A PRACTICAL ASYMPTOTIC VARIANCE ESTIMATOR FOR TWO-STEP SEMIPARAMETRIC ESTIMATORS

By

Daniel Ackerberg, Xiaohong Chen, and Jinyong Hahn

May 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1803

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators

Daniel Ackerberg, UCLA
Xiaohong Chen, Yale University
Jinyong Hahn*, UCLA

First Version: March 20, 2009; Final Version: July, 2010.
Forthcoming in the *Review of Economics and Statistics*

Abstract

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

EconLit Subject Descriptor: C140

1 Introduction

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

*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.

(2002, 2007), Jofre-Bonet and Pesendorfer (2003), Bajari, Benkard, and Levin (2007), Pakes, Ostrovsky, and Berry (2007), Pesendorfer and Schmidt-Dengler (2007), Bajari, Hong, Krainer, and Nekipelov (2008), and Bajari, Chernozhukov, Hong, and Nekipelov (2010)). These two-step semiparametric estimators are 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 involves estimating 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, e.g. the number of terms in a series approximation or the flexibility of a sieve. But naïve 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.

There is a long line of theoretical literature that derives expressions for semiparametric asymptotic variances of two-step estimators (Newey (1994), Andrews (1994), Newey and McFadden (1994), Ai and Chen (2007), Chen, Linton and van Keilegom (2003), Ichimura and Lee (2010), to name a few). Some of these papers also show how to consistently estimate the asymptotic variances. While these theoretical results are useful, their implementation is typically not straightforward in practice. These limitations have often lead applied researchers to use the bootstrap to estimate asymptotic variances (e.g., Ryan (2006), Ellickson and Misra (2008), Macieira (2008)), but this can be computationally demanding and may also be difficult to justify theoretically.[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. On the other hand, the asymptotic variances of the two estimators will be different, as researcher A is in a parametric world where the total number of unknown parameters is constant (and finite), while researcher B is in a semiparametric world where the total number of unknown parameters is increasing to infinity.

Our results concern the *estimated* asymptotic variance of the structural parameters. We show, 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 2, see, e.g. Murphy and Topel (1985), or Newey and McFadden (1994)). In other words, researcher A and researcher B will obtain numerically identical variance estimates (for the structural parameters). This is true even though they are estimating different objects asymptotically – the true asymptotic parametric variance vs. the true asymptotic semiparametric variance of the finite dimensional structural parameters of interest. To the best of our knowledge, Newey (1994, Section 6) 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 quick review of the standard two-step parametric approach in Section 2. Section 3 presents equivalence results for models where the first-stage sieve nonparametric estimation is based on conditional moment restrictions. Section 4 considers the case where first-stage sieve nonparametric estimation is based on a maximum-likelihood like criterion. Section 5 considers various extensions of the result, e.g. to situations where the second stage is overidentified, and gives explicit examples of applications of our approach to the IO and Labor literatures discussed above. Section 6 concludes.

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

In this section, we provide a brief review of how to estimate the asymptotic variance of two-step parametric M-estimators. We assume that a researcher estimates a finite dimensional parameter vector θ using a first-step M-estimator (e.g. OLS, NLLS, MLE, method of moments). This estimate is then plugged into a second-step M-estimator which is used to estimate another 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 even be found in standard textbooks such as Wooldridge (2002, Chapter 12.4).

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

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

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

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

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

In order to assess the asymptotic variance of $\sqrt{n}\left(\widehat{\beta}-\beta_*\right)$ 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]

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

where

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

Note that (3) is equivalent to $\frac{1}{n}\sum_{i=1}^{n}\varphi\left(z_i,\widehat{\theta}\right)=0$ and $\frac{1}{n}\sum_{i=1}^{n}\psi\left(z_i,\widehat{\beta},\widehat{\theta}\right)=0$. Therefore, the $\widehat{\theta}$ and $\widehat{\beta}$ that solve (3) are numerically identical to $\widehat{\theta}$ and $\widehat{\beta}$ that solve (1) and (2). Letting $\alpha=(\beta',\theta')'$ and recognizing that $\widehat{\alpha}=\left(\widehat{\beta}',\widehat{\theta}'\right)'$ is an M-estimator, we can then use standard arguments[8] to compute the asymptotic variance of $\sqrt{n}\left(\widehat{\alpha}-\alpha_*\right)$ i.e. a consistent estimator of the asymptotic variance of $\sqrt{n}\left(\widehat{\alpha}-\alpha_*\right)$ is given by

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

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

3 Estimator of Asymptotic Variance of Two-Step Semiparametric Estimators

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

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

Researcher B, on the other hand, makes the correct nonparametric assumption that the number of terms in the series approximation increases to infinity as an appropriate function of the sample size. Therefore, Researcher B would like to compute a consistent estimator of the asymptotic variance of $\widehat{\beta}$ 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 (i.e., the theoretical formula expressed in population expectations) will generally be different from Researcher B's. In other words, Researcher A is trying to estimate a different theoretical variance object than Researcher B.[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 non-parametric 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 non-parametric function.

