estimator of asymptotic variance is (7) with

$$\begin{aligned} \frac{\partial g(z_i, \widehat{\beta}, \widehat{\theta})}{\partial(\beta', \theta')} &= \begin{bmatrix} 0 & -p_1^{K_1}(x_{1,i}) \left(p_1^{K_1}(x_{1,i}) \right)' & & \\ & & \ddots & \\ 0 & & & -p_L^{K_L}(x_{L,i}) \left(p_L^{K_L}(x_{L,i}) \right)' \\ \frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta'} & \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_1} p_1^{K_1}(x_{1,i})' & \dots & \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_L} p_L^{K_L}(x_{L,i})' \end{bmatrix} \\ &= \begin{bmatrix} 0 & -P_i P_i' \\ q_i' & Q_i' \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} g(z_i, \widehat{\beta}, \widehat{\theta}) &= \begin{bmatrix} p_1^{K_1}(x_{1,i}) \left(y_{1i} - h_1(x_{1,i}, \widehat{\theta}_{(1)}) \right) \\ \vdots \\ p_L^{K_L}(x_{L,i}) \left(y_{Li} - h_L(x_{Li}, \widehat{\theta}_{(L)}) \right) \\ m(z_i, \widehat{\beta}, h_1(x_{1,i}, \widehat{\theta}_{(1)}), \dots, h_L(x_{Li}, \widehat{\theta}_{(L)})) \end{bmatrix} \\ &= \begin{bmatrix} P_i & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} y_i - h_i \\ m_i \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} P_i &= \begin{bmatrix} p_1^{K_1}(x_{1,i}) & & 0 \\ & \ddots & \\ 0 & & p_L^{K_L}(x_{L,i}) \end{bmatrix} \\ q_i' &= \frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta'} \\ Q_i' &= \begin{bmatrix} \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_1} p_1^{K_1}(x_{1,i})' & \dots & \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_L} p_L^{K_L}(x_{L,i})' \end{bmatrix} \\ y_i - h_i &= \begin{bmatrix} y_{1i} - h_1(x_{1i}, \widehat{\theta}_{(1)}) \\ \vdots \\ y_{Li} - h_L(x_{Li}, \widehat{\theta}_{(L)}) \end{bmatrix} \\ m_i &= m(z_i, \widehat{\beta}, h_1(x_{1i}, \widehat{\theta}_{(1)}), \dots, h_L(x_{Li}, \widehat{\theta}_{(L)})). \end{aligned}$$

We now apply Lemma 1 to characterize the upper-left block of the estimated variance matrix. For this purpose, we let

$$\begin{aligned} \mathbb{A} &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} 0 \\ q_i' \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{q}' \end{bmatrix} \\ \mathbb{B} &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} -P_i P_i' \\ Q_i' \end{bmatrix} = \begin{bmatrix} -\frac{1}{n} \sum_{i=1}^n P_i P_i' \\ \bar{Q}' \end{bmatrix} \\ \mathbb{F} &= \frac{1}{n} \sum_{i=1}^n g(z_i, \widehat{\beta}, \widehat{\theta}) g(z_i, \widehat{\beta}, \widehat{\theta})' \end{aligned}$$

and

$$\Upsilon^{-1} = \begin{bmatrix} \left(\frac{1}{n} \sum_{i=1}^n P_i P_i'\right)^{-1} & 0 \\ 0 & I_d \end{bmatrix}$$

In the minimization problem of the first step, we see that the objective function is

$$(\mathbb{A}_j - \mathbb{B}c)' \Upsilon^{-1} (\mathbb{A}_j - \mathbb{B}c) = c' \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) c + \bar{q}'_j \bar{q}_j - 2\bar{q}'_j \bar{Q}' c + c' \bar{Q} \bar{Q}' c \quad (16)$$

Therefore, we can see that $c_j^* = \left(\left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) + \bar{Q} \bar{Q}' \right)^{-1} \bar{Q} \bar{q}_j$ or

$$c^* = \left(\left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) + \bar{Q} \bar{Q}' \right)^{-1} \bar{Q} \bar{q}'$$

Also, we have

$$(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) = (c^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) c^* + (\bar{q} - \bar{Q}' c^*)' (\bar{q} - \bar{Q}' c^*) \equiv \hat{\Lambda}_p$$

and

$$\begin{aligned} (\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} g(z_i, \beta, \theta) &= [(c^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) \quad \bar{q}' - (c^*)' \bar{Q}] \begin{bmatrix} \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right)^{-1} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} P_i & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} y_i - h_i \\ m_i \end{bmatrix} \\ &= [(c^*)' P_i \quad \bar{q}' - (c^*)' \bar{Q}] \begin{bmatrix} y_i - h_i \\ m_i \end{bmatrix} \\ &= [(c^*)' P_i (y_i - h_i) \quad \bar{q}' m_i - (c^*)' \bar{Q} m_i] \end{aligned}$$

and

$$\begin{aligned} (\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} \mathbb{F} \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) &= \frac{1}{n} \sum_{i=1}^n \left(\left(\hat{\mathbb{A}} - \hat{\mathbb{B}}c^* \right)' \mathbb{F} g_i \right) \left(\left(\hat{\mathbb{A}} - \hat{\mathbb{B}}c^* \right)' \mathbb{F} g_i \right)' \\ &= (\hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i (y_i - h_i)' (y_i - h_i) P_i' \right) \hat{c}^* + (\bar{q} - \bar{Q}' \hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n m_i m_i' \right) (\bar{q} - \bar{Q}' \hat{c}^*) \\ &\equiv \hat{\Omega}_p \end{aligned}$$

The practitioner's estimator \hat{V}_p for the asymptotic variance of $\hat{\beta}$ is then equal to

$$\hat{V}_p = \hat{\Lambda}_p^{-1} \hat{\Omega}_p (\hat{\Lambda}_p^{-1})'$$

Now, we note that Ai and Chen's first step minimization problem solves for c_j^* that minimizes

$$\frac{1}{n} \sum_{i=1}^n (P_i' c)' (P_i' c) + \left(\frac{1}{n} \sum_{i=1}^n (q_{ij} - Q_i' c) \right)^2 = c' \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) c + \bar{q}'_j \bar{q}_j - 2\bar{q}'_j \bar{Q}' c + c' \bar{Q} \bar{Q}' c \quad (17)$$

We can see that the same \hat{c}^* as above solves the practitioner's problem (16). Ai and Chen's estimator then requires calculating

$$\begin{aligned} \hat{\Delta}_{\hat{w}^*}(z_i) &= \begin{bmatrix} P_i' \hat{c}^* \\ \bar{q} - \bar{Q}' \hat{c}^* \end{bmatrix} \\ \frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{w}^*}(z_i) \right)' \hat{\Delta}_{\hat{w}^*}(z_i) &= (\hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) \hat{c}^* + (\bar{q} - \bar{Q}' \hat{c}^*)' (\bar{q} - \bar{Q}' \hat{c}^*) \end{aligned}$$

and

$$\begin{aligned}\widehat{\Omega} &= \frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{w}^*}(z_i) \right)' \rho(z_i, \widehat{\alpha}) \rho(z_i, \widehat{\alpha})' \widehat{\Delta}_{\widehat{w}^*}(z_i) \\ &= (\widehat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i (y_i - h_i)' (y_i - h_i) P_i' \right) \widehat{c}^* + \left(\bar{q} - \bar{Q}' \widehat{c}^* \right)' \left(\frac{1}{n} \sum_{i=1}^n m_i m_i' \right) \left(\bar{q} - \bar{Q}' \widehat{c}^* \right)\end{aligned}$$

Note that

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{w}^*}(z_i) \right)' \widehat{\Delta}_{\widehat{w}^*}(z_i) &= \widehat{\Lambda}_p \\ \widehat{\Omega} &= \widehat{\Omega}_p\end{aligned}$$

It follows that the practitioner's estimator of the asymptotic variance is numerically equal to Ai and Chen's.

C Understanding Newey's (1994) Asymptotic Variance Formula

Newey's result We consider a simple model where the true unknown function h_* is scalar-valued and solves $E[y_i - h_*(x_i) | x_i] = 0$, and the true β_* solves $E[m(z_i, \beta_*, h_*(x_i))] = 0$.

