A PRACTICAL ASYMPTOTIC VARIANCE ESTIMATOR FOR TWO-STEP SEMIPARAMETRIC ESTIMATORS

Daniel Ackerberg, Xiaohong Chen, and Jinyong Hahn*

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 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 semiparametric procedures.

I. Introduction

MANY recently introduced empirical methodologies use two-step semiparametric estimation approaches. In the first step, certain functions are estimated nonparametrically. In the second step, structural and causal parameters are estimated parametrically, using the nonparametric estimates from the first stage as inputs. Such estimators have been used in both the treatment effect literature to estimate average treatment effects (Hahn, 1998; Hirano, Imbens, & Ridder, 2003) and in the labor and industrial organization (IO) literatures to estimate rich, often dynamic, structural models (Hotz and Miller, 1993; Hotz et al., 1994; Olley & Pakes, 1996; Aguirregabiria & Mira, 2002, 2007; Jofre-Bonet & Pesendorfer, 2003; Bajari, Benkard, & Levin, 2007; Pakes, Ostrovsky, & Berry, 2007; Pesendorfer & Schmidt-Dengler, 2008; Bajari, Hong, Krainer, & Nekipelov, 2008; Bajari, Chernozhukov, Hong, & Nekipelov, 2010). These two-step semiparametric estimators often have significant computational advantages over one-step estimators.

These methods often rely crucially on being nonparametric in the first step. For example, in the approach of Hotz and Miller (1993), the first step is to estimate reduced-form policy functions that arise from the equilibrium 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, for example, the number of terms in a series approximation or the flexibility of a sieve. But naive parametric specification of these reduced-form policy functions is likely to contradict the 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. This requires one to explicitly consider the problem's semiparametric nature when estimating the variances of the estimated finite-dimensional (structural) parameters.

A long line of theoretical literature derives expressions for semiparametric asymptotic variances of two-step estimators (Newey, 1994; Andrews, 1994; Newey & McFadden, 1994; Ai & Chen, 2007; Chen, Linton, & van Keilegom, 2003; Ichimura & Lee, 2010, to name a few). Some of this work also shows how to consistently estimate the asymptotic variances. While these theoretical results are useful, their implementation is typically not straightforward in practice. These limitations have often led applied researchers to use the bootstrap to estimate asymptotic variances (Ryan, 2006; Ellickson & Misra, 2008; Macieira, 2008), but this can be computationally demanding and may also be difficult to justify theoretically.[2]

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 semiparametrically, 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. But the asymptotic variances of the two estimators will be different because 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, perhaps 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 II; see Murphy & Topel, 1985, or Newey & McFadden, 1994). (In the online appendix, we explain how Newey's asymptotic variance formula and Ai and Chen's asymptotic variance formula are related to each other.) In other words,

Received for publication October 6, 2009. Revision accepted for publication October 18, 2010.

* Ackerberg: University of Michigan; Chen: Yale University; Hahn: UCLA.

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. An online appendix is available at http:www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00251.

[1] Imposing 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.

[2] 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 Gonçalves & White, 2005, for a discussion).

The Review of Economics and Statistics, May 2012, 94(2): 481–498

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 versus the true asymptotic semiparametric variance of the finite-dimensional structural parameters of interest. To the best of our knowledge, Newey (1994) was the first to recognize this equivalence[3] in a simple example involving one infinite-dimensional parameter, which is estimated by least squares using a series approximation in the first step.[4] We go one step further and generalize his insight to other classes of two-step semiparametric estimators, including models with multiple nonpametric 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. These equivalence results are useful for applied researchers, since they imply that one can obtain estimates of standard errors for the finite-dimensional structural parameters using well-known and simple formulas from the parametric literature.[5] We hope that this simplicity will promote the use of asymptotic semiparametric variance estimates and lessen the need for computationally burdensome bootstrapping.[6]

We start with a brief review of the standard two-step parametric approach in section II. Section III presents equivalence results for models where the first-stage sieve nonparametric estimation is based on conditional moment restrictions. Section IV considers the case where first-stage sieve nonparametric estimation is based on a maximum likelihood–like criterion. Section V considers various extensions of the result, for example, to situations where the second stage is overidentified and gives explicit examples of applications of our approach to the IO and labor literature discussed above. Section VI concludes.

[3] The "equivalence" throughout the paper refers to the equivalence between Newey (1994)/Ai and Chen (2007) variance estimators and Murphy and Topel (1985)/Newey and McFadden (1994) variance estimators. There are obviously other consistent estimators of the relevant asymptotic variances.

[4] Imbens and Wooldridge (2005) conjectured an equivalence in propensity score estimation.

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

[6] 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 addressed within the context of higher-order analysis, which we leave to future research.

II. 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 finite-dimensional parameter vector θ using a first-step M-estimator (for example, OLS, NLLS, MLE, method of moments). This estimate is then plugged into a second-step M-estimator, which is used to estimate another finite-dimensional parameter vector β. The question is whether and how the estimation error of the first-step M-estimator $\widehat{\theta}$ affects the asymptotic variance of the second-step M-estimator $\widehat{\beta}$. 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 $\widehat{\beta}$ are now so well understood that they can be found in standard textbooks such as Wooldridge (2002).

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

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

In the case where $\widehat{\theta}$ 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

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

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

In order to assess the asymptotic variance of $\sqrt{n}(\widehat{\beta}-\beta_*)$ that correctly reflects the estimation error of $\widehat{\theta}$, a researcher can consider the two-step estimator as a component of a one-step M-estimator,[7]

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

where

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

Note that equation (3) is equivalent to $\frac{1}{n}\sum_{i=1}^{n}\varphi(z_i,\widehat{\theta})=0$ and $\frac{1}{n}\sum_{i=1}^{n}\psi(z_i,\widehat{\beta},\widehat{\theta})=0$. Therefore, the $\widehat{\theta}$ and $\widehat{\beta}$ that solve

[7] This formulation assumes exact identification, $\dim(\varphi)=\dim(\theta)$ and $\dim(\psi)=\dim(\beta)$. We consider an overidentified situation in section VC.

equation (3) are numerically identical to $\widehat{\theta}$ and $\widehat{\beta}$ that solve equation (1) and (2). Letting $\alpha = (\beta', \theta')'$ and recognizing that $\widehat{\alpha} = (\widehat{\beta}', \widehat{\theta}')'$ is an M-estimator, we can then use standard arguments[8] to compute the asymptotic variance of $\sqrt{n}(\widehat{\alpha} - \alpha_*)$, that is, a consistent estimator of the asymptotic variance of $\sqrt{n}(\widehat{\alpha} - \alpha_*)$ is given by

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

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

III. 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 $\widehat{\beta}$ of some finite-dimensional parameter vector. We assume that the first-step nonparametric estimation is implemented by the method of sieves, such as 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 $\widehat{\beta}$ using the formula discussed in section II.