3.1 Dependence of Second-Stage on the Non-Parametric Function

Consider a model given by the following moment restrictions

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

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

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

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 Naïve practitioner's estimator We now consider how the semiparametric estimators $\widehat{\beta}$ and $\widehat{V}$ relate to what one obtains if the estimation problem is approached from a purely parametric perspective (i.e. Researcher A). First, note that a parametric estimator based on the parametric specification $h_l(x_l) = p_{l,1}(x_l)\,\theta_{(l),1}+\cdots+p_{l,K_l}(x_l)\,\theta_{(l),K_l} = h_l\left(x_l,\theta_{(l)}\right)$ (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 $E\left[g\left(z_i,\beta_*,\theta_*\right)\right]=0$, where

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

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

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

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

3.2 Extension: Dependence of Second-Stage on Full Non-Parametric Function

Consider a model where

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

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

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

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

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

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

4 Estimator of Asymptotic Variance of Sieve MLE

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

Assume that we observe z_i for each individual. We further assume that z_i are independent and identically distributed.[10] The log likelihood of the data $\{z_i\}_{i=1}^n$ is given by $\frac{1}{n}\sum_{i=1}^n \ell\left(z_i, \beta, h\left(\cdot\right)\right)$, where $\beta \in \mathcal{B}$ is a vector of finite-dimensional parameter of interest and $h \in \mathcal{H}$ is a vector of L real-valued unknown functions (i.e., $h\left(\cdot\right) = \left(h_1\left(\cdot\right), \ldots, h_L\left(\cdot\right)\right)$ and each $h_l\left(\cdot\right)$ could depend on different argument x_l for $l = 1, ..., L$). We take $h\left(\cdot\right)$ 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 $\sup_{(\beta,h)\in\mathcal{B}\times\mathcal{H}} E\left[\ell\left(z_i, \beta, h\left(\cdot\right)\right)\right]$. 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

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

with $p_l^{K_l}(x_l) = \left(p_{l,1}\left(x_l\right), ..., p_{l,K_l}\left(x_l\right)\right)'$, where $K_l = K_{l,n}$ is a function of n although it is perceived to be fixed for our fictitious practitioner. Denote $\theta = \left(\theta'_{(1)}, \ldots, \theta'_{(L)}\right)'$ which is a $K \times 1-$vector with $K = K_1 + \cdots + K_L$. The parametric practitioner would estimate

$(\beta_*, \theta_*) = \text{argmax}_{\beta,\theta} E[\ell(z_i, \beta, \theta)]$ via parametric MLE, and obtain:

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

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 $\left(\widehat{\beta}, \widehat{\theta}\right)'$ can be consistently estimated by the following matrix:

$$\left[\begin{array}{cc} \frac{1}{n}\sum_{i=1}^n \frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\beta}\frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\beta'} & \frac{1}{n}\sum_{i=1}^n \frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\beta}\frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\theta'} \\ \frac{1}{n}\sum_{i=1}^n \frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\theta}\frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\beta'} & \frac{1}{n}\sum_{i=1}^n \frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\theta}\frac{d\ell\left(z_i,\widehat{\beta},\widehat{\theta}\right)}{d\theta'} \end{array}\right]^{-1},$$

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 outer-product is used for calculation of information, "parametric" inference for β is numerically identical to semiparametric inference.

5 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.

5.1 First Step with Restriction

As another extension, we can consider a model where

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

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

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

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 (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.)

5.2 Nonparametric Sieve M-Estimation As First Step

Next consider semiparametric two-step estimation where the first-step involves nonparametric sieve, maximum-likelihood-like, M-estimation in the first step. Again, these nonparametric estimators are plugged into a parametric moment equation to compute an estimator $\widehat{\beta}$ 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:

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

where $\ell(z_i, h)$ is any criterion function and $h = (h_1, ..., h_L)$ 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 $\sqrt{n}\left(\widehat{\beta} - \beta_*\right)$, 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.

5.3 Overidentified Second Step

So far, we have implicitly assumed that the second step is exactly identified, i.e., $\dim(m) = \dim(\beta)$. 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 $\frac{1}{n}\sum_{i=1}^n m\left(z_i, \beta, \widehat{h}\right)$ as close to zero as possible. This is usually done by

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

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

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 $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n m\left(z_i, \beta_0, \widehat{h}\right)$, 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, i.e., by the sieve method as discussed at the end of Section 3.1.

2. Using an arbitrary weight matrix W, minimize the sample moment $\left(\frac{1}{n}\sum_{i=1}^{n} m\left(z_i,\beta,\widehat{h}\right)\right)' W \left(\frac{1}{n}\sum_{i=1}^{n} m\left(z_i,\beta,\widehat{h}\right)\right)$ 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 $\frac{1}{n}\sum_{i=1}^{n}\left(m\left(z_i,\overline{\beta},\widehat{h}\right)-\widehat{\mu}\right)$ equal to zero (this estimation problem is exactly identified (and trivial))

4. Again consider $\overline{\beta}$ to be fixed. Apply Murphy-Topel, i.e., the naïve practitioner's estimator of the asymptotic variance discussed in Section 3.1, to the moment conditions corresponding to Steps 1 and 3, i.e.