Newey (1994) considers a method of moment estimator $\widehat{\beta}$ that solves

$$\frac{1}{n} \sum_{i=1}^n m(z_i, \widehat{\beta}, \widehat{h}) = 0,$$

where \widehat{h} is some nonparametric estimator of h_* . Newey (1994) shows that the asymptotic variance of $\sqrt{n}(\widehat{\beta} - \beta_*)$ is the asymptotic variance of

$$-(M_\beta)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \{m(z_i, \beta_*, h_*) + E[D(z) | x = x_i] (y_i - h_*(x_i))\} \right) \quad (18)$$

where $M_\beta = E \left[\frac{\partial m(z_i, \beta_*, h_*)}{\partial \beta'} \right]$ and $D(z) = \partial m(z, \beta_*, h(x)) / \partial h|_{h=h_*}$.

Then a consistent estimator for the semiparametric asymptotic variance is equal to

$$\left(\widehat{M}_\beta \right)^{-1} \frac{1}{n} \sum_{i=1}^n \left(m(z_i, \widehat{\beta}, \widehat{h}) + \widehat{E}[D(z) | x_i] (y_i - \widehat{h}(x_i)) \right)^2 \left(\widehat{M}_\beta' \right)^{-1} \quad (19)$$

where $\widehat{M}_\beta = \frac{1}{n} \sum_{i=1}^n \frac{\partial m(z_i, \widehat{\beta}, \widehat{h})}{\partial \beta'}$ and $\widehat{E}[D(z) | x_i]$ is some nonparametric estimator of $E[D(z) | x = x_i]$. (For notational simplicity, we assume that m is scalar-valued.)

In order to prove (18), it suffices to characterize the asymptotic distribution of $\frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, \widehat{h})$. (This is because we have

$$\sqrt{n}(\widehat{\beta} - \beta_*) = -(M_\beta)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, \widehat{h}) \right) + o_p(1)$$

under regularity conditions.)

Newey (1994) basically writes

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, \widehat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{m(z_i, \beta_*, h_*) + a(z_i)\}$$

and devotes the rest of his paper characterizing the adjustment $a(z_i)$ to the influence function. We follow Newey's (1994) notation for convenience of readers. From Newey (p. 1360), we can see that, for $D(z, h) = D(z) h(v)$

with $D(z) = \partial m(z, h(v))/\partial h|_{h=h_*}$, we have his equation (4.1) satisfied. As is discussed on the same page, we now assume that $h_*(x) = E[y|x]$ for some y and x . Now we follow his equation (4.4), and see if we can find

$$E[D(z)\tilde{g}(x)] = E[\delta(x)\tilde{g}(x)] \text{ for all } \tilde{g}.$$

Obviously the answer is given by $\delta(x) = E[D(z)|x]$. Then according to Newey's (1994) Proposition 4, we can see that $a(z) = \delta(x)(y - E[y|x])$ or

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, \hat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (m(z_i, \beta_*, h_*) + E[D(z)|x=x_i](y_i - h_*(x_i)) + o_p(1)). \quad (20)$$

A Naïve practitioner's estimator Now we pose the following question. Let's assume that a practitioner fits a "flexible" but finite-dimensional parametric model $h(x, \theta)$ for $E[y|x]$. In other words, he will believe that $h_*(x) = E[y|x] = h(x, \theta_*)$. The practitioner pretends that his parametric model is a correct one. He will then assume that the population analog of his parametric strategy is $\theta_* = \operatorname{argmin}_{\theta} E[(y - h(x, \theta))^2]$. We will further suppose that $h(x_i, \theta) = p^K(x_i)' \theta = p_1(x_i)\theta_1 + \dots + p_K(x_i)\theta_K$ where $p^K(x) = (p_1(x), \dots, p_K(x))'$, where K is finite and fixed.

We now argue that a consistent estimator that this practitioner will use is the outer product of

$$-\left(\frac{1}{n} \sum_{i=1}^n \frac{\partial m(z_i, \hat{\beta}, h(x_i, \hat{\theta}))}{\partial \beta'}\right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (m(z_i, \hat{\beta}, h(x_i, \hat{\theta})) + \hat{E}^*[D(z_i)|p^K(x_i)](y_i - h(x_i, \hat{\theta})))\right)$$

where

$$\begin{aligned} & \hat{E}^*[D(z_i)|p^K(x_i)](y_i - h(x_i, \hat{\theta})) \\ & \equiv p^K(x_i)'(P'P)^{-1} \left(\sum_{i=1}^n p^K(x_i) D(z_i)\right) \left(y_i - p^K(x_i)'(P'P)^{-1} \left(\sum_{i=1}^n p^K(x_i) y_i\right)\right) \end{aligned} \quad (21)$$

and $P = [p^K(x_1), \dots, p^K(x_n)]'$. Because the practitioner believes that $\theta_* = \operatorname{argmin}_{\theta} E[(y - h(x, \theta))^2]$, he would believe that the corresponding moment equation is

$$E\left[\frac{\partial h(x, \theta_*)}{\partial \theta}(y - h(x, \theta_*))\right] = 0$$

With this in mind, he will conclude that

$$\sqrt{n}(\hat{\theta} - \theta_*) = \left(E\left[\frac{\partial h(x, \theta_*)}{\partial \theta} \frac{\partial h(x, \theta_*)}{\partial \theta'}\right]\right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial h(x_i, \theta_*)}{\partial \theta}(y_i - h(x_i, \theta_*))\right) + o_p(1)$$

He will then proceed and conclude that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, h(x_i, \hat{\theta})) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, h(x_i, \theta_*)) \\ &+ \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial m(z_i, \beta_*, h(x_i, \theta_*))}{\partial h} \frac{\partial h(x_i, \theta_*)}{\partial \theta'}\right) (\hat{\theta} - \theta_*) + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, h(x_i, \theta_*)) \\ &+ \left(\frac{1}{n} \sum_{i=1}^n D(z_i) \frac{\partial h(x_i, \theta_*)}{\partial \theta'}\right) \sqrt{n}(\hat{\theta} - \theta_*) + o_p(1) \end{aligned} \quad (22)$$

Now, in his mind, he will think that

$$\begin{aligned}
& \left(\frac{1}{n} \sum_{i=1}^n D(z_i) \frac{\partial h(x_i, \theta_*)}{\partial \theta'} \right) \sqrt{n} (\hat{\theta} - \theta_*) \\
&= E \left[D(z_i) \frac{\partial h(x_i, \theta_*)}{\partial \theta'} \right] \sqrt{n} (\hat{\theta} - \theta_*) + o_p(1) \\
&= E \left[D(z_i) \frac{\partial h(x_i, \theta_*)}{\partial \theta'} \right] \left(E \left[\frac{\partial h(x, \theta_*)}{\partial \theta} \frac{\partial h(x, \theta_*)}{\partial \theta'} \right] \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial h(x_i, \theta_*)}{\partial \theta} (y_i - h(x_i, \theta_*)) \right) + o_p(1) \quad (23)
\end{aligned}$$

We now see that, if we regress $D(z_i)$ on $\frac{\partial h(x, \theta_*)}{\partial \theta}$ in the population, the coefficient is equal to

$$\left(E \left[\frac{\partial h(x, \theta_*)}{\partial \theta} \frac{\partial h(x, \theta_*)}{\partial \theta'} \right] \right)^{-1} E \left[\frac{\partial h(x_i, \theta_*)}{\partial \theta} D(z_i) \right]$$

and the fitted value is equal to

$$\begin{aligned}
& \frac{\partial h(x_i, \theta_*)}{\partial \theta'} \left(E \left[\frac{\partial h(x, \theta_*)}{\partial \theta} \frac{\partial h(x, \theta_*)}{\partial \theta'} \right] \right)^{-1} E \left[\frac{\partial h(x_i, \theta_*)}{\partial \theta} D(z_i) \right] \\
&= E \left[D(z_i) \frac{\partial h(x_i, \theta_*)}{\partial \theta'} \right] \left(E \left[\frac{\partial h(x, \theta_*)}{\partial \theta} \frac{\partial h(x, \theta_*)}{\partial \theta'} \right] \right)^{-1} \frac{\partial h(x_i, \theta_*)}{\partial \theta} \quad (24)
\end{aligned}$$

So, let's write

$$E^* \left[D(z_i) \left| \frac{\partial h(x_i, \theta_*)}{\partial \theta} \right. \right] = E \left[D(z_i) \frac{\partial h(x_i, \theta_*)}{\partial \theta'} \right] \left(E \left[\frac{\partial h(x, \theta_*)}{\partial \theta} \frac{\partial h(x, \theta_*)}{\partial \theta'} \right] \right)^{-1} \frac{\partial h(x_i, \theta_*)}{\partial \theta} \quad (25)$$

where E^* denotes the best linear predictor. Combining (23) – (25), we can then see the practitioner's thought process would lead to the expression