Researcher B 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 $\widehat{\beta}$ using a formula that correctly reflects $\widehat{\beta}$'s semiparametric nature. Because the two researchers are considering different asymptotic sequences, researcher A's asymptotic variance formula (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 from researcher B.[9] 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 consider two separate cases. In the first case, the second-stage moment equation depends on the nonparametric function only through its value evaluated at the particular observation. In the second case, the second-stage moment equation depends on the entire functional form of the nonparametric function.

A. *Dependence of Second Stage on the Nonparametric Function*

Consider a model given by the following moment restrictions:

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

The $h_1(\cdot), \ldots, h_L(\cdot)$ functions are the nonparametric components in the model. β is the finite-dimensional component of the model. Note that the conditioning variables $x_{1i}, \ldots, x_{Li}$ are allowed to differ from each other. We also allow the dimensions of $x_{1i}, \ldots, x_{Li}$ to differ. Unlike the second case discussed in this section, the second-stage moment equation m depends on the nonparametric components only through their values $h_1(x_{1i}), \ldots, h_L(x_{Li})$.

The practitioner nonparametrically estimates $h_{1*}(x_{1i}), \ldots, h_{L*}(x_{Li})$ with the estimators $\widehat{h}_1(x_{1i}), \ldots, \widehat{h}_L(x_{Li})$, and then estimates β_* with the $\widehat{\beta}$ that solves

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

Ai and Chen (2007) show that $\widehat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal under certain regularity conditions, and propose a consistent estimator $\widehat{V}$ of the asymptotic variance. (See Appendix A for details.) Ai and Chen assume that nonparametric estimation is implemented by the "sieve" approach, where each $h_l(x_l)$ is approximated by a polynomial function $p_{l,1}(x_l)\theta_{(l),1} + \cdots + p_{l,K_{l,n}}(x_l)\theta_{(l),K_{l,n}}$.

A Naive 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

[8] See Wooldridge (2002).

[9] Researcher 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 $\widehat{\beta}$ in the second step will be inconsistent in general because her first-step estimator will not converge to the true nonparametric object.

purely parametric perspective (researcher A). First, note that a parametric estimator based on the parametric specification $h_l(x_l) = p_{l,1}(x_l)\theta_{(l),1} + \cdots + p_{l,K_l}(x_l)\theta_{(l),K_l} = h_l(x_l, \theta_{(l)})$ (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

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

where $\alpha = (\beta', \theta')'$, $\theta = (\theta'_{(1)}, \ldots, \theta'_{(L)})'$, and for $l = 1, \ldots, L$, $h_l(x_{li}, \theta_{(l)}) = p_l^{K_l}(x_{l,i})'\theta_{(l)}$ with $p_l^{K_l}(x_{l,i}) = (p_{l,1}(x_{l,i}), \ldots, p_{l,K_l}(x_{l,i}))'$. 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,

$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(z_i, \widehat{\alpha})}{\partial \alpha'}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i, \widehat{\alpha}) g(z_i, \widehat{\alpha})'\right) \times \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(z_i, \widehat{\alpha})'}{\partial \alpha}\right)^{-1}, \quad (5)$$

and as in section II, an estimator $\widehat{V}_p$ of the parametric asymptotic variance of $\widehat{\beta}$ can be obtained from the upper-left corner of equation (5).

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 C 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 is 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.

B. *Extension: Dependence of Second Stage on Full Nonparametric Function*

Consider a model where

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

Note the important difference between this model and the previous one. 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.

Now assume that a practitioner takes a parametric perspective with $h_\theta(\cdot) = p_1(\cdot)\theta_1 + \cdots + p_K(\cdot)\theta_K$, 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 $\mathbf{m}(z_i, \alpha)' = [\mathbf{m}_1(z_i, \alpha), \ldots, \mathbf{m}_K(z_i, \alpha)]$, and for $k = 1, \ldots, K$,

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

Here, the pathwise derivatives are defined as

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

As before, the numerical equivalence goes through. (See Appendix C.) Thus we can conclude the upper-left ($\dim(\beta) \times \dim(\beta)$) block of the parametric variance estimator,

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

is numerically identical to a valid consistent estimator of the asymptotic semiparametric variance.

IV. 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 III, 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.[10] 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 (that is, $h(\cdot) = (h_1(\cdot), \ldots, h_L(\cdot))$, and each $h_l(\cdot)$ could depend on different argument x_l for $l = 1, \ldots, L$). 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 $\widehat{\beta}$ of β is a sample counterpart. In Appendix D, we propose a consistent estimator $\widehat{V}_{smle}$ of the asymptotic variance of $\widehat{\beta}$.[11]

We now discuss the practical implications. Consider a fictitious practitioner who assumes that h can 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 maximization problem. Would his standard error for $\widehat{\beta}$ be identical to ours?

As in the previous section, the practitioner would write

$$\begin{aligned} h_l(x_l) &= p_{l,1}(x_l)\theta_{(l),1} + \cdots + p_{l,K_l}(x_l)\theta_{(l),K_l} \\ &= p_l^{K_l}(x_l)'\theta_{(l)} \text{ for } \theta_{(l)} = (\theta_{(l),1}, \ldots, \theta_{(l),K_l})' \end{aligned}$$

with $p_l^{K_l}(x_l) = (p_{l,1}(x_l), \ldots, p_{l,K_l}(x_l))'$, where $K_l = K_{l,n}$ is a function of n, although it is perceived to be fixed for our fictitious practitioner. Denote $\theta = (\theta_{(1)}', \ldots, \theta_{(L)}')'$, 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:

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

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.

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 $\widehat{V}_p$ of the above matrix, which can be computed by the partitioned inverse formula.

It turns out 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. (See Appendix D.) We conclude that as long as the outer product is used for calculating of information, "parametric" inference for β is numerically identical to semiparametric inference.

V. 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.

A. *First Step with Restriction*

As another extension, we can consider a model where

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

where the dimensions of $x_{1i}, \ldots, x_{Li}$ are restricted to be identical, and for simplicity we assume $h_*(\cdot)$ is a scalar-valued function.

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. A natural estimator would minimize

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

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

[11] We provide a proof of the consistency of $\widehat{V}_{smle}$ along with regularity conditions in Appendix D 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 Chen, 2007, and Chen, Fan, & 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.

The practitioner's moment condition is then $E[g(z_i, \beta_*, \theta_*)] = 0$, where

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

where $h(x_{li}, \theta) = p^K(x_{l,i})'\theta$. It follows that the practitioner's estimator of asymptotic variance is equation (5).

Again, it turns out that the numerical equivalence continues to hold, and we obtain the practical conclusion that researchers wanting to do semiparametric inference need not explicitly consider the semiparametric nature of the problem in estimation. (See Appendix E for a proof.)