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

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

5. Solve the minimization problem

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

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

6. Compute

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

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 $\left(\widehat{h},\widehat{\mu}\right)$ in Step 4.)

This algorithm is the procedure of a naïve 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 $\frac{1}{n}\sum_{i=1}^{n} m\left(z_i,\beta_0,\widehat{h}\right)$, 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.

5.4 Example: Estimation of Average Treatment Effects

There is a large body of literature on estimation of average treatment effects. We discuss two estimators that fit into our framework. Consider the effect of a treatment on some outcome variable of interest. Let d_i denote the dummy variable such that $d_i = 1$ when treatment is given to the 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 only observes d_i and $y_i \equiv d_i y_{1i} + (1 - d_i)\, y_{0i}$. On the other hand, the *average treatment effect* $\beta \equiv E\left[y_{1i} - y_{0i}\right]$ can be identified and consistently estimated when d_i is assigned independent of (y_{0i}, y_{1i}). Extending this idea, Hahn (1998) and Hirano, Imbens, and Ridder (2003) proposed estimators of the average treatment effect when the treatment d_i is assigned independent of (y_{0i}, y_{1i}) given the observed covariates x_i.

Hahn's (1998) estimator is

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

and Hirano, Imbens, and Ridder'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 $E\left[d_i y_i \mid x_i\right]$, $E\left[(1 - d_i)\, y_i \mid x_i\right]$, and $E\left[d_i \mid x_i\right]$. We can easily recognize that they fit into our framework discussed in Section 3.

Hirano, Imbens, and Ridder (2003) also consider an estimator where the propensity score $p(x_i) = E\left[d_i \mid x_i\right]$ is estimated by nonparametric maximum likelihood estimation with a Logit specification. This alternative estimator fits into our framework in Section 5.2. We note that our result there can in principle accommodate the case where the propensity score is specified as a Probit model, which has some minor theoretical significance because the proof in Hirano, Imbens and Ridder (2003) can address only a Logit specification.[14]

Note that implementation of Murphy-Topel would require writing down moments. As for

Hahn's (1998) estimator, the moments are

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

and as for Hirano, Imbens, and Ridder's (2003) estimator, they are

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

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.

5.5 Example: 2-Step Estimation of Dynamic Models

There is a large recent literature on two-step semiparametric estimation of single agent dynamic programming problems and dynamic games, including Hotz and Miller (1993, 1994), Aguirregabiria and Mira (2002, 2007), Jofre-Bonet and Pesendorfer (2003), Bajari, Benkard, and Levin (2007), Pakes, Ostrovsky, and Berry (2007), Pesendorfer and Schmidt-Dengler (2008), and Bajari, Chernozhukov, Hong, and Nekipelov (2010). The basic idea behind these estimators is that "reduced form" policy functions describing optimal agent behavior can be non-parametrically estimated in a first stage.[15] These estimated policy functions can then be used as an input into in 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 agent(s)' dynamic programming problems, hence reducing computational burden relative to one step estimation. **I**n 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

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

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

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

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

The Hotz-Miller (1993) inversion implies that

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

Now, consider (6) evaluated at $a_t = 1$, i.e.

$$\overline{V}\left(x_t, 1; \beta\right) = U\left(x_t, 1; \beta_U\right) + \delta \int \left[0.5772 + \ln\left(e^{\overline{V}\left(x_{t+1}, 0; \beta\right)} + e^{\overline{V}\left(x_{t+1}, 1; \beta\right)}\right)\right] F_x\left(\mathrm{d}x_{t+1} \middle| x_t, 1; \beta_F\right)$$

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

$$\Pr\left(a_t = 1 \middle| x_t; \beta\right) = \frac{\exp\left(U\left(x_t, 1; \beta_U\right) + \delta \int \left[0.5772 + \ln\left(1 + \frac{\Pr\left(a_{t+1} = 1 | x_{t+1}; \beta\right)}{1 - \Pr\left(a_{t+1} = 1 | x_{t+1}; \beta\right)}\right)\right] F_x\left(\mathrm{d}x_{t+1} \middle| x_t, 1; \beta_F\right)\right)}{1 + \exp\left(U\left(x_t, 1; \beta_U\right) + \delta \int \left[0.5772 + \ln\left(1 + \frac{\Pr\left(a_{t+1} = 1 | x_{t+1}; \beta\right)}{1 - \Pr\left(a_{t+1} = 1 | x_{t+1}; \beta\right)}\right)\right] F_x\left(\mathrm{d}x_{t+1} \middle| x_t, 1; \beta_F\right)\right)}. \tag{8}$$

Equation (8) implies that

$$E\left[a_t - \left.\frac{\exp\left(U\left(x_t, 1; \beta_U\right) + \delta \int \left[0.5772 + \ln\left(1 + \frac{\Pr(a=1|z)}{1 - \Pr(a=1|z)}\right)\right] F_x\left(\mathrm{d}z \middle| x_t, 1; \beta_F\right)\right)}{1 + \exp\left(U\left(x_t, 1; \beta_U\right) + \delta \int \left[0.5772 + \ln\left(1 + \frac{\Pr(a=1|z)}{1 - \Pr(a=1|z)}\right)\right] F_x\left(\mathrm{d}z \middle| x_t, 1; \beta_F\right)\right)} \right| x_t\right] = 0,$$

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),[16] as long as we have a first stage nonparametric estimator of $\Pr\left(a = 1 \middle| z\right)$ and parametric estimator of β_F.