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^n m(z_i, \beta_*, h(x_i, \hat{\theta})) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(m(z_i, \beta_*, h(x_i, \theta_*)) + E^* \left[D(z_i) \left| \frac{\partial h(x_i, \theta_*)}{\partial \theta} \right. \right] (y_i - h(x_i, \theta_*)) \right) + o_p(1) \quad (26)
\end{aligned}$$

We now compare (20) with (26). It is easy to see that, except for $E[D(z)|x=x_i]$ in (20) and $E^* \left[D(z_i) \left| \frac{\partial h(x_i, \theta_*)}{\partial \theta} \right. \right]$ in (26), the formulae that the practitioner uses for asymptotic variance calculation are identical. Obviously, we need to ask the question when $E^* \left[D(z_i) \left| \frac{\partial h(x_i, \theta_*)}{\partial \theta} \right. \right]$ can be interpreted to be an approximation of $E[D(z)|x=x_i]$. This is easy. Suppose that $h(x_i, \theta) = p^K(x_i)' \theta$. Then

$$\frac{\partial h(x_i, \theta_*)}{\partial \theta} = p^K(x_i)$$

so the best linear predictor $E^* \left[D(z_i) \left| \frac{\partial h(x_i, \theta_*)}{\partial \theta} \right. \right]$ is essentially the least squares operation on $p^K(x_i)$, which can be interpreted to be an approximation to $E[D(z)|x=x_i]$ as long as K is large enough.

A consistent estimator for the “parametric” asymptotic variance is equal to

$$\left(\widehat{M}_\beta \right)^{-1} \frac{1}{n} \sum_{i=1}^n \left(m(z_i, \hat{\beta}, h(x_i, \hat{\theta})) + \widehat{E}^* [D(z_i) | p^K(x_i)] (y_i - h(x_i, \hat{\theta})) \right)^2 \left(\widehat{M}_\beta' \right)^{-1}. \quad (27)$$

Numerical equivalence when \hat{h} is a sieve estimator When will Newey's estimator (19) of the semiparametric asymptotic variance (18) be numerically identical to the practitioner's parametric variance estimator (27)? If we are to use a sieve estimator with basis $p^K(x_i)$ to compute $\hat{h}(x_i) = \hat{E}[y|x = x_i]$ and $\hat{E}[D(z)|x_i]$ in Newey's (19), it can be easily seen that

$$\begin{aligned} & \hat{E}[D(z)|x_i] \left(y_i - \hat{E}[y|x = x_i] \right) \\ &= p^K(x_i)' (P'P)^{-1} \left(\sum_{i=1}^n p^K(x_i) D(z_i) \right) \left(y_i - p^K(x_i)' (P'P)^{-1} \left(\sum_{i=1}^n p^K(x_i) y_i \right) \right), \end{aligned} \quad (28)$$

which is numerically identical to (21). It follows that Newey's estimator (19) is numerically identical to (27) when a sieve least squares estimator is used for \hat{h} and $\hat{E}[D(z)|x_i]$. (In fact, Murphy and Topel's (1985) estimator is identical to (27).)

D Discussion of Ai and Chen (2007)

Ai and Chen's (2007) sieve estimator of the asymptotic variance may appear somewhat mysterious. It is in fact a sample counterpart of the population characterization of the asymptotic variance involving a minimization problem. In order to gain some intuition, we consider the following simple example model:

$$E[y_i - h_*(x_i)|x_i] = 0, \quad E[m(z_i, \beta_*, h_*(x_i))] = 0. \quad (29)$$

Ai and Chen' (2007) modified sieve minimum distance (SMD) estimator²⁶ for $\alpha_* = (\beta_*, h_*)$ boils down to

$$\left(\hat{\beta}, \hat{h} \right) = \underset{(\beta, h) \in \mathcal{B} \times \mathcal{H}_n}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i))^2 + \left\| \frac{1}{n} \sum_{i=1}^n m(z_i, \beta, h(x_i)) \right\|^2 \right\},$$

which amounts to estimating $h(x_i)$ by the method of sieves and then estimating β in the moment equation $E[m(z_i, \beta_*, h_*(x_i))] = 0$ plugging in the first step nonparametric estimator. In other words, it is exactly the same setup as that in Newey (1994). Ai and Chen (2007)'s asymptotic variance V for their $\hat{\beta}$ can be characterized by the following algorithm, where we assume that $\dim(\beta) = 1$ and scalar-valued h for notational simplicity:

1. Compute w^* to solve

$$\inf_w E \left[(w(x_i))^2 + \left(E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} w(x_i) \right] \right)^2 \right].$$

2. Calculate

$$\Delta_{w^*}(z_i) = \begin{bmatrix} w^*(x_i) \\ E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} w^*(x_i) \right] \end{bmatrix},$$

and

$$\rho(z_i, \alpha_*) = \begin{bmatrix} y_i - h_*(x_i) \\ m(z_i, \beta_*, h_*(x_i)) \end{bmatrix}.$$

3. Calculate

$$V = \left(E \left[\Delta_{w^*}(z_i)' \Delta_{w^*}(z_i) \right] \right)^{-1} \operatorname{Var} \left(\Delta_{w^*}(z_i)' \rho(z_i, \alpha_*) \right) \left(E \left[\Delta_{w^*}(z_i)' \Delta_{w^*}(z_i) \right] \right)^{-1}. \quad (30)$$

²⁶See their equation (5).

For this simple example model (29), it can be shown that the solutions in the above Steps 1 - 3 are

$$w^*(x_i) = \frac{E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} \right]}{1 + E \left[(E[D(z)|x=x_i])^2 \right]} E[D(z)|x=x_i],$$

$$\Delta_{w^*}(z_i) = \begin{bmatrix} \frac{E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} \right]}{1 + E \left[(E[D(z)|x=x_i])^2 \right]} E[D(z)|x=x_i] \\ \frac{E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} \right]}{1 + E \left[(E[D(z)|x=x_i])^2 \right]} \end{bmatrix},$$

$$\Delta_{w^*}(z_i)' \rho(z_i, \alpha_*) = \frac{E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} \right]}{1 + E \left[(E[D(z)|x=x_i])^2 \right]} (m(z_i, \beta_*, h_*(x_i)) + E[D(z)|x=x_i] (y_i - h_*(x_i))),$$

and

$$V = \frac{\text{Var} [m(z_i, \beta_*, h_*(x_i)) + E[D(z)|x=x_i] (y_i - h_*(x_i))]}{\left(E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} \right] \right)^2},$$

where $D(z) = \partial m(z, \beta_*, h(x)) / \partial h|_{h=h_*}$. In particular, we see that Ai and Chen's asymptotic variance V is identical to Newey's (1994) asymptotic variance (18) for this example model (29). We note that analytic characterization of $w^*(\cdot)$ hence population asymptotic variance V is not always easy for general semiparametric models considered in Ai and Chen (2007). Their sieve estimator of the asymptotic variance V simply uses a sample counterpart of the population minimization problem to bypass such a difficulty.

E Estimator of Asymptotic Variance of Sieve MLE (9)

We first recall the asymptotic variance of the sieve MLE $\hat{\beta}$, and then argue that the estimator \hat{V}_{smlc} of the asymptotic variance of $\hat{\beta}$ is consistent.