B. Nonparametric Sieve M-Estimation as a 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 $\widehat{\beta}$ 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:

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

where $\ell(z_i, h)$ is any criterion function and $h = (h_1, \ldots, h_L)$ is a vector of L unknown real-valued functions, each $h_l(\cdot)$ potentially depending on different arguments.[12] We propose a sieve estimator $\widehat{\beta}$, the characterization of the asymptotic variance V of $\sqrt{n}(\widehat{\beta} - \beta_*)$, as well as a consistent estimator $\widehat{V}$ of V in Appendix F.

As before, we note that the $\widehat{\beta}$ is numerically equivalent to the parametric estimator based on the parametric specification $h(\cdot) = p_1(\cdot)\theta_1 + \cdots + p_K(\cdot)\theta_K$, 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. As before, it can be shown that the sieve estimator $\widehat{V}$ of the asymptotic variance of $\widehat{\beta}$ is numerically identical to the well-known Murphy and Topel's (1985) formula. (See Appendix F.) We again obtain the practical conclusion that researchers wanting to do semiparametric inference need not explicitly consider the semiparametric nature of the problem in estimation.

[12] This problem does not fit into the framework of Ai and Chen (2007). To our knowledge, the result below is new to the literature.

C. Overidentified Second Step

So far, we have implicitly assumed that the second step is exactly identified: $\dim(m) = \dim(\beta)$. We now discuss the extension to the case where $\dim(m) > \dim(\beta)$. For simplicity of presentation, we will assume that the nonparametric component estimated in the first step is scalar valued, and is identified from the moment restriction $E[y - h_0(x)|x] = 0$. In the second step, we estimate β based on the moment restriction $E[m(z, \beta_0, h_0)] = 0$. Because h_0 is not known, we estimate β_0 by making the sample analog $\frac{1}{n}\sum_{i=1}^n m(z_i, \beta, \widehat{h})$ as close to 0 as possible. This is usually done by

$$\min_{\beta} \left(\frac{1}{n}\sum_{i=1}^n m(z_i, \beta, \widehat{h})\right)' \widehat{\Omega}^{-1} \left(\frac{1}{n}\sum_{i=1}^n m(z_i, \beta, \widehat{h})\right)$$

for some appropriate "weight matrix" $\widehat{\Omega}^{-1}$, that is, GMM. If we choose the probability limit Ω of $\widehat{\Omega}$ to be equal to the asymptotic variance matrix of $\frac{1}{n}\sum_{i=1}^n m(z_i, \beta_0, \widehat{h})$, then we can easily infer that the asymptotic variance of the resultant estimator is equal to $(M'\Omega^{-1}M)^{-1}$, where $M = E[\partial m(z_i, \beta_0, h_0)/\partial \beta']$ can be consistently estimated by $\widehat{M} = \frac{1}{n}\sum_{i=1}^n \partial m(z_i, \widetilde{\beta}, \widehat{h})/\partial \beta'$ given any arbitrary consistent estimator $\widetilde{\beta}$ of β_0.[13]

Therefore, for two-step estimation, the only thing that matters is consistent estimation of Ω because the rest is taken care of by the usual GMM formula. For this purpose, we write $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n m(z_i, \beta_0, \widehat{h})$ and understand Ω to be the asymptotic variance of $\widehat{\mu}$. If β_0 were known, we could estimate Ω using the Murphy-Topel formula applied to the "parameter" μ_0 in the moment restrictions

$$E[y - h_0(x)|x] = 0,$$
$$E[m(z_i, \beta_0, h_0) - \mu_0] = 0.$$

Thus, to derive a feasible estimator of Ω (and then β_0), we propose the following algorithm:

1. Estimate $\widehat{h}$ as before (by the sieve method as discussed at the end of section IIIA).
2. Using an arbitrary weight matrix W, minimize the sample moment $(\frac{1}{n}\sum_{i=1}^n m(z_i, \beta, \widehat{h}))'W(\frac{1}{n}\sum_{i=1}^n m(z_i, \beta, \widehat{h}))$ over β to obtain a preliminary estimator $\overline{\beta}$ of β_0.
3. Pretend that $\overline{\beta} = \beta_0$. "Estimate" $\widehat{\mu}$ by setting the sample moment $\frac{1}{n}\sum_{i=1}^n (m(z_i, \overline{\beta}, \widehat{h}) - \widehat{\mu})$ equal to 0 (this estimation problem is exactly identified—and trivial).
4. Again consider $\overline{\beta}$ to be fixed. Apply Murphy-Topel, (the naive practitioner's estimator of the asymptotic variance discussed in section IIIA), to the moment conditions corresponding to steps 1 and 3,

$$E[y - h_0(x)|x] = 0,$$
$$E[m(z_i, \overline{\beta}, h_0) - \mu_0] = 0,$$

[13] See Chen et al. (2003).

to obtain an estimate of the asymptotic variance matrix of $\widehat{\mu}$. Call this $\widehat{\Omega}$.

5. Solve the minimization problem,

$$\min_{\beta} \left(\frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta, \widehat{h})\right)' \widehat{\Omega}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta, \widehat{h})\right).$$

Call the solution $\widehat{\beta}$.

6. Compute

$$\left(\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(z_i, \widehat{\beta}, \widehat{h})}{\partial \beta'}\right)' \widehat{\Omega}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(z_i, \widehat{\beta}, \widehat{h})}{\partial \beta'}\right)\right)^{-1},$$

for a consistent estimator of the asymptotic variance of $\widehat{\beta}$. (Note that step 6 does not require applying Murphy-Topel a second time. This is because in this approach, the effect of the variance in $\widehat{h}$ on $\widehat{\beta}$ is summarized in the $\widehat{\Omega}$ obtained from using Murphy-Topel on $(\widehat{h}, \widehat{\mu})$ in step 4.)

This algorithm is the procedure of a naive practitioner, who equates the sieve estimation of $h_0(x)$ with parametric estimation. Yet at the same time, $\widehat{\Omega}$ is a consistent estimator of the asymptotic variance matrix of $\frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, \widehat{h})$, where $\widehat{h}$ is interpreted to be nonparametric, so the algorithm produces a correct semiparametric method of inference. As such, the result in this section can be understood to be a natural extension of the previous equivalence results.

D. *Example: Estimation of Average Treatment Effects*

The body of literature on estimation of average treatment effects is large. 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 ith 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 ith individual to increase by $y_{1i} - y_{0i}$. Thus, $y_{1i} - y_{0i}$ can be called the treatment effect for the ith individual (see, e.g., Rubin, 1974). Individual treatment effect cannot be observed, though, because the econometrician observes only d_i and $y_i \equiv d_i y_{1i} + (1 - d_i) y_{0i}$. 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 et al. (2003) proposed estimators of the average treatment effect when treatment d_i is assigned independent of (y_{0i}, y_{1i}) given the observed covariates x_i.

Hahn's (1998) estimator is

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

and Hirano et al.'s (2003) estimator is

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

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 $E[d_i | x_i]$. We can easily recognize that they fit into our framework discussed in section III.