Note that this moment condition depends on the non-parametric function $\Pr\left(a = 1 \middle| x\right)$ 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 3.2. Since x is a continuous variable, $\Pr\left(a = 1 \middle| x\right)$ can be estimated non-parametrically either with a linear series approximation, or, following our results in Section 5.2, in other (sufficiently flexible) ways, e.g. a sieve Logit or sieve Probit. So in sum, one can obtain semi-parametric 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 2.

6 Concluding Remarks

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

We derived the numerical equivalence for two classes of semiparametric two-step estimators: the first class involves first-stage sieve nonparametric estimation based on conditional moment restrictions;[17] 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 our current paper already cover a very wide range of practical applications of two-step semiparametric estimation.

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

Appendix

A Some Details for Section 3

Model in Section 3.1 We show that the two-step estimator considered in Section 3.1 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:

$$\mathcal{H}_{l,n}=\left\{h_l : h_l(x_l)=p_{l,1}(x_l)\,\theta_{(l),1}+\cdots+p_{l,K_{l,n}}(x_l)\,\theta_{(l),K_{l,n}}=h_l\left(x_l,\theta_{(l)}\right)\right\}, \tag{9}$$

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

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

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 $\sqrt{n}\left(\widehat{\beta}-\beta_*\right)$, which we now describe. For simplicity of notation, we will write

$$r(z_i,\alpha_*)=\begin{bmatrix} y_{1i}-h_{1*}(x_{1i})\\ \vdots \\ y_{Li}-h_{L*}(x_{Li})\end{bmatrix} \tag{10}$$

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 (9) for $l = 1, ..., L$, Ai and Chen's estimator $\widehat{V}$ of the asymptotic variance of $\widehat{\beta}$ can be computed using the following algorithm:

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

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

2. Compute

$$\rho\left(z_i, \widehat{\alpha}\right) = \left[\begin{array}{c} r\left(z_i, \widehat{\alpha}\right) \\ m\left(z_i, \widehat{\alpha}\right) \end{array} \right],$$

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

and

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

3. Compute

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

Model in Section 3.2 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 $\frac{\partial m(z_i,\widehat{\alpha})}{\partial h} w_j\left(x_i\right)$ in Section 3 by $\frac{\partial m(z_i,\widehat{\alpha})}{\partial h}\left[w_j\left(\cdot\right)\right]$. Let the sieve space be $\mathcal{H}_n = \{h : h(\cdot) = \theta_1 p_1\left(\cdot\right) + \cdots + \theta_{K_n} p_{K_n}\left(\cdot\right)\}$. 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 $\widehat{\mathbf{w}}^* = (\widehat{w}_1^*, ..., \widehat{w}_d^*)$ for $j = 1, ..., d$ as

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

2. Compute

$$\rho\left(z_i, \widehat{\alpha}\right) = \left[\begin{array}{c} y_i - \widehat{h}\left(x_i\right) \\ m\left(z_i, \widehat{\beta}, \widehat{h}\right) \end{array} \right],$$

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

and

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

3. Compute

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

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*

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

for some symmetric positive definite matrix Υ. Let c_j^* denote the solution, and let $c^*=\left[c_1^*,\ldots,c_{d_1}^*\right]$.
***Step2**: Compute*

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

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

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

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

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

and

$$\left(\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} [\mathbb{A} - \mathbb{B}c^*, \mathbb{B}]\right)^{-1}$$
$$= \begin{bmatrix} \left((\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*)\right)^{-1} & -\left((\mathbb{A} - \mathbb{B}c^*)' \Upsilon^{-1} (\mathbb{A} - \mathbb{B}c^*)\right)^{-1} (c^*)' \\ 0 & (\mathbb{B}'\Upsilon^{-1}\mathbb{B})^{-1} \end{bmatrix} \quad (11)$$

Now, we have

$$[\mathbb{A}, \mathbb{B}]^{-1} \mathbb{F} \begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix}^{-1}$$
$$= \left(\Upsilon^{-1} [\mathbb{A}, \mathbb{B}]\right)^{-1} \left(\Upsilon^{-1}\mathbb{F}\Upsilon^{-1}\right) \left(\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1}\right)^{-1}$$
$$= \left(\begin{bmatrix} (\mathbb{A} - \mathbb{B}c^*)' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} [\mathbb{A}, \mathbb{B}]\right)^{-1} \left(\begin{bmatrix} (\mathbb{A} - \mathbb{B}c^*)' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1}\mathbb{F}\Upsilon^{-1} [\mathbb{A} - \mathbb{B}c^*, \mathbb{B}]\right) \left(\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix} \Upsilon^{-1} [\mathbb{A} - \mathbb{B}c^*, \mathbb{B}]\right)^{-1} \quad (12)$$