Below is an argument leading to the characterization of the asymptotic variance. We follow Chen and Shen's (1998) notation. For any $\alpha = (\beta, h) \in \mathcal{A} = \mathcal{B} \times \mathcal{H}$, let $\alpha(\alpha_*, \tau) \in \mathcal{A}$ is a path in τ connecting α_* and α such that $\alpha(\alpha_*, 0) = \alpha_*$ and $\alpha(\alpha_*, 1) = \alpha$. Let

$$\ell'_{\alpha_*}[z, \alpha - \alpha_*] = \lim_{\tau \rightarrow 0} \frac{\ell(z, \alpha(\alpha_*, \tau)) - \ell(z, \alpha_*)}{\tau} = \frac{d\ell(z, \alpha_*)}{d\beta'} (\beta - \beta_*) + \frac{d\ell(z, \alpha_*)}{dh} [h - h_*],$$

where when $h() = (h_1, \dots, h_L)$ we have

$$\frac{d\ell(z, \alpha_*)}{dh} [h - h_*] = \sum_{l=1}^L \frac{d\ell(z, \alpha_*)}{dh_l} [h_l - h_{*l}].$$

For any $\alpha, \bar{\alpha} \in \mathcal{A}$, denote $\ell'_{\alpha_*}[z, \alpha - \bar{\alpha}] = \ell'_{\alpha_*}[z, \alpha - \alpha_*] - \ell'_{\alpha_*}[z, \bar{\alpha} - \alpha_*]$, and define the metric $\|\cdot\|$ as

$$\|\alpha - \bar{\alpha}\| = \sqrt{E \left[(\ell'_{\alpha_*}[z, \alpha - \bar{\alpha}])^2 \right]}$$

which defines the Hilbert space on the closure of the linear span of $\mathcal{A} - \{\alpha_*\}$ with the inner product

$$\langle v, \bar{v} \rangle = E \left[\ell'_{\alpha_*}[z, v] \cdot \ell'_{\alpha_*}[z, \bar{v}] \right].$$

For each component β_j of β , let w_j^* denote the solution to

$$w_j^* = \arg \inf_{w \in \mathcal{H}} E \left[\left(\frac{d\ell(z, \alpha_*)}{d\beta_j} - \frac{d\ell(z, \alpha_*)}{dh} [w] \right)^2 \right] \quad \text{for } j = 1, \dots, d.$$

Denote

$$\Delta(z, \alpha_*) = \begin{bmatrix} \frac{d\ell(z, \alpha_*)}{d\beta_1} - \frac{d\ell(z, \alpha_*)}{dh} [w_1^*] \\ \vdots \\ \frac{d\ell(z, \alpha_*)}{d\beta_d} - \frac{d\ell(z, \alpha_*)}{dh} [w_d^*] \end{bmatrix},$$

and

$$\mathcal{I} \equiv E [\Delta(z_i, \alpha_*) \Delta(z_i, \alpha_*)'].$$

Consider the smooth functional $f(\alpha) = \lambda' \beta$ for some $\lambda \in R^d$ with $\lambda \neq 0$. Also let $\mathbf{w}^* = (w_1^*, \dots, w_d^*)$ and $v^* = (v_\beta^*, v_h^*)$ with

$$v_\beta^* = (E [\Delta(z, \alpha_*) \Delta(z, \alpha_*)'])^{-1} \lambda = (\mathcal{I})^{-1} \lambda, \quad v_h^* = -\mathbf{w}^* \times v_\beta^*.$$

We then have

$$\begin{aligned} & \langle (v_\beta^*, v_h^*), \alpha - \alpha_* \rangle \\ &= E \left[\left(\frac{d\ell(z, \alpha_*)}{d\beta'} v_\beta^* + \frac{d\ell(z, \alpha_*)}{dh} [v_h^*] \right) \left(\frac{d\ell(z, \alpha_*)}{d\beta'} (\beta - \beta_*) + \frac{d\ell(z, \alpha_*)}{dh} [h - h_*] \right) \right] \\ &= E \left[(\Delta(z, \alpha_*)' v_\beta^*) \left(\Delta(z, \alpha_*)' (\beta - \beta_*) + \frac{d\ell(z, \alpha_*)}{dh} [\mathbf{w}^* \times (\beta - \beta_*)] + \frac{d\ell(z, \alpha_*)}{dh} [h - h_*] \right) \right] \\ &= (v_\beta^*)' E [\Delta(z, \alpha_*) \Delta(z, \alpha_*)'] (\beta - \beta_*) \\ &= \lambda' (\beta - \beta_*) = f(\alpha) - f(\alpha_*) \end{aligned}$$

By Chen and Shen (1998, Theorem 2), we obtain that

$$\sqrt{n} \lambda' (\hat{\beta} - \beta_*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell'_{\alpha_*} [z_i, v^*] + o_p(1)$$

where

$$\ell'_{\alpha_*} [z_i, v^*] = \Delta(z_i, \alpha_*)' v_\beta^* = \Delta(z_i, \alpha_*)' (E [\Delta(z_i, \alpha_*) \Delta(z_i, \alpha_*)'])^{-1} \lambda.$$

In other words, we have

$$\sqrt{n} (\hat{\beta} - \beta_*) \rightarrow N(0, \mathcal{I}^{-1}), \quad \text{with } \mathcal{I} = E [\Delta(z_i, \alpha_*) \Delta(z_i, \alpha_*)'],$$

which provides an intuitive reason why the sieve estimator \hat{V}_{smle} given in (9) is a plausible estimator of \mathcal{I}^{-1} .

Below, we provide a rigorous proof for the consistency of (9). In the following we let $\|\cdot\|_s$ denote a metric (e.g., the supreme norm or the mean squared metric) on $\mathcal{A} = \Theta \times \mathcal{H}$. Denote $\mathcal{N}_0 = \{\alpha \in \mathcal{A} : \|\alpha - \alpha_*\|_s = o(1)\}$ and $\mathcal{W}_n = \{w \in \mathcal{H}_n : \|w\|_s \leq \text{const.} < \infty\}$. Also denote $g_j(z, \alpha, w) = \frac{d\ell(z, \alpha)}{d\beta_j} - \frac{d\ell(z, \alpha)}{dh} [w]$.

We impose the following assumptions:

Assumption A.1 (1) $\|v^*\|^2 = \lambda' \mathcal{I}^{-1} \lambda < \infty$; (2) There is a $v_n^* = (v_\beta^*, -\mathbf{w}_n^* v_\beta^*)$ with $\mathbf{w}_n^* = (w_{n1}^*, \dots, w_{nd}^*)$, $w_{nj}^* \in \mathcal{H}_n$ for all $j = 1, \dots, d$, such that $\|v_n^* - v^*\| = o(1)$.

Assumption A.2 For all $j = 1, \dots, d$, (1) $E \left[\sup_{\alpha \in \mathcal{N}_0, w \in \mathcal{W}_n} |g_j(z, \alpha, w)|^2 \right] \leq \text{const.} < \infty$; (2) there is a finite constant $\kappa > 0$ such that $|g_j(z, \alpha, w) - g_j(z, \alpha_*, w)| \leq U(z, w) \times \|\alpha - \alpha_*\|_s^\kappa$ for some $E \left[\sup_{w \in \mathcal{W}_n} |U(z, w)|^2 \right] \leq \text{const.} < \infty$.

Lemma 2 Let $\hat{\alpha} = (\hat{\beta}, \hat{h})$ be the sieve MLE such that $\|\hat{\alpha} - \alpha_0\|_s = o_P(1)$. Suppose that $\{z_i\}$ is i.i.d. and assumptions A.1-A.2 hold. If $K_n \rightarrow \infty$, $K_n/n \rightarrow 0$, then: $\hat{V}_{smle} = \mathcal{I}^{-1} + o_P(1)$.

Proof. Assumption A.2 implies that for all $j = 1, \dots, d$, $\left\{ \left(\frac{d\ell(z, \alpha)}{d\beta_j} - \frac{d\ell(z, \alpha)}{dh} [w] \right)^2 : \alpha \in \mathcal{N}_0, w \in \mathcal{W}_n \right\}$ is a Glivenko-Cantelli class. Thus, uniformly over $w \in \mathcal{H}_n$,

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \left(\frac{d\ell(z_i, \hat{\alpha})}{d\beta_j} - \frac{d\ell(z_i, \hat{\alpha})}{dh} [w] \right)^2 - E \left(\frac{d\ell(z_i, \alpha_*)}{d\beta_j} - \frac{d\ell(z_i, \alpha_*)}{dh} [w] \right)^2 \\
&= E_{z_i} \left[\left(\frac{d\ell(z_i, \hat{\alpha})}{d\beta_j} - \frac{d\ell(z_i, \hat{\alpha})}{dh} [w] \right)^2 \right] - E_{z_i} \left[\left(\frac{d\ell(z_i, \alpha_*)}{d\beta_j} - \frac{d\ell(z_i, \alpha_*)}{dh} [w] \right)^2 \right] + o_p(1) \\
&\leq \sqrt{E_{z_i} \left(\left\{ \frac{d\ell(z_i, \hat{\alpha})}{d\beta_j} - \frac{d\ell(z_i, \hat{\alpha})}{dh} [w] \right\} - \left\{ \frac{d\ell(z_i, \alpha_*)}{d\beta_j} - \frac{d\ell(z_i, \alpha_*)}{dh} [w] \right\} \right)^2} \\
&\times \sqrt{2E \left(\sup_{\alpha \in \mathcal{N}_0, w \in \mathcal{H}_n} \left| \frac{d\ell(z_i, \alpha)}{d\beta_j} - \frac{d\ell(z_i, \alpha)}{dh} [w] \right|^2 \right)} \\
&= o_P(1),
\end{aligned}$$