Hirano et al. (2003) 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. This alternative estimator fits into our framework in section VB. 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 et al. (2003) can address only a logit specification.[14]

Note that implementation of Murphy-Topel would require writing down moments. As for Hahn's (1998) estimator, the moments are

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

and as for Hirano et al.'s (2003) estimator, they moments are

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

Replacing $h_1(x_i)$, $h_2(x_i)$, and $p(x_i)$ by parametric models and applying Murphy-Topel, we can obtain the asymptotic variance consistently.

E. *Example: Two-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) and Hotz et al. (1994), Aguirregabiria and Mira (2002, 2007), Jofre-Bonet and Pesendorfer (2003), Bajari et al. (2007), Pakes et al. (2007), Pesendorfer and Schmidt-Dengler (2008), and Bajari et al. (2010). The basic idea behind these estimators is that "reduced-form" policy functions describing optimal agent behavior can be nonparametrically estimated in a first stage.[15] These estimated policy functions can then

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

[15] Aguirregabiria (1999), Ryan (2006), Collard-Wexler (2006), Dunne, Klimek, Roberts, and Xu (2006), Sweeting (2007), Macieira (2008), Ellickson and Misra (2008), Snider (2008), and Ryan and Tucker (2008) are some examples of empirical applications of these methods.

be used as an input into a second-stage objective function that can be used to estimate a finite-dimensional structural parameter. Calculating this second-stage objective function typically does not require solving agents' dynamic programming problems, hence reducing computational burden relative to one-step estimation. In the following, we give a simple example.to illustrate how our results might be applied in some of these contexts.

Suppose a single agent makes a binary discrete choice $a_t \in \{0, 1\}$ in each period t. The state $x_t \in \mathbb{R}^J$ evolves according to distribution $F_x(x_{t+1}|x_t, a_t; \beta_F)$. Single-period utility is given by $U(x_t, a_t; \beta_U) + \epsilon_{a_t,t}$, where $\epsilon_{a_t,t}$ are i.i.d. type 1 extreme value utility shocks associated with each choice. β_F and β_U are finite vectors of structural parameters.

The Bellman equation for this problem is

$$V(x_t, \epsilon_t; \beta) = \max_{a_t \in \{0,1\}} \left\{ U(x_t, a_t; \beta_U) + \epsilon_{a_t,t} + \delta \int V(x_{t+1}, \epsilon_{t+1}; \theta) F_\epsilon(d\epsilon_{t+1}) \times F_x(dx_{t+1}|x_t, a_t; \beta_F) \right\}.$$

Following Rust (1987), define the alternative-specific value function,

$$\begin{aligned} \overline{V}(x_t, a_t; \beta) &= U(x_t, a_t; \beta_U) \\ &\quad + \delta \int V(x_{t+1}, \epsilon_{t+1}; \beta) F_\epsilon(d\epsilon_{t+1}) \\ &\quad \times F_x(dx_{t+1}|x_t, a_t; \beta_F) \\ &= U(x_t, a_t; \beta_U) + \delta \int \max_{a_{t+1} \in \{0,1\}} \{\overline{V}(x_{t+1}, a_{t+1}; \beta) \\ &\quad + \epsilon_{a_t,t}\} F_\epsilon(d\epsilon_{t+1}) F_x(dx_{t+1}|x_t, a_t; \beta_F) \\ &= U(x_t, a_t; \beta_U) \\ &\quad + \delta[0.5772 + \ln(e^{\overline{V}(x_{t+1},0;\beta)} + e^{\overline{V}(x_{t+1},1;\beta)})] \\ &\quad \times F_x(dx_{t+1}|x_t, a_t; \beta_F), \end{aligned} \tag{6}$$

and assume a renewal model in which $U(x_t, 0; \beta_U)$ and $F_x(x_{t+1}|x_t, 0; \beta_F)$ do not depend on x_t (action $a_t = 0$ "renews" the model). This allows us to normalize $\overline{V}(x_t, 0; \beta) = 0$ at all x_t.

The Hotz-Miller (1993) inversion implies that

$$\begin{aligned} \overline{V}(x_t, 1; \beta) - \overline{V}(x_t, 0; \beta) &= \overline{V}(x_t, 1; \beta) \\ &= \ln\left(\frac{\Pr(a_t = 1|x_t; \beta)}{1 - \Pr(a_t = 1|x_t; \beta)}\right). \end{aligned} \tag{7}$$

Now consider equation (6) evaluated at $a_t = 1$,

$$\begin{aligned} \overline{V}(x_t, 1; \beta) &= U(x_t, 1; \beta_U) \\ &\quad + \delta[0.5772 + \ln(e^{\overline{V}(x_{t+1},0;\beta)} + e^{\overline{V}(x_{t+1},1;\beta)})] \\ &\quad \times F_x(dx_{t+1}|x_t, 1; \beta_F). \end{aligned}$$

Substituting in equation (7) on both sides, using the normalization $\overline{V}(x_t, 0; \beta) = 0$, and rearranging results in

$$\Pr(a_t = 1|x_t; \beta) = \frac{\exp\left(U(x_t, 1; \beta_U) + \delta \int \lambda(x_{t+1}) F_x(dx_{t+1}|x_t, 1; \beta_F)\right)}{1 + \exp\left(U(x_t, 1; \beta_U) + \delta \int \lambda(x_{t+1}) F_x(dx_{t+1}|x_t, 1; \beta_F)\right)}. \tag{8}$$

where we define

$$\lambda(x_{t+1}) = 0.5772 + \ln\left(1 + \frac{\Pr(a_{t+1} = 1|x_{t+1}; \beta)}{1 - \Pr(a_{t+1} = 1/x_{t+1}; \beta)}\right)$$

Equation (8) implies that

$$E\left[a_t - \frac{\exp\left(U(x_t, 1; \beta_U) + \delta \int \lambda(z) F_x(dz|x_t, 1; \beta_F)\right)}{1 + \exp\left(U(x_t, 1; \beta_U) + \delta \int \lambda(z) F_x(dz|x_t, 1; \beta_F)\right)} \middle| x_t \right] = 0,$$

which can be used as a basis of second-step estimation for the structural parameters β_U in $U(x_t, 1; \beta_U)$ (and the discount factor δ if desired), as long as we have a first-stage nonparametric estimator of $\Pr(a = 1|z)$ and parametric estimator of β_F.[16]

Note that this moment condition depends on the nonparametric function $\Pr(a = 1|x)$ at all values of x, not only at the realized value of the conditioning variable x_t. Thus, this fits into the model of section IIIB. Since x is a continuous variable, $\Pr(a = 1|x)$ can be estimated nonparametrically either with a linear series approximation, or, following our results in section VB, in other (sufficiently flexible) ways, such as a sieve logit or sieve probit. So in sum, one can obtain semiparametric standard errors of the structural parameters in $U(x_t, 1; \beta_U)$ by simply treating the chosen sieves as parametric functions and applying the well-known parametric methodology of section II.