Using (11), it can be shown that the upper left block of (12) is equal to

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

which proves the validity of the algorithm. ■

C Proof of Numerical Equivalence Result in Section 3

We now prove the first main numerical equivalence result stated in Section 3.1. We assume that the practitioner adopts the parametric specification $h_l\left(x_{l,i}, \theta_{(l)}\right) = p_l^{K_l}(x_{l,i})' \theta_{(l)}$, 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 (5) with

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

and

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

where

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

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

$$
\begin{aligned}
\mathbb{A} &= \frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c} 0 \\ q_i' \end{array}\right] = \left[\begin{array}{c} 0 \\ \overline{q}' \end{array}\right] \\
\mathbb{B} &= \frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c} -P_iP_i' \\ Q_i' \end{array}\right] = \left[\begin{array}{c} -\frac{1}{n}\sum_{i=1}^{n}P_iP_i' \\ \overline{Q}' \end{array}\right] \\
\mathbb{F} &= \frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\widehat{\beta},\widehat{\theta}\right) g\left(z_i,\widehat{\beta},\widehat{\theta}\right)'
\end{aligned}
$$

and

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

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

$$
\left(\mathbb{A}_j-\mathbb{B}c\right)'\Upsilon^{-1}\left(\mathbb{A}_j-\mathbb{B}c\right) = c'\left(\frac{1}{n}\sum_{i=1}^{n}P_iP_i'\right)c + \overline{q}_j'\overline{q}_j - 2\overline{q}_j\overline{Q}'c + c'\overline{QQ}'c \tag{13}
$$

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

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

and

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

and

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

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

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

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

$$\frac{1}{n}\sum_{i=1}^n (P_i'c)' (P_i'c) + \left(\frac{1}{n}\sum_{i=1}^n (q_{ij} - Q_i'c)\right)^2 = c' \left(\frac{1}{n}\sum_{i=1}^n P_i P_i'\right) c + \overline{q}_j'\overline{q}_j - 2\overline{q}_j\overline{Q}'c + c'\overline{Q}\overline{Q}'c \quad (14)$$

We can see that the same $\widehat{c}^*$ as above solves the practitioner's problem (13). Ai and Chen's estimator then requires calculating

$$\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) = \begin{bmatrix} P_i'\widehat{c}^* \\ \overline{q} - \overline{Q}'\widehat{c}^* \end{bmatrix}$$

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

and

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

Note that

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

$$\widehat{\Omega}=\widehat{\Omega}_p$$

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 3.2, all we need to do is to note that the same argument goes through writing

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

and

$$g\left(z_i,\widehat{\beta},\widehat{\theta}\right)=\left[\begin{array}{c}p^K(x_i)\left(y_i-\widehat{h}_\theta(x_i)\right)\\ m\left(z_i,\widehat{\beta},\widehat{h}_\theta\right)\end{array}\right]=\left[\begin{array}{cc}P_i & 0\\ 0 & I_d\end{array}\right]\left[\begin{array}{c}y_i-h_i\\ m_i\end{array}\right]$$

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

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 (i.e., $h(\cdot)=(h_1(\cdot),\ldots,h_L(\cdot))$ and each $h_l(\cdot)$ could depend on different argument x_l for $l=1,...,L$). 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 $\sup_{(\beta,h)\in\mathcal{B}\times\mathcal{H}} E\left[\ell\left(z_i,\beta,h\left(\cdot\right)\right)\right]$. 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 $\left(\widehat{\beta},\widehat{h}\right)$ is the solution to $\max_{(\beta,h)\in\mathcal{B}\times\mathcal{H}_n}\frac{1}{n}\sum_{i=1}^{n}\ell\left(z_i,\beta,h\left(\cdot\right)\right)$. Shen's result (1997) implies that $\widehat{\beta}$ is $\sqrt{n}$-consistent, asymptotically normal and semiparametrically efficient (under regularity conditions).

In the rest of this section, we will 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 $\alpha\left(\alpha_*,\tau\right)\in\mathcal{A}$ is a path in τ connecting α_* and α such that $\alpha\left(\alpha_*,0\right)=\alpha_*$ and $\alpha\left(\alpha_*,1\right)=\alpha$. Let

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

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

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

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

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

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

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

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

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

Denote

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

and

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

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

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

We then have

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

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

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

where

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

In other words, we have

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

which provides an intuitive reason why the sieve estimator $\widehat{V}_{smle}$ given in (15) 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^*, $j = 1, ..., d$:

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

2. Compute

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

3. Compute

$$\widehat{V}_{smle} = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{\Delta}(z_i) \widehat{\Delta}(z_i)' \right)^{-1}. \tag{15}$$

Below, we provide a proof for the consistency of (15). In the following we let $\|\cdot\|_s$ denote a metric (e.g., the supreme norm or the mean squared metric) on $\mathcal{A} = \Theta \times \mathcal{H}$. Denote $\mathcal{N}_0 = \{\alpha \in \mathcal{A} : \|\alpha - \alpha_*\|_s = o(1)\}$ and $\mathcal{W}_n = \{w \in \mathcal{H}_n : \|w\|_s \leq const. < \infty\}$. Also denote $g_j(z, \alpha, w) = \frac{d\ell(z,\alpha)}{d\beta_j} - \frac{d\ell(z,\alpha)}{dh}[w]$.