where the last equality also follows from assumption A.2. Here, E_{z_i} denotes the expectation taken only respect to z_i regarding $\hat{\alpha}$ as a nonstochastic constant. Thus,

$$\begin{aligned}
\min_{w \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left(\frac{d\ell(z_i, \hat{\alpha})}{d\beta_j} - \frac{d\ell(z_i, \hat{\alpha})}{dh} [w] \right)^2 &= \min_{w \in \mathcal{H}_n} E \left[\left(\frac{d\ell(z_i, \alpha_*)}{d\beta_j} - \frac{d\ell(z_i, \alpha_*)}{dh} [w] \right)^2 \right] + o_p(1) \\
&= \inf_{w \in \mathcal{H}} E \left[\left(\frac{d\ell(z_i, \alpha_*)}{d\beta_j} - \frac{d\ell(z_i, \alpha_*)}{dh} [w] \right)^2 \right] + o_p(1),
\end{aligned}$$

where the second equation follows from assumption A.1. The lemma now follows immediately. ■

F Proof for Section 6.2 on Restricted First Step

We first describe Ai and Chen's sieve estimator of the semiparametric asymptotic variance of $\hat{\beta}$ for this restricted case. For simplicity of notation, we will write

$$r(z_i, \alpha_*) = \begin{bmatrix} y_{1i} - h_*(x_{1i}) \\ \vdots \\ y_{Li} - h_*(x_{Li}) \end{bmatrix}$$

Assuming that $\mathcal{H}_n = \{h : h(x) = p_1(x)\theta_1 + \dots + p_{K_n}(x)\theta_{K_n}\}$, Ai and Chen's estimator \hat{V} of the asymptotic variance of $\hat{\beta}$ can be computed by the following algorithm:

1. Compute $\hat{\mathbf{w}}^* = (\hat{w}_1^*, \dots, \hat{w}_d^*)$ for $j = 1, \dots, d$ as

$$\hat{w}_j^* = \operatorname{argmin}_{w \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left\{ \left(\frac{\partial r(z_i, \hat{\alpha})}{\partial \beta_j} - \sum_{l=1}^L \frac{\partial r(z_i, \hat{\alpha})}{\partial h_l} w_j(x_{l,i}) \right)' \left(\frac{\partial r(z_i, \hat{\alpha})}{\partial \beta_j} - \sum_{l=1}^L \frac{\partial r(z_i, \hat{\alpha})}{\partial h_l} w_j(x_{l,i}) \right) + \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta_j} - \sum_{l=1}^L \frac{\partial m(z_i, \hat{\alpha})}{\partial h_l} w_j(x_{l,i}) \right) \right)^2 \right\}.$$

(We write $h_*(x_{li}) = h_{l*}(x_{li})$ for ease of accounting.)

2. Compute

$$\rho(z_i, \hat{\alpha}) = \begin{bmatrix} r(z_i, \hat{\alpha}) \\ m(z_i, \hat{\alpha}) \end{bmatrix},$$

$$\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) = \begin{bmatrix} \sum_{l=1}^L \left(\frac{\partial r(z_i, \widehat{\alpha})}{\partial \beta_1} - \sum_{l=1}^L \frac{\partial r(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_j^*(x_{l,i}) \right) & \cdots & \sum_{l=1}^L \left(\frac{\partial r(z_i, \widehat{\alpha})}{\partial \beta_d} - \sum_{l=1}^L \frac{\partial r(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_j^*(x_{l,i}) \right) \\ \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta_1} - \sum_{l=1}^L \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_j^*(x_{l,i}) \right) & \cdots & \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta_d} - \sum_{l=1}^L \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_l} \widehat{w}_j^*(x_{l,i}) \right) \end{bmatrix}$$

and

$$\widehat{\Omega} = \frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)' \rho(z_i, \widehat{\alpha}) \rho(z_i, \widehat{\alpha})' \widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i)$$

3. Compute

$$\widehat{V} = \left(\frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)' \widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)^{-1} \widehat{\Omega} \left(\frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)' \widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) \right)^{-1}.$$

Next, we assume that the practitioner adopts the parametric specification $h(x_{l,i}, \theta) = p^K(x_{l,i})' \theta$, where $p^K(x) = (p_1(x), \dots, p_K(x))'$, where $K = K_n$ is a function of n although it is perceived to be fixed for our fictitious practitioner. Note that the practitioner's estimator is identical to the modified SMD estimator. The practitioner's moment condition is then

$$g(z_i, \beta, \theta) = \begin{bmatrix} p^K(x_{1,i})(y_{1i} - h(x_{1i}, \theta)) + \cdots + p^K(x_{L,i})(y_{Li} - h(x_{Li}, \theta)) \\ m(z_i, \beta, h(x_{1i}, \theta), \dots, h(x_{Li}, \theta)) \end{bmatrix}$$

where $h(x_{li}, \theta) = p^K(x_{li})' \theta$. (For ease of accounting, we sometimes write $h(x_{li}, \theta) = h_l(x_{li}, \theta)$.) It follows that the practitioner's estimator of asymptotic variance is (7) with

$$\begin{aligned} \frac{\partial g(z_i, \widehat{\beta}, \widehat{\theta})}{\partial (\beta', \theta')} &= \begin{bmatrix} 0 & -p^K(x_{1,i})(p^K(x_{1,i}))' - \cdots - p^K(x_{L,i})(p^K(x_{L,i}))' \\ \frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta'} & \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_1} p_1^K(x_i)' + \cdots + \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_L} p_L^K(x_i)' \end{bmatrix} \\ &\equiv \begin{bmatrix} 0 & -P_i P_i' \\ q_i' & Q_i' \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} g(z_i, \widehat{\beta}, \widehat{\theta}) &= \begin{bmatrix} p_1^K(x_{1,i})(y_{1i} - h(x_{1i}, \widehat{\theta})) + \cdots + p_L^K(x_{L,i})(y_{Li} - h(x_{Li}, \widehat{\theta})) \\ m(z_i, \widehat{\beta}, h(x_{1i}, \widehat{\theta}), \dots, h(x_{Li}, \widehat{\theta})) \end{bmatrix} \\ &\equiv \begin{bmatrix} P_i & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} y_i - h_i \\ m_i \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} P_i &= [p^K(x_{1,i}) \quad \cdots \quad p^K(x_{L,i})] \\ q_i' &= \frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta'} \\ Q_i' &= \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_1} p_1^K(x_i)' + \cdots + \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_L} p_L^K(x_i)' \\ y_i - h_i &= \begin{bmatrix} y_{1i} - h(x_{1i}, \widehat{\theta}) \\ \vdots \\ y_{Li} - h(x_{Li}, \widehat{\theta}) \end{bmatrix} \\ m_i &= m(z_i, \widehat{\beta}, h(x_{1i}, \widehat{\theta}), \dots, h(x_{Li}, \widehat{\theta})) \end{aligned}$$

We now apply Lemma 1 to characterize the upper-left block of the estimated variance matrix. For this purpose, we let

$$\begin{aligned}\mathbb{A} &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} 0 \\ q'_i \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{q}' \end{bmatrix} \\ \mathbb{B} &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} -P_i P'_i \\ Q'_i \end{bmatrix} = \begin{bmatrix} -\frac{1}{n} \sum_{i=1}^n P_i P'_i \\ \bar{Q}' \end{bmatrix} \\ \mathbb{F} &= \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\beta}, \hat{\theta}) g(z_i, \hat{\beta}, \hat{\theta})'\end{aligned}$$

and

$$\Upsilon^{-1} = \begin{bmatrix} \left(\frac{1}{n} \sum_{i=1}^n P_i P'_i\right)^{-1} & 0 \\ 0 & I_d \end{bmatrix}$$

In the minimization problem of the first step, we see that the objective function is

$$(\mathbb{A}_j - \mathbb{B}c)' \Upsilon^{-1} (\mathbb{A}_j - \mathbb{B}c) = c' \left(\frac{1}{n} \sum_{i=1}^n P_i P'_i \right) c + \bar{q}'_j \bar{q}_j - 2\bar{q}_j \bar{Q}' c + c' \bar{Q} \bar{Q}' c \quad (31)$$