VI. Conclusion

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 estimator is numerically equivalent to a consistent sieve estimator of the semiparametric asymptotic variance. This numerical equivalence is significant because

[16] Note that this does not identify the structural parameter $U(x_t, 0; \theta_U)$ (since by assumption $U(x_t, 0; \theta_U)$ does not depend on x_t, this is just a scalar U_0). This parameter satisfies

$$U_0 = -\beta \int \left[0.5772 + \ln\left(1 + \frac{\Pr(a = 1|z)}{1 - \Pr(a = 1|z))}\right)\right] F_x(dz|x_t, 0; \theta_F).$$

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,[17] and the second class involves first-stage sieve nonparametric estimation based on a maximum likelihood–like criterion.[18] 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 (2010). Nevertheless, we believe that the numerical equivalence results in this paper already cover a very wide range of practical applications of two-step semiparametric estimation.

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 (1986) 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 such an information bound in complicated structural models.

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

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

REFERENCES

Aguirregabiria, V., "The Dynamics of Markups and Inventories in Retailing Firms," *Review of Economic Studies* 66 (1999), 275–308.

Aguirregabiria, V., and P. Mira, "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica* 70 (2002), 1519–1543.

——— "Sequential Estimation of Dynamic Discrete Games," *Econometrica* 75 (2007), 1–53.

Ai, C., and X. Chen, "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," *Journal of Econometrics* 141 (2007), 5–43.

Andrews, D., "Asymptotics for Semi-Parametric Econometric Models via Stochastic Equicontinuity, *Econometrica* 62 (1994), 43–72.

Bajari, P., C. L. Benkard, and J. Levin, "Estimating Dynamic Models of Imperfect Competition," *Econometrica* 75 (2007), 1331–1370.

Bajari, P., V. Chernozhukov, H. Hong, and D. Nekipelov, "Nonparametric and Semi-Parametric Analysis of a Dynamic Game Model," unpublished working paper (2010).

Bajari, P., H. Hong and D. Nekipelov, "Estimating Static Models of Strategic Interactions," *Journal of Business and Economic Statistics* 28 (2010), 469–482.

Chamberlain, G., "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics* 32 (1986), 189–218.

Chen, X., "Large Sample Sieve Estimation of Semi-Nonparametric Models," in James J. Heckman and Edward E. Lleamer (Eds.), *The Handbook of Econometrics*, Vol. 6B (Amsterdam: North-Holland, 2007).

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

Chen, X., and X. Shen, "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica* 66 (1998), 289–314.

Chen, X., Y. Fan, and V. Tsyrennikov, "Efficient Estimation of Semiparametric Multivariate Copula Models," *Journal of the American Statistical Association* 101 (2006), 1228–1240.

Collard-Wexler, A., "Demand Fluctuations and Plant Turnover in Ready Mix Concrete," NYU Stern, unpublished working paper (2006).

Dunne, T., S. Klimek, M. Roberts, and Y. Xu, "Entry and Exit in Geographic Markets," PSU, unpublished working paper (2006).

Ellickson, P., and S. Misra, "Supermarket Pricing Strategies," *Marketing Science* 27 (2008), 811–828.

Gonçalves, S., and H. White, "Bootstrap Standard Error Estimation for Linear Regressions," *Journal of the American Statistical Association* 100 (2005), 970–979.

Hahn, J., "The Efficiency Bound of the Mixed Proportional Hazard Model," *Review of Economic Studies* 61 (1994), 607–629.

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

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

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

Hotz, V. J., R. A. Miller, S. Sanders, and J. Smith, "A Simulation Estimator for Dynamic Models of Discrete Choice," *Review of Economic Studies* 61 (1994), 265–289.

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

Imbens, G., and J. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," Harvard University and Michigan State University, working paper (2005).

Jofre-Bonet, M., and M. Pesendorfer, "Estimation of a Dynamic Auction Game," *Econometrica* 71 (2003), 1443–1489.

Macieria, J., "Extending the Frontier: A Structural Model of Investment and Technological Competition in the Supercomputer Industry," mimeograph, Virginia Tech.

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

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

——— "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62 (1994), 1349–1382.

Newey, W. K., and D. F. McFadden, "Large Sample Estimation and Hypothesis Testing," in R. F. Engle III and D. F. McFadden (Eds.), *The Handbook of Econometrics*, vol. 4. (Amsterdam: North-Holland, 1994).

Olley, G. S., and A. Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica* 64 (1996), 1263–1297.

Pagan, A., "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review* 25 (1984), 221–247.

Pakes, A., M. Ostrovsky, and S. Berry, "Simple Estimators for the Parameters of Discrete Dynamic Games, with Entry/Exit Examples," *RAND Journal of Economics* 38 (2007), 373–399.

Pesendorfer, M., and P. Schmidt-Dengler, "Asymptotic Least Squares Estimators for Dynamic Games," *Review of Economic Studies* 75 (2008), 901–928.

Rubin, D., "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology* 66 (1974), 688–701.

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

Ryan, S., "The Costs of Environmental Regulation in a Regulated Industry," MIT unpublished working paper (2006).

Ryan, S., and C. Tucker, "Heterogeneity and the Dynamics of Technology Adoption," MIT, unpublished working paper (2008).

Shen, X., "On Methods of Sieves and Penalization," *Annals of Statistics* 25 (1997), 2555–2591.
Snider, C., "Predatory Incentives and Predation Policy: The American Airlines Case," University of Minnesota, unpublished working paper (2008).
Sweeting, A., "Dynamic Product Repositioning in Differentiated Product Industries: The Case of Format Switching in the Commercial Radio Industry," Duke University, unpublished working paper (2007).
Wooldridge, J. M., *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press, 2002).

APPENDIX A

Some Details for Section III

Model in Section IIIA.

We show that the two-step estimator considered in section IIIA is numerically identical to Ai and Chen's (2007) modified SMD estimator as long as $\widehat{h}_1(x_{1i}), \ldots, \widehat{h}_L(x_{Li})$ are approximated using the method of sieves.[19] The modified SMD estimator solves the minimization problem,

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

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

$$\mathcal{H}_{l,n} = \{h_l : h_l(x_l) = p_{l,1}(x_l)\theta_{(l),1} + \cdots + p_{l,K_{l,n}}(x_l)\theta_{(l),K_{l,n}} = h_l(x_l, \theta_{(l)})\}, \tag{A1}$$

we can see that the modified SMD is numerically equivalent to the following multistep estimator:

$$\widehat{\theta}_{(l)} = \underset{\theta_{(l),1},\ldots,\theta_{(l),K_{l,n}}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_{li} - (p_{l,1}(x_l)\theta_{(l),1} + \cdots + p_{l,K_{l,n}}(x_l)\theta_{(l),K_{l,n}}))^2, \quad l = 1, \ldots, L,$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} m(z_i, \widehat{\beta}, h_1(x_{1i}, \widehat{\theta}_{(L)}), \ldots, h_L(x_{Li}, \widehat{\theta}_{(L)})).$$