We impose the following assumptions:

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

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

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

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

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

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

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

where the second equation follows from assumption A.1. 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 on the other hand, 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 $\frac{d\ell(z,\beta_*,\theta_*)}{d\theta'}$:

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

where

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

and

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

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

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

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

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

2. Compute

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

3. Compute

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

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

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

Therefore, the minimization problem over $c \in R^K$ is in fact identical to the minimization problem over all linear combinations $w_{(l)} = p_l^{K_l}\left(\cdot\right)' c_{(l)}$, which in turn is identical to the minimization over $w = \left(w_{(1)}, \ldots, w_{(L)}\right) \in \mathcal{H}_n = \mathcal{H}_{1,n} \times \cdots \times \mathcal{H}_{L,n}$, with $\mathcal{H}_{l,n}$ given by (9) for $l = 1, ..., L$. 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.

E Proof for Section 5.1 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 will write

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

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 $\widehat{\mathbf{w}}^* = (\widehat{w}_1^*, ..., \widehat{w}_d^*)$ for $j = 1, ..., d$ as
$$\widehat{w}_j^* = \underset{w \in \mathcal{H}_n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \begin{array}{c} \left(\frac{\partial r(z_i,\widehat{\alpha})}{\partial \beta_j} - \sum_{l=1}^{L} \frac{\partial r(z_i,\widehat{\alpha})}{\partial h_l} w_j(x_{l,i})\right)' \left(\frac{\partial r(z_i,\widehat{\alpha})}{\partial \beta_j} - \sum_{l=1}^{L} \frac{\partial r(z_i,\widehat{\alpha})}{\partial h_l} w_j(x_{l,i})\right) \\ + \left(\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial m(z_i,\widehat{\alpha})}{\partial \beta_j} - \sum_{l=1}^{L} \frac{\partial m(z_i,\widehat{\alpha})}{\partial h_l} w_j(x_{l,i})\right)\right)^2 \end{array} \right\}.$$
(We write $h_*(x_{li}) = h_{l*}(x_{li})$ for ease of accounting.)

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

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

Next, we assume that the practitioner adopts the parametric specification $h(x_{l,i}, \theta) = p^K(x_{l,i})' \theta$, where $p^K(x) = (p_1(x), ..., p_K(x))'$, 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 (5) with
$$\frac{\partial g\left(z_i, \widehat{\beta}, \widehat{\theta}\right)}{\partial(\beta', \theta')} = \begin{bmatrix} 0 & -p^K(x_{1,i})\left(p^K(x_{1,i})\right)' - \cdots - p^K(x_{L,i})\left(p^K(x_{L,i})\right)' \\ \frac{\partial m(z_i,\widehat{\alpha})}{\partial \beta'} & \frac{\partial m(z_i,\widehat{\alpha})}{\partial h_1} p_1^K(x_i)' + \cdots + \frac{\partial m(z_i,\widehat{\alpha})}{\partial h_L} p_L^K(x_i)' \end{bmatrix}$$
$$\equiv \begin{bmatrix} 0 & -P_i P_i' \\ q_i' & Q_i' \end{bmatrix}$$

and

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

where

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

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

$$
\begin{aligned}
\mathbb{A} &= \frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c} 0 \\ q_i' \end{array}\right] = \left[\begin{array}{c} 0 \\ \overline{q}' \end{array}\right] \\
\mathbb{B} &= \frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c} -P_iP_i' \\ Q_i' \end{array}\right] = \left[\begin{array}{c} -\frac{1}{n}\sum_{i=1}^{n}P_iP_i' \\ \overline{Q}' \end{array}\right] \\
\mathbb{F} &= \frac{1}{n}\sum_{i=1}^{n}g\left(z_i,\widehat{\beta},\widehat{\theta}\right)g\left(z_i,\widehat{\beta},\widehat{\theta}\right)'
\end{aligned}
$$

and

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

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

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

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

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

and

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

and

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

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

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

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

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

We can see that the same $\widehat{c}^*$ as above solves the practitioner's problem (13). Ai and Chen's estimator then requires calculating

$$\widehat{\Delta}_{\widehat{\mathbf{w}}^*}(z_i) = \begin{bmatrix} P_i'\widehat{c}^* \\ \overline{q} - \overline{Q}'\widehat{c}^* \end{bmatrix},$$

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

and

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

Note that

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

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

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

F Proof for Section 5.2 on First Step Sieve M-Estimation

We propose the following sieve estimator:

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

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

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

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 (i.e., $\dim(\beta) = 1$) 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

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

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

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

for

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

We can write

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

where

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

and

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

It follows that the influence function is

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

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

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

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

$$\inf_{w \in \mathcal{H}} E\left[\left(-\frac{\partial^2 \ell(z_i, h_*)}{\partial h \partial h}[w, w]\right) + \left(E\left[\frac{\partial m(z_i, \alpha_*)}{\partial \beta} - \frac{\partial m(z_i, \alpha_*)}{\partial h}[w]\right]\right)^2\right]. \tag{18}$$

Equivalence of these two asymptotic variance characterizations For the simple case of scalar $h()$ function of x, the optimization problem (18) can be solved in closed form. Note that

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

has a solution equal to

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

so that

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

Note that

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

and

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

Then,

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

and

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

from which we obtain

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

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

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

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

   and

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

2. Compute

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

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

   and

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

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 $E\left[g\left(z_i,\beta_*,\theta_*\right)\right]=0$, where

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

and $h\left(\cdot,\theta\right)=p^K\left(\cdot\right)'\theta$. Here, both β and θ are finite dimensional parameters such that $\dim\left(g\right)=\dim\left(\beta\right)+\dim\left(\theta\right)$. A consistent estimator of $\widehat{\alpha}=\left(\widehat{\beta},\widehat{\theta}'\right)'$ is given by the usual formula (which is (5)):

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

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

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

and

$$g\left(z_i,\beta,\theta\right)=\left[\begin{array}{c} -\frac{\partial\ell(z_i,h(x_i,\theta))}{\partial h}p^K\left(x_i\right) \\ m\left(z_i,\beta,h\left(x_i,\theta\right)\right)\end{array}\right]$$

Using the notation in Lemma 1, we let

$$\begin{aligned}\mathbb{A}&=\frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c} 0 \\ q_i'\end{array}\right]=\left[\begin{array}{c} 0 \\ \overline{q}'\end{array}\right]\\ \mathbb{B}&=\frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c} R_i \\ Q_i'\end{array}\right]=\left[\begin{array}{c} \overline{R} \\ \overline{Q}'\end{array}\right]\\ \mathbb{F}&=\frac{1}{n}\sum_{i=1}^{n}g\left(z_i,\widehat{\beta},\widehat{\theta}\right)g\left(z_i,\widehat{\beta},\widehat{\theta}\right)'\end{aligned}$$

and

$$\Upsilon^{-1}=\left[\begin{array}{cc} \overline{R}^{-1} & 0 \\ 0 & I_d\end{array}\right]$$

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

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

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

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

We also have

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

and

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

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

Notes

[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, e.g., Gonçalves and White (2005) for a recent discussion).

[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. On the other hand, the semiparametric formula in principle addresses nonparametric first step sieve estimation with 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.

[7]This formulation assumes exact identification, i.e., $\dim(\varphi) = \dim(\theta)$ and $\dim(\psi) = \dim(\beta)$. We consider an overidentified situation in Section 5.3.

[8]See Wooldridge (2002, Chapter 12.3).

[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.

[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, e.g., Chen (2007, remark 4.2), Chen, Fan and Tsyrennikov (2006). For most other results in this paper, we do not provide any rigorous asymptotic theory, which is already done in the existing literature.

[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.

[13]See Chen, Linton and van Keilegom (2003).

[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.

[16]Note that this doesn't 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

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

[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.

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

References

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

[2] Aguirregabiria, V. and P. Mira (2007): "Sequential Estimation of Dynamic Discrete Games," *Econometrica* 75, pp. 1–53.

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

[4] Andrews, D. (1994) "Asymptotics for Semi-parametric Econometric Models via Stochastic Equicontinuity", *Econometrica* 62, pp. 43-72.

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

[6] Bajari, P., V. Chernozhukov, and H. Hong (2005): "Semiparametric Estimation of a Dynamic Game of Incomplete Information," unpublished working paper, Duke University.

[7] Bajari, P., V. Chernozhukov, H. Hong, D. Nekipelov (2010): "Nonparametric and Semiparametric Analysis of a Dynamic Game Model," unpublished working paper.

[8] Barron, A. and C. Sheu (1991): "Approximation of density functions by sequence of exponential families", *The Annals of Statistics* 19, pp. 1347-1369.

[9] Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," Journal of Econometrics 32, pp. 189-218.

[10] Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models", chapter 76 in *The Handbook of Econometrics*, Vol. 6B, eds. James J. Heckman and Edward E. Lleamer, North-Holland.

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

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

[13] Chen, X., O. Linton and I. van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function is not Smooth", *Econometrica* 71, pp. 1591-1608.

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

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

[16] Ellickson, P. and Misra, S. (2008): "Supermarket Pricing Strategies," *Marketing Science* 27, pp. 811-828

[17] Efromovich, S. (1999): *Nonparametric Curve Estimation*, Springer Series in Statistics: New York.

[18] Gallant, A.R. and D. Nychka (1987): "Semi-non-parametric maximum likelihood estimation", *Econometrica* 55, pp. 363-390.