Therefore, we can see that $c_j^* = \left(\left(\frac{1}{n} \sum_{i=1}^n P_i P'_i \right) + \bar{Q} \bar{Q}' \right)^{-1} \bar{Q} \bar{q}_j$ or

$$c^* = \left(\left(\frac{1}{n} \sum_{i=1}^n P_i P'_i \right) + \bar{Q} \bar{Q}' \right)^{-1} \bar{Q} \bar{q}'$$

Also, we have

$$(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) = (c^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i P'_i \right) c^* + (\bar{q} - \bar{Q}' c^*)' (\bar{q} - \bar{Q}' c^*) \equiv \hat{\Lambda}_p$$

and

$$\begin{aligned}(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} g(z_i, \beta, \theta) &= [(c^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i P'_i \right) \quad \bar{q}' - (c^*)' \bar{Q}] \begin{bmatrix} \left(\frac{1}{n} \sum_{i=1}^n P_i P'_i\right)^{-1} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} P_i & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} y_i - h_i \\ m_i \end{bmatrix} \\ &= [(c^*)' P_i \quad \bar{q}' - (c^*)' \bar{Q}] \begin{bmatrix} y_i - h_i \\ m_i \end{bmatrix} \\ &= [(c^*)' P_i (y_i - h_i) \quad \bar{q}' m_i - (c^*)' \bar{Q} m_i]\end{aligned}$$

and

$$\begin{aligned}(\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} \mathbb{F} \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) &= \frac{1}{n} \sum_{i=1}^n \left(\left(\hat{\mathbb{A}} - \hat{\mathbb{B}}c^* \right)' \mathbb{F} g_i \right) \left(\left(\hat{\mathbb{A}} - \hat{\mathbb{B}}c^* \right)' \mathbb{F} g_i \right)' \\ &= (\hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i (y_i - h_i)' (y_i - h_i) P_i' \right) \hat{c}^* + (\bar{q} - \bar{Q}' \hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n m_i m_i' \right) (\bar{q} - \bar{Q}' \hat{c}^*) \\ &\equiv \hat{\Omega}_p\end{aligned}$$

The practitioner's parametric estimator \hat{V}_p for the parametric asymptotic variance of $\hat{\beta}$ is then equal to

$$\hat{V}_p = \hat{\Lambda}_p^{-1} \hat{\Omega}_p (\hat{\Lambda}_p^{-1})'$$

Finally, we note that Ai and Chen's first step minimization problem solves for c_j^* that minimizes

$$\frac{1}{n} \sum_{i=1}^n (P_i' c)' (P_i' c) + \left(\frac{1}{n} \sum_{i=1}^n (q_{ij} - Q_i' c) \right)^2 = c' \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) c + \bar{q}_j' \bar{q}_j - 2 \bar{q}_j' \bar{Q}' c + c' \bar{Q} \bar{Q}' c. \quad (32)$$

We can see that the same \hat{c}^* as above solves the practitioner's problem (16). Ai and Chen's estimator then requires calculating

$$\begin{aligned} \hat{\Delta}_{\hat{w}^*}(z_i) &= \begin{bmatrix} P_i' \hat{c}^* \\ \bar{q} - \bar{Q}' \hat{c}^* \end{bmatrix}, \\ \frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{w}^*}(z_i) \right)' \hat{\Delta}_{\hat{w}^*}(z_i) &= (\hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i P_i' \right) \hat{c}^* + (\bar{q} - \bar{Q}' \hat{c}^*)' (\bar{q} - \bar{Q}' \hat{c}^*), \end{aligned}$$

and

$$\begin{aligned} \hat{\Omega} &= \frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{w}^*}(z_i) \right)' \rho(z_i, \hat{\alpha}) \rho(z_i, \hat{\alpha})' \hat{\Delta}_{\hat{w}^*}(z_i) \\ &= (\hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n P_i (y_i - h_i)' (y_i - h_i) P_i' \right) \hat{c}^* + (\bar{q} - \bar{Q}' \hat{c}^*)' \left(\frac{1}{n} \sum_{i=1}^n m_i m_i' \right) (\bar{q} - \bar{Q}' \hat{c}^*). \end{aligned}$$

Note that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \left(\hat{\Delta}_{\hat{w}^*}(z_i) \right)' \hat{\Delta}_{\hat{w}^*}(z_i) &= \hat{\Lambda}_p, \\ \hat{\Omega} &= \hat{\Omega}_p. \end{aligned}$$

It follows that the practitioner's estimator of the parametric asymptotic variance is *numerically equal to* Ai and Chen's sieve estimator of the semi-parametric asymptotic variance.

G Proof for Section 6.3 on First Step Sieve M-Estimation

In order to simplify presentation we assume that β is a scalar (i.e., $\dim(\beta) = 1$) and h is a scalar function of x . Then, under standard regularity conditions, we show that $\hat{\beta}$ is \sqrt{n} -consistent and asymptotically normal, and solve its asymptotic variance analytically. Below we provide two ways to characterize the asymptotic variance of $\hat{\beta}$.

Explicit characterization of the influence function Asymptotic variance can be obtained by explicitly characterizing the influence function of

$$\frac{1}{n} \sum_{i=1}^n m(z_i, \beta_*, \hat{h}(x_i)).$$

Define the functional $f : \mathcal{H} \rightarrow \mathbb{R}$ as $f(h) = E[m(z_i, \beta_*, h(x_i))]$. Using Chen and Shen (1998), we then have

$$\begin{aligned} f'[\alpha - \alpha_*] &= E \left[\frac{\partial m(z_i, \beta_*, h_*(x_i))}{\partial h} (h(x_i) - h_*(x_i)) \right] \\ &= E \left[\frac{\partial \ell(z_i, h_*(x_i))}{\partial h} u^*(x_i) \frac{\partial \ell(z_i, h_*(x_i))}{\partial h} (h(x_i) - h_*(x_i)) \right] \end{aligned}$$

for

$$u^* = E \left[\left(\frac{\partial \ell(z_i, h_*(x_i))}{\partial h} \right)^2 \middle| x_i \right]^{-1} E \left[\frac{\partial m(z_i, \beta_*, h_*(x_i))}{\partial h} \middle| x_i \right].$$

We can write

$$f' [\alpha - \alpha_*] = \langle v^*, \alpha - \alpha_* \rangle$$

where

$$v^* = \mathcal{I}(x_i)^{-1} M_h(x_i),$$

and

$$\begin{aligned} \mathcal{I}(x_i) &= E \left[\left(\frac{\partial \ell(z_i, h_*(x_i))}{\partial h} \right)^2 \middle| x_i \right] = -E \left[\frac{\partial^2 \ell(z_i, h_*(x_i))}{\partial h^2} \middle| x_i \right], \\ M_h(x_i) &= E \left[\frac{\partial m(z_i, \beta_*, h_*(x_i))}{\partial h} \middle| x_i \right], \quad M_\beta = E \left[\frac{\partial m(z_i, \beta_*, h_*(x_i))}{\partial \beta} \right]. \end{aligned}$$

It follows that the influence function is

$$\frac{\partial \ell(z_i, h_*(x_i))}{\partial h} [v^*] = \frac{\partial \ell(z_i, h_*(x_i))}{\partial h} \mathcal{I}(x_i)^{-1} M_h(x_i)$$

It follows that, as long as stochastic equicontinuity is satisfied, $\sqrt{n}(\hat{\beta} - \beta_*) \rightarrow N(0, V)$, where

$$V = \frac{E \left[\left(m(z_i, \beta_*, h_*(x_i)) + \frac{\partial \ell(z_i, h_*(x_i))}{\partial h} \mathcal{I}(x_i)^{-1} M_h(x_i) \right)^2 \right]}{M_\beta^2}.$$

Ai and Chen (2007) style asymptotic variance characterization If we adopt the approach of Ai and Chen (2007), we have $\sqrt{n}(\hat{\beta} - \beta_*) \rightarrow N(0, v_\beta^* \Omega v_\beta^*)$, where

$$v_\beta^* = \left(E \left[\left(-\frac{\partial^2 \ell(z_i, h_*)}{\partial h \partial h} [\mathbf{w}^*, \mathbf{w}^*] \right) + \left(E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} [\mathbf{w}^*] \right] \right)^2 \right] \right)^{-1},$$

and

$$\Omega = \text{Var} \left(\frac{\partial \ell(z_i, h_*)}{\partial h} [\mathbf{w}^*] + \left(E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} [\mathbf{w}^*] \right] \right) m(z_i, \alpha_*) \right),$$

and \mathbf{w}^* solves

$$\inf_{w \in \mathcal{H}} E \left[\left(-\frac{\partial^2 \ell(z_i, h_*)}{\partial h \partial h} [w, w] \right) + \left(E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} [w] \right] \right)^2 \right]. \quad (33)$$