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

$$r(z_i, \alpha_*) = \begin{bmatrix} y_{1i} - h_{1*}(x_{1i}) \\ \vdots \\ y_{Li} - h_{L*}(x_{Li}) \end{bmatrix}, \tag{A2}$$

where $\alpha_* = (\beta_*, h_*)$ and h is an abbreviation of $(h_1, \ldots, h_L)$. We adopt a similar convention for $\widehat{h}$. Denote $\widehat{\alpha} = (\widehat{\beta}, \widehat{h})$. Assuming the sieve space $\mathcal{H}_n = \mathcal{H}_{1,n} \times \cdots \times \mathcal{H}_{L,n}$ with $\mathcal{H}_{l,n}$ given by equation (A1) for $l = 1, \ldots, L$, Ai and Chen's estimator $\widehat{V}$ of the asymptotic variance of $\widehat{\beta}$ can be computed using the following algorithm:

[19] See their equation (5) or their plug-in estimation equations (6) and (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 $\widehat{\mathbf{w}}^* = (\widehat{w}_1^*, \ldots, \widehat{w}_d^*)$ that solves for $j = 1, \ldots, d$,

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

2. Compute

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

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

and

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

3. Compute

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

Model in Section IIIB.

This model still fits into the framework of Ai and Chen (2007). According to their asymptotic variance formula for their modified SMD estimator $\widehat{\beta}$, to consider this model, we simply have to replace the term $\frac{\partial m(z_i,\widehat{\alpha})}{\partial h} w_j(x_i)$ in section III by $\frac{\partial m(z_i,\widehat{\alpha})}{\partial h}[w_j(\cdot)]$. Let the sieve space be $\mathcal{H}_n = \{h : h(\cdot) = \theta_1 p_1(\cdot) + \cdots + \theta_{K_n} p_{K_n}(\cdot)\}$. Ai and Chen's sieve estimator $\widehat{V}$ of the asymptotic variance of $\widehat{\beta}$ can then be computed by the following algorithm:

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

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

2. Compute

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

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

and

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

3. Compute

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

APPENDIX B

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 jth 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 $c^*=[c_1^*,\ldots,c_{d_1}^*]$.
Step 2: 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

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

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

$$\left(\begin{bmatrix}\mathbb{A}'\\ \mathbb{B}'\end{bmatrix}\Upsilon^{-1}[\mathbb{A}-\mathbb{B}c^*,\mathbb{B}]\right)^{-1} = \begin{bmatrix}((\mathbb{A}-\mathbb{B}c^*)'\Upsilon^{-1}(\mathbb{A}-\mathbb{B}c^*))^{-1} & -((\mathbb{A}-\mathbb{B}c^*)'\Upsilon^{-1}(\mathbb{A}-\mathbb{B}c^*))^{-1}(c^*)'\\ 0 & (\mathbb{B}'\Upsilon^{-1}\mathbb{B})^{-1}\end{bmatrix}. \tag{B1}$$

Now we have

$$\begin{aligned}&[\mathbb{A},\mathbb{B}]^{-1}\mathbb{F}\begin{bmatrix}\mathbb{A}'\\ \mathbb{B}'\end{bmatrix}^{-1}\\ &= (\Upsilon^{-1}[\mathbb{A},\mathbb{B}])^{-1}(\Upsilon^{-1}\mathbb{F}\Upsilon^{-1})\left(\begin{bmatrix}\mathbb{A}'\\ \mathbb{B}'\end{bmatrix}\Upsilon^{-1}\right)^{-1}\\ &= \left(\begin{bmatrix}(\mathbb{A}-\mathbb{B}c^*)'\\ \mathbb{B}'\end{bmatrix}\Upsilon^{-1}[\mathbb{A},\mathbb{B}]\right)^{-1}\\ &\quad\times\left(\begin{bmatrix}(\mathbb{A}-\mathbb{B}c^*)'\\ \mathbb{B}'\end{bmatrix}\Upsilon^{-1}\mathbb{F}\Upsilon^{-1}[\mathbb{A}-\mathbb{B}c^*,\mathbb{B}]\right)\\ &\quad\times\left(\begin{bmatrix}\mathbb{A}'\\ \mathbb{B}'\end{bmatrix}\Upsilon^{-1}[\mathbb{A}-\mathbb{B}c^*,\mathbb{B}]\right)^{-1}.\end{aligned} \tag{B2}$$

Using equation (B1), it can be shown that the upper-left block of equation (B2) 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.

APPENDIX C

Proof of Numerical Equivalence Result in Section III

We now prove the first main numerical equivalence result stated in section IIIA. 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 $l=1.,,.L$, 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 equation (5) with

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

and

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

where

$$P_i = \begin{bmatrix}p_1^{K_1}(x_{1,i}) & & 0\\ & \ddots & \\ 0 & & p_L^{K_L}(x_{L,i})\end{bmatrix}$$

$$q_i' = \frac{\partial m(z_i,\widehat{\alpha})}{\partial\beta'}$$

$$Q_i' = \left[\frac{\partial m(z_i,\widehat{\alpha})}{\partial h_1}p_1^{K_1}(x_{1,i})' \quad \cdots \quad \frac{\partial m(z_i,\widehat{\alpha})}{\partial h_L}p_L^{K_L}(x_{L,i})'\right]$$
$$y_i - h_i = \begin{bmatrix} y_{1i} - h_1(x_{1i},\widehat{\theta}_{(1)}) \\ \vdots \\ y_{Li} - h_L(x_{Li},\widehat{\theta}_{(L)}) \end{bmatrix}$$
$$m_i = m(z_i,\widehat{\beta},h_1(x_{1i},\widehat{\theta}_{(1)}),\ldots,h_L(x_{Li},\widehat{\theta}_{(L)})).$$

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

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

and

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

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

$$(\mathbb{A}_j - \mathbb{B}c)'\Upsilon^{-1}(\mathbb{A}_j - \mathbb{B}c) = c'\left(\frac{1}{n}\sum_{i=1}^n P_iP_i'\right)c + \overline{q}_j'\overline{q}_j - 2\overline{q}_j\overline{Q}'c + c'\overline{QQ}'c. \tag{C1}$$

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

$$c^* = \left(\left(\frac{1}{n}\sum_{i=1}^n P_iP_i'\right) + \overline{QQ}'\right)^{-1}\overline{Q}\overline{q}'.$$

Also, we have

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

and

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

and

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

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

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

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

$$\frac{1}{n}\sum_{i=1}^n (P_i'c)'(P_i'c) + \left(\frac{1}{n}\sum_{i=1}^n (q_{ij} - Q_i'c)\right)^2 = c'\left(\frac{1}{n}\sum_{i=1}^n P_iP_i'\right)c + \overline{q}_j'\overline{q}_j - 2\overline{q}_j\overline{Q}'c + c'\overline{QQ}'c. \tag{C2}$$

We can see that the same $\widehat{c}^*$ as above solves the practitioner's problem, equation (C2). Ai and Chen's estimator then requires calculating

$$\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) = \begin{bmatrix} P_i'\widehat{c}^* \\ \overline{q} - \overline{Q}'\widehat{c}^* \end{bmatrix}$$
$$\frac{1}{n}\sum_{i=1}^n (\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i))'\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) = (\widehat{c}^*)'\left(\frac{1}{n}\sum_{i=1}^n P_iP_i'\right)\widehat{c}^* + (\overline{q} - \overline{Q}'\widehat{c}^*)'(\overline{q} - \overline{Q}'\widehat{c}^*)$$

and

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

Note that

$$\frac{1}{n}\sum_{i=1}^n (\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i))'\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) = \widehat{\Lambda}_p,$$
$$\widehat{\Omega} = \widehat{\Omega}_p.$$

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

As for the proof of the result in section IIIB, all we need to do is to note that the same argument goes through writing

$$\frac{\partial g(z_i,\widehat{\beta},\widehat{\theta})}{\partial(\beta',\theta')} = \begin{bmatrix} 0 & -p^K(x_i)p^K(x_i)' \\ \frac{\partial m(z_i,\widehat{\beta},\widehat{h}_\theta)}{\partial\beta'} & \mathbf{m}(z_i,\widehat{\beta},\widehat{\theta})' \end{bmatrix} = \begin{bmatrix} 0 & -P_iP_i' \\ q_i' & Q_i' \end{bmatrix}$$

and

$$g(z_i,\widehat{\beta},\widehat{\theta}) = \begin{bmatrix} p^K(x_i)(y_i - \widehat{h}_\theta(x_i)) \\ m(z_i,\widehat{\beta},\widehat{h}_\theta) \end{bmatrix} = \begin{bmatrix} P_i & 0 \\ 0 & I_d \end{bmatrix}\begin{bmatrix} y_i - h_i \\ m_i \end{bmatrix},$$

with $P_i = p^K(x_i)$, $q_i' = \frac{\partial m(z_i,\widehat{\beta},\widehat{h}_\theta)}{\partial\beta'}$, $Q_i' = \mathbf{m}(z_i,\widehat{\beta},\widehat{\theta})'$, $y_i - h_i = y_i - \widehat{h}_\theta(x_i)$, and $m_i = m(z_i,\widehat{\beta},\widehat{h}_\theta)$.

APPENDIX D

Estimator of Asymptotic Variance of Sieve MLE

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 $(h(\cdot) = (h_1(\cdot), \ldots, h_L(\cdot))$ and each $h_l(\cdot)$ could depend on different argument x_l for $l = 1, \ldots, L$). 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 $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_L$ is replaced by a sieve parameter space $\mathcal{H}_n = \mathcal{H}_{1,n} \times \cdots \times \mathcal{H}_{L,n}$. In other words, the sieve MLE $(\widehat{\beta}, \widehat{h})$ is the solution to $\max_{(\beta,h)\in\mathcal{B}\times\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^n \ell(z_i, \beta, h(\cdot))$. Shen's (1997) result implies that $\widehat{\beta}$ is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient (under regularity conditions).

In the rest of this section, we recall the asymptotic variance of the sieve MLE $\widehat{\beta}$, present the estimator $\widehat{V}_{smle}$ of the asymptotic variance of $\widehat{\beta}$, and then argue that $\widehat{V}_{smle}$ 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

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

where, when $h() = (h_1, \ldots, h_L)$, 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 $\alpha, \overline{\alpha} \in \mathcal{A}$, denote $\ell'_{\alpha_*}[z, \alpha - \overline{\alpha}] = \ell'_{\alpha_*}[z, \alpha - \alpha_*] - \ell'_{\alpha_*}[z, \overline{\alpha} - \alpha_*]$, and define the metric $\|\cdot\|$ as

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

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

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

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

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

Denote

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

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 $\mathbf{w}^* = (w_1^*, \ldots, w_d^*)$ 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

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

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

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

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

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

which provides an intuitive reason that the sieve estimator $\widehat{V}_{smle}$ given in equation (D1) below is a plausible estimator of $\mathcal{I}^{-1}$.

We now present the estimator $\widehat{V}_{smle}$ of the asymptotic variance of $\widehat{\beta}$:

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

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

2. Compute

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

3. Compute

$$\widehat{V}_{smle} = \left(\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}(z_i)\widehat{\Delta}(z_i)'\right)^{-1}. \tag{D1}$$

Below, we provide a proof for the consistency of equation (D1). In the following we let $\|\cdot\|_s$ denote a metric (for example, 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 $\mathcal{W}_n = \{w \in \mathcal{H}_n : \|w\|_s \leq const. < \infty\}$. 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 A1. (1) $||v^*||^2 = \lambda'\mathcal{I}^{-1}\lambda < \infty$. (2) There is a $v_n^* = (v_\beta^*, -\mathbf{w}_n^* v_\beta^*)$ with $\mathbf{w}_n^* = (w_{n1}^*, \ldots, w_{nd}^*)$, $w_{nj}^* \in \mathcal{H}_n$ for all $j = 1, \ldots, d$, such that $||v_n^* - v^*|| = o(1)$.

Assumption A2. For all $j = 1, \ldots, d$, (1) $E[\sup_{\alpha\in\mathcal{N}_0, w\in\mathcal{W}_n} |g_j(z, \alpha, w)|^2] \leq const. < \infty$. (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 $E[\sup_{w\in\mathcal{W}_n} |U(z, w)|^2] \leq const. < \infty$.

Lemma 2. *Let $\widehat{\alpha} = (\widehat{\beta}, \widehat{h})$ be the sieve MLE such that $\|\widehat{\alpha} - \alpha_0\|_s = o_P(1)$. Suppose that $\{z_i\}$ is i.i.d. and assumptions A1 and A2 hold. If $K_n \to \infty$, $K_n/n \to 0$, then $\widehat{V}_{smle} = \mathcal{I}^{-1} + o_P(1)$.*

Proof. Assumption A2 implies that for all $j = 1, \ldots, d$, $\{(\frac{d\ell(z,\alpha)}{d\beta_j} - \frac{d\ell(z,\alpha)}{dh}[w])^2 : \alpha \in \mathcal{N}_0, w \in \mathcal{W}_n\}$ is a Glivenko-Cantelli class. Thus, uniformly over $w \in \mathcal{H}_n$,

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

where the last equality also follows from assumption A2. Here, E_{z_i} denotes the expectation taken only with respect to z_i regarding $\widehat{\alpha}$ as a nonstochastic constant. Thus,

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

where the second equation follows from assumption A1. The lemma now follows immediately.