Equivalence of these two asymptotic variance characterizations For the simple case of scalar $h()$ function of x , the optimization problem (33) can be solved in closed form. Note that

$$\begin{aligned} & E \left[\left(-\frac{\partial^2 \ell(z_i, h_*)}{\partial h \partial h} [w, w] \right) \right] + \left(E \left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} [w] \right] \right)^2 \\ &= E \left[\mathcal{I}(x_i) w(x_i)^2 \right] + (M_\beta - E[M_h(x_i) w(x_i)])^2 \end{aligned}$$

has a solution equal to

$$w^*(x_i) = \left(M_\beta - \frac{M_\beta E \left[\frac{M_h(x)^2}{\mathcal{I}(x)} \right]}{1 + E \left[\frac{M_h(x)^2}{\mathcal{I}(x)} \right]} \right) \frac{M_h(x_i)}{\mathcal{I}(x_i)} = \frac{M_\beta}{1 + E \left[\frac{M_h(x)^2}{\mathcal{I}(x)} \right]} \frac{M_h(x_i)}{\mathcal{I}(x_i)} \equiv \frac{M_\beta}{\Xi} \frac{M_h(x_i)}{\mathcal{I}(x_i)}$$

so that

$$\begin{aligned} (v_\beta^*)^{-1} &= \frac{M_\beta^2}{\left(1 + E\left[\frac{M_h(x)^2}{\mathcal{I}(x)}\right]\right)^2} E\left[\frac{M_h(x)^2}{\mathcal{I}(x)}\right] + \left(M_\beta - \frac{M_\beta}{1 + E\left[\frac{M_h(x)^2}{\mathcal{I}(x)}\right]} E\left[\frac{M_h(x)^2}{\mathcal{I}(x)}\right]\right)^2 \\ &= \frac{M_\beta^2}{\left(1 + E\left[\frac{M_h(x)^2}{\mathcal{I}(x)}\right]\right)^2} E\left[\frac{M_h(x)^2}{\mathcal{I}(x)}\right] + \frac{M_\beta^2}{\left(1 + E\left[\frac{M_h(x)^2}{\mathcal{I}(x)}\right]\right)^2} = \frac{M_\beta^2}{\Xi} \end{aligned}$$

Note that

$$\frac{\partial \ell(z_i, h_*)}{\partial h} [\mathbf{w}^*] = \frac{\partial \ell(z_i, h_*)}{\partial h} \frac{M_\beta}{\Xi} \frac{M_h(x_i)}{\mathcal{I}(x_i)}$$

and

$$E\left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} [w]\right] = M_\beta \left(1 - \frac{1}{\Xi} E\left[\frac{M_h(x_i)^2}{\mathcal{I}(x_i)}\right]\right) = \frac{M_\beta}{\Xi}$$

Then,

$$\begin{aligned} &\frac{\partial \ell(z_i, h_*)}{\partial h} [\mathbf{w}^*] + \left(E\left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} [\mathbf{w}^*]\right]\right) m(z_i, \alpha_*) \\ &= \frac{M_\beta}{\Xi} \left(\frac{\partial \ell(z_i, h_*)}{\partial h} \frac{M_h(x_i)}{\mathcal{I}(x_i)} + m(z_i, \alpha_*)\right) \end{aligned}$$

and

$$\begin{aligned} \Omega &= \text{Var}\left(\frac{\partial \ell(z_i, h_*)}{\partial h} [\mathbf{w}^*] + \left(E\left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h} [\mathbf{w}^*]\right]\right) m(z_i, \alpha_*)\right) \\ &= \frac{M_\beta^2}{\Xi^2} \text{Var}\left(\frac{\partial \ell(z_i, h_*)}{\partial h} \frac{M_h(x_i)}{\mathcal{I}(x_i)} + m(z_i, \alpha_*)\right) \end{aligned}$$

from which we obtain

$$v_\beta^* \Omega v_\beta^* = \frac{\text{Var}\left(\frac{\partial \ell(z_i, h_*)}{\partial h} \frac{M_h(x_i)}{\mathcal{I}(x_i)} + m(z_i, \alpha_*)\right)}{M_\beta^2} = V.$$

Numerical equivalence We now apply Lemma 1 to characterize the upper-left block of the estimated variance matrix. For this purpose, we assume that the practitioner adopts the parametric specification $h(x, \theta) = p^K(x)' \theta$, with $p^K(x) = (p_1(x), \dots, p_K(x))'$, where $K = K_n$ is a function of n although it is perceived to be fixed for our fictitious practitioner. Then:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \frac{\partial g(z_i, \beta, \theta)}{\partial (\beta, \theta')} &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} 0 & -\frac{\partial^2 \ell(z_i, h(x_i, \theta))}{\partial h^2} p^K(x_i) p^K(x_i)' \\ \frac{\partial m(z_i, \beta, h(x_i, \theta))}{\partial \beta} & \frac{\partial m(z_i, \beta, h(x_i, \theta))}{\partial h} p^K(x_i)' \end{bmatrix} \\ &\equiv \begin{bmatrix} 0 & \bar{R} \\ \bar{q} & \bar{Q}' \end{bmatrix} \end{aligned}$$

and

$$g(z_i, \beta, \theta) = \begin{bmatrix} -\frac{\partial \ell(z_i, h(x_i, \theta))}{\partial h} p^K(x_i) \\ m(z_i, \beta, h(x_i, \theta)) \end{bmatrix}$$

Using the notation in Lemma 1, we let

$$\begin{aligned}\mathbb{A} &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} 0 \\ q'_i \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{q}' \end{bmatrix} \\ \mathbb{B} &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} R_i \\ Q'_i \end{bmatrix} = \begin{bmatrix} \bar{R} \\ \bar{Q}' \end{bmatrix} \\ \mathbb{F} &= \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\beta}, \hat{\theta}) g(z_i, \hat{\beta}, \hat{\theta})'\end{aligned}$$

and

$$\Upsilon^{-1} = \begin{bmatrix} \bar{R}^{-1} & 0 \\ 0 & I_d \end{bmatrix}$$

In the minimization problem of the first step in the lemma, we see that the objective function is

$$\begin{aligned}& (\mathbb{A} - \mathbb{B}c)'\ (\mathbb{A} - \mathbb{B}c) \\ &= c'\bar{R}c + \bar{q}^2 - 2\bar{q}\bar{Q}'c + c'\bar{Q}\bar{Q}'c \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \left(-\frac{\partial^2 \ell(z_i, \hat{\alpha})}{\partial h \partial h} (p^K(x_i)'c)^2 \right) + \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{\partial m(z_i, \hat{\alpha})}{\partial \beta} - \frac{\partial m(z_i, \hat{\alpha})}{\partial h} p^K(x_i)'c \right) \right)^2 \right\}\end{aligned}$$

which is identical to the minimization in our algorithm. We therefore obtain

$$\hat{v}_p^{-1} \equiv (\mathbb{A} - \mathbb{B}c^*)'\ \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) = (c^*)'\ \bar{R}c^* + (\bar{q} - \bar{Q}'c^*)' (\bar{q} - \bar{Q}'c^*) = (\hat{v}_\beta^*)^{-1}$$

We also have

$$\begin{aligned}& (\mathbb{A} - \mathbb{B}c^*)'\ \Upsilon^{-1} g(z_i, \beta, \theta) \\ &= \begin{bmatrix} -(c^*)' & \bar{q} - (c^*)'\bar{Q} \end{bmatrix} \begin{bmatrix} -\frac{\partial \ell(z_i, h(x_i, \theta))}{\partial h} p^K(x_i) \\ m(z_i, \beta, h(x_i, \theta)) \end{bmatrix} \\ &= \frac{\partial \ell(z_i, h(x_i, \theta))}{\partial h} (p^K(x_i)'c^*) + (\bar{q} - (c^*)'\bar{Q}) m(z_i, \alpha)\end{aligned}$$

and

$$\begin{aligned}\hat{\Omega}_p &\equiv (\mathbb{A} - \mathbb{B}c^*)'\ \Upsilon^{-1} \mathbb{F} \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*) \\ &= \frac{1}{n} \sum_{i=1}^n \left((\hat{\mathbb{A}} - \hat{\mathbb{B}}c^*)'\ \Upsilon^{-1} g_i \right) \left((\hat{\mathbb{A}} - \hat{\mathbb{B}}c^*)'\ \Upsilon^{-1} g_i \right)' \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial \ell(z_i, h(x_i, \theta))}{\partial h} (p^K(x_i)'c^*) + (\bar{q} - (c^*)'\bar{Q}) m(z_i, \beta, h(x_i, \theta)) \right)^2 \\ &\equiv \hat{\Omega}\end{aligned}$$

By Lemma 1, the practitioner's estimator \hat{V}_p for the asymptotic variance of $\hat{\beta}$ is then equal to

$$\hat{V}_p = \hat{v}_p \hat{\Omega}_p \hat{v}_p$$

Because $\hat{v}_p = \hat{v}_\beta^*$ and $\hat{\Omega}_p = \hat{\Omega}$, we get the desired conclusion that $\hat{V} = \hat{V}_p$.