We now argue that $\widehat{V}_p$ is exactly identical to $\widehat{V}_{smle}$. We recall that the practitioner's asymptotic variance for $\widehat{\beta}$, V_p, is simply the upper-left block of the above variance and covariance matrix

$$\begin{bmatrix} E\left[\frac{d\ell(z,\beta_*,\theta_*)}{d\beta}\frac{d\ell(z,\beta_*,\theta_*)}{d\beta'}\right] & E\left[\frac{d\ell(z,\beta_*,\theta_*)}{d\beta}\frac{d\ell(z,\beta_*,\theta_*)}{d\theta'}\right] \\ E\left[\frac{d\ell(z,\beta_*,\theta_*)}{d\theta}\frac{d\ell(z,\beta_*,\theta_*)}{d\beta'}\right] & E\left[\frac{d\ell(z,\beta_*,\theta_*)}{d\theta}\frac{d\ell(z,\beta_*,\theta_*)}{d\theta'}\right] \end{bmatrix}^{-1}.$$

The partitioned inverse formula, has another interpretation as the inverse of the variance of the least squares projection residual of $\frac{d\ell(z,\beta_*,\theta_*)}{d\beta}$ on $\frac{d\ell(z,\beta_*,\theta_*)}{d\theta'}$:

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

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

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

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:

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 $h_l(x_l) = p_l^{K_l}(x_l)'\theta_{(l)}$, $\theta = (\theta_{(1)}', \ldots, \theta_{(L)}')'$ and $c = (c_{(1)}', \ldots, c_{(L)}')'$, 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 $w = (w_{(1)}, \ldots, w_{(L)}) \in \mathcal{H}_n = \mathcal{H}_{1,n} \times \cdots \times \mathcal{H}_{L,n}$, with $\mathcal{H}_{l,n}$ given by equation (9) for $l = 1, \ldots, L$. 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.

APPENDIX E

Proof for Section VA on Restricted First Step

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

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

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

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

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

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

2. Compute

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

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

and

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

3. Compute

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

Next, we assume that the practitioner adopts the parametric specification $h(x_{l,i}, \theta) = p^K(x_{l,i})'\theta$, where $p^K(x) = (p_1(x), \ldots, p_K(x))'$, 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

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

where $h(x_{li}, \theta) = p^K(x_{l,i})'\theta$. (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 equation (5) with

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

and

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

where

$$P_i = [p^K(x_{1,i}) \quad \cdots \quad p^K(x_{L,i})],$$

$$q_i' = \frac{\partial m(z_i, \widehat{\alpha})}{\partial \beta'},$$

$$Q_i' = \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_1} p_1^K(x_i)' + \cdots + \frac{\partial m(z_i, \widehat{\alpha})}{\partial h_L} p_L^K(x_i)',$$

$$y_i - h_i = \begin{bmatrix} y_{1i} - h(x_{1i}, \widehat{\theta}) \\ \vdots \\ y_{Li} - h(x_{Li}, \widehat{\theta}) \end{bmatrix},$$

$$m_i = m(z_i, \widehat{\beta}, h(x_{1i}, \widehat{\theta}), \ldots, h(x_{Li}, \widehat{\theta})).$$

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

$$\mathbb{A} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} 0 \\ q_i' \end{bmatrix} = \begin{bmatrix} 0 \\ \overline{q}' \end{bmatrix},$$

$$\mathbb{B} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} -P_i P_i' \\ Q_i' \end{bmatrix} = \begin{bmatrix} -\frac{1}{n} \sum_{i=1}^{n} P_i P_i' \\ \overline{Q}' \end{bmatrix},$$

$$\mathbb{F} = \frac{1}{n} \sum_{i=1}^{n} g(z_i, \widehat{\beta}, \widehat{\theta}) g(z_i, \widehat{\beta}, \widehat{\theta})',$$

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

$$(\mathbb{A}_j - \mathbb{B}c)' \Upsilon^{-1} (\mathbb{A}_j - \mathbb{B}c) = c' \left(\frac{1}{n} \sum_{i=1}^{n} P_i P_i' \right) c + \overline{q}_j' \overline{q}_j - 2 \overline{q}_j \overline{Q}' c + c' \overline{Q} \overline{Q}' c. \tag{E1}$$

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

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

Also, we have

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

and

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

and

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

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

$$\widehat{V}_p=\widehat{\Lambda}_p^{-1}\widehat{\Omega}_p\big(\widehat{\Lambda}_p^{-1}\big)'.$$

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

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\big(P_i'c\big)'\big(P_i'c\big)+\left(\frac{1}{n}\sum_{i=1}^{n}\big(q_{ij}-Q_i'c\big)\right)^2&=c'\left(\frac{1}{n}\sum_{i=1}^{n}P_iP_i'\right)c\\
&\quad+\overline{q}_j'\overline{q}_j-2\overline{q}_j\overline{Q}'c+c'\overline{QQ}'c. \quad \text{(E2)}
\end{aligned}
$$

We can see that the same $\widehat{c}^*$ as above solves the practitioner's problem, equation (C1). Ai and Chen's estimator then requires calculating

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

and

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

Note that

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

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

APPENDIX F

Proof for Section VB on First-Step Sieve M-Estimation

We propose the following sieve estimator:

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

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

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

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

Explicit Characterization of the Influence Function

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

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

Define the functional $f:\mathcal{H}\to\mathbb{R}$ 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

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

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, $\sqrt{n}(\widehat{\beta}-\beta_*)\to N(0,V)$, 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 $\sqrt{n}(\widehat{\beta}-\beta_*) \to N(0, v_\beta^* \Omega v_\beta^*)$, 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

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

and $\mathbf{w}^*$ solves

$$\inf_{w \in \mathcal{H}} E\left[\left(-\frac{\partial^2 \ell(z_i, h_*)}{\partial h \partial h}[w, w] \right) + \left(E\left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h}[w] \right] \right)^2 \right]. \quad \text{(F1)}$$

Equivalence of These Two Asymptotic Variance Characterizations

For the simple case of scalar $h()$ function of x, the optimization problem (F1) can be solved in closed form. Note that

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

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

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

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,

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

and

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

from which we obtain

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

Consistent Estimator of the Asymptotic Variance

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

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

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

and

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

2. Compute

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

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

and

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

3. Compute

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

Numerical Equivalence

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, equation (5),

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

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

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 $p^K(x) = (p_1(x), \ldots., p_K(x))'$, where $K = K_n$ is a function of n although it is perceived to be fixed for our fictitious practitioner. Then:

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

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

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

and

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

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

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

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

$$\begin{aligned} \widehat{v}_p^{-1} &\equiv (\mathbb{A} - \mathbb{B}c^*)'\Upsilon^{-1}(\mathbb{A} - \mathbb{B}c^*) = (c^*)'\overline{R}c^* + (\overline{q} - \overline{Q}'c^*)'(\overline{q} - \overline{Q}'c^*) \\ &= \left(\widehat{v}_\beta^*\right)^{-1}. \end{aligned}$$

We also have

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

and

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

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

$$\widehat{V}_p = \widehat{v}_p \widehat{\Omega}_p \widehat{v}_p.$$

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