References

- [1] Aguirregabiria, V. and P. Mira (2002): “Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models,” *Econometrica* 70, pp. 1519–1543.
- [2] Aguirregabiria, V. and P. Mira (2007): “Sequential Estimation of Dynamic Discrete Games,” *Econometrica* 75, pp. 1–53.
- [3] Ai, C. and X. Chen (2007): “Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables,” *Journal of Econometrics* 141, pp. 5 – 43.
- [4] Andrews, D. (1994) “Asymptotics for Semi-parametric Econometric Models via Stochastic Equicontinuity”, *Econometrica* 62, pp. 43-72.
- [5] Bajari, P., C.L. Benkard, and J. Levin (2007): “Estimating Dynamic Models of Imperfect Competition,” *Econometrica* 75, pp. 1331–1370.
- [6] Bajari, P., V. Chernozhukov, and H. Hong (2005): “Semiparametric Estimation of a Dynamic Game of Incomplete Information,” unpublished working paper, Duke University.
- [7] Bajari, P., V. Chernozhukov, H. Hong, D. Nekipelov (2008): “Nonparametric and Semiparametric Analysis of a Dynamic Game Model,” unpublished working paper.
- [8] Barron, A. and C. Sheu (1991): “Approximation of density functions by sequence of exponential families”, *The Annals of Statistics* 19, pp. 1347-1369.
- [9] Chamberlain, G. (1986): “Asymptotic efficiency in semi-parametric models with censoring,” *Journal of Econometrics* 32, pp. 189-218.
- [10] Chen, X. (2007): “Large Sample Sieve Estimation of Semi-nonparametric Models”, chapter 76 in *The Handbook of Econometrics*, Vol. 6B, eds. James J. Heckman and Edward E. Leamer, North-Holland.
- [11] Chen, X. and X. Shen (1998): “Sieve Extremum Estimates for Weakly Dependent Data,” *Econometrica* 66, pp. 289 – 314.
- [12] Chen, X., Y. Fan and V. Tsyrennikov (2006) “Efficient Estimation of Semiparametric Multivariate Copula Models”, *Journal of the American Statistical Association* 101, pp. 1228-1240.
- [13] Chen, X., O. Linton and I. van Keilegom (2003): “Estimation of Semiparametric Models when the Criterion Function is not Smooth”, *Econometrica* 71, pp. 1591-1608.
- [14] Collard-Wexler, A. (2006): “Demand Fluctuations and Plant Turnover in Ready Mix Concrete,” unpublished working paper, NYU Stern.
- [15] Dunne, T., Klimek, S., Roberts, M. and Y. Xu (2006): “Entry and Exit in Geographic Markets,” unpublished working paper, PSU.
- [16] Ellickson, P. and Misra, S. (2008): “Supermarket Pricing Strategies,” *Marketing Science* 27, pp. 811-828
- [17] Efromovich, S. (1999): *Nonparametric Curve Estimation*, Springer Series in Statistics: New York.
- [18] Gallant, A.R. and D. Nychka (1987): “Semi-non-parametric maximum likelihood estimation”, *Econometrica* 55, pp. 363-390.
- [19] Genovese, C. and L. Wasserman (2000): “Rates of Convergence for the Gaussian Mixture Sieve”, *The Annals of Statistics* 28, pp. 1105-1127.

- [20] Gonçalves, S. and H. White (2005): “Bootstrap Standard Error Estimation for Linear Regressions,” *Journal of the American Statistical Association* 100, pp. 970-979.
- [21] Hahn, J. (1994): “The Efficiency Bound of the Mixed Proportional Hazard Model,” *Review of Economic Studies* 61, pp. 607-629.
- [22] Hahn, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica* 66, pp.315-331.
- [23] Hirano, K., Imbens, G. W., and Ridder, G., (2003): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica* 71, pp.1161-1189.
- [24] Hotz, V.J. and R.A. Miller (1993): “Conditional Choice Probabilities and the Estimation of Dynamic Models,” *Review of Economic Studies* 60, pp. 497–529.
- [25] Hotz, V.J., Miller, R.A., Sanders, S., and J. Smith (1994): “A Simulation Estimator for Dynamic Models of Discrete Choice,” *Review of Economic Studies* 61, pp. 265-289
- [26] Ichimura, H. and S. Lee (2008): “Characterization of the Asymptotic Distribution of Semiparametric M-Estimators”, unpublished manuscript, University of Tokyo and University College London.
- [27] Imbens, G. and J. Wooldridge (2005): "Recent Developments in the Econometrics of Program Evaluation", working paper version, Harvard University and Michigan State University.
- [28] Jofre-Bonet, M. and M. Pesendorfer (2003): "Estimation of a Dynamic Auction Game," *Econometrica* 71, pp. 1443 – 1489.
- [29] Macieria, J. (2008): “Extending the Frontier: A Structural Model of Investment and Technological Competition in the Supercomputer Industry,” mimeo, Virginia Tech
- [30] Murphy, K. M. and R. H. Topel (1985): “Estimation and Inference in Two-Step Econometric Models,” *Journal of Business and Economic Statistics* 3, pp. 370 – 379.
- [31] Newey, W.K. (1984): “A Method of Moments Interpretation of Sequential Estimators,” *Economics Letters* 14, pp. 201 – 206.
- [32] Newey, W.K. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica* 62, pp. 1349 – 1382.
- [33] Newey, W.K. and D. F. McFadden (1994): “Large sample estimation and hypothesis testing” , in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.
- [34] Olley, G.S. and A. Pakes (1996): “The Dynamics of Productivity in the Telecommunications Equipment Industry The Dynamics of Productivity in the Telecommunications Equipment Industry,” *Econometrica* 64, pp. 1263–1297.
- [35] Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review* 25, pp. 221-247.
- [36] Pakes, A. and G.S. Olley (1995): “A Limit Theorem for a Smooth Class of Semiparametric Estimators,” *Journal of Econometrics* 65, pp. 295–332.
- [37] Pakes, A., Ostrovsky, M, and S. Berry (2007): “Simple Estimators for the Parameters of Discrete Dynamic Games, with Entry/Exit Examples,” *RAND Journal of Economics* 38, pp. 373-399.
- [38] Pesendorfer, M. and P. Schmidt-Dengler (2008): “Asymptotic Least Squares Estimators for Dynamic Games,” *Review of Economic Studies* 75, pp. 901–928.

- [39] Robinson, P. (1988) “Root-N-Consistent Semiparametric Regression”, *Econometrica*, 56, pp. 931-954.
- [40] Rubin, D. (1974): “Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies,” *Journal of Educational Psychology* 66, pp. 688-701.
- [41] Ryan, S. (2006): “The Costs of Environmental Regulation in a Regulated Industry,” unpublished working paper, MIT.
- [42] Ryan, S. and C. Tucker (2008): “Heterogeneity and the Dynamics of Technology Adoption,” unpublished working paper, MIT.
- [43] Shen, X. (1997) “On Methods of Sieves and Penalization”, *The Annals of Statistics* 25, pp. 2555-2591.
- [44] Snider, C. (2008): “Predatory Incentives and Predation Policy: The American Airlines Case,” unpublished working paper, Minnesota.
- [45] Stone, C.J. (1990): “Large-sample inference for log-spline models”, *The Annals of Statistics* 18, pp. 717-741.
- [46] Sweeting, A. (2007): “Dynamic Product Repositioning in Differentiated Product Industries: The Case of Format Switching in the Commercial Radio Industry,” unpublished working paper, Duke.
- [47] Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.