[19] Genovese, C. and L. Wasserman (2000): "Rates of Convergence for the Gaussian Mixture Sieve", *The Annals of Statistics* 28, pp. 1105-1127.

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

[21] Hahn, J. (1994): "The Efficiency Bound of the Mixed Proportional Hazard Model," *Review of Economic Studies* 61, pp. 607-629.

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

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

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

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

[26] Ichimura, H. and S. Lee (2010): "Characterization of the Asymptotic Distribution of Semiparametric M-Estimators", forthcoming in *Journal of Econometrics*.

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

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

[29] Macieria, J. (2008): "Extending the Frontier: A Structural Model of Investment and Technological Competition in the Supercomputer Industry," mimeo, Virginia Tech

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

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

[32] Newey, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, pp. 1349 – 1382.

[33] Newey, W.K. and D. F. McFadden (1994): "Large sample estimation and hypothesis testing", in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.

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

[35] Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review* 25, pp. 221-247.

[36] Pakes, A. and G.S. Olley (1995): "A Limit Theorem for a Smooth Class of Semiparametric Estimators," *Journal of Econometrics* 65, pp. 295–332.

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

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

[39] Robinson, P. (1988) "Root-N-Consistent Semiparametric Regression", *Econometrica,* 56, pp. 931-954.

[40] Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66, pp. 688-701.

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

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

[43] Shen, X. (1997) "On Methods of Sieves and Penalization", *The Annals of Statistics* 25, pp. 2555-2591.

[44] Snider, C. (2008): "Predatory Incentives and Predation Policy: The American Airlines Case," unpublished working paper, Minnesota.

[45] Stone, C.J. (1990): "Large-sample inference for log-spline models", *The Annals of Statistics* 18, pp. 717-741.

[46] Sweeting, A. (2007): "Dynamic Product Repositioning in Differentiated Product Industries: The Case of Format Switching in the Commercial Radio Industry," unpublished working paper, Duke.

[47] Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data,* Cambridge: MIT Press.

A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators: Supplementary Appendix

Daniel Ackerberg, Xiaohong Chen, and Jinyong Hahn

A Understanding Newey's (1994) Asymptotic Variance Formula

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

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

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

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

$$-\left(M_\beta\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(z_i, \beta_*, h_*) + E\left[D(z)\right| x = x_i]\left(y_i - h_*(x_i)\right)\}\right) \tag{19}$$

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

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

$$\left(\widehat{M_\beta}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(m\left(z_i, \widehat{\beta}, \widehat{h}\right) + \widehat{E}\left[D(z)\right| x_i]\left(y_i - \widehat{h}(x_i)\right)\right)^2\left(\widehat{M_\beta'}\right)^{-1} \tag{20}$$

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

In order to prove (19), it suffices to characterize the asymptotic distribution of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(z_i, \beta_*, \widehat{h}\right)$. (This is because we have

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

under regularity conditions.)

Newey (1994) basically writes

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

and devotes the rest of his paper characterizing the adjustment $a\left(z_i\right)$ to the influence function. We follow Newey's (1994) notation for convenience of readers. From Newey (p. 1360), we can see that, for $D\left(z, h\right) = D\left(z\right) h\left(v\right)$ with $D\left(z\right) = \left.\partial m\left(z, h\left(v\right)\right)/\partial h\right|_{h=h_*}$, we have his equation (4.1) satisfied. As is discussed on the same page, we now assume that $h_*(x) = E\left[\left. y\right| x\right]$ for some y and x. Now we follow his equation (4.4), and see if we can find

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

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

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(z_i, \beta_*, \widehat{h}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m\left(z_i, \beta_*, h_*\right) + E\left[\left. D\left(z\right)\right| x = x_i\right]\left(y_i - h_*(x_i)\right) + o_p\left(1\right).\right. \tag{21}$$

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

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

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

where

$$\begin{aligned}&\widehat{E}^*\left[D\left(z_i\right)\left|p^K\left(x_i\right)\right.\right]\left(y_i - h\left(x_i, \widehat{\theta}\right)\right)\\ &\equiv p^K\left(x_i\right)'\left(P'P\right)^{-1}\left(\sum_{i=1}^{n} p^K\left(x_i\right) D\left(z_i\right)\right)\left(y_i - p^K\left(x_i\right)'\left(P'P\right)^{-1}\left(\sum_{i=1}^{n} p^K\left(x_i\right) y_i\right)\right)\end{aligned} \tag{22}$$

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

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

With this in mind, he will conclude that

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

He will then proceed and conclude that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(z_i, \beta_*, h\left(x_i, \widehat{\theta}\right)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i, \beta_*, h(x_i,\theta_*)) \\
&\quad + \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial m(z_i,\beta_*,h(x_i,\theta_*))}{\partial h}\frac{\partial h(x_i,\theta_*)}{\partial \theta'}\right)\left(\widehat{\theta} - \theta_*\right) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i, \beta_*, h(x_i,\theta_*)) \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^{n} D(z_i)\frac{\partial h(x_i,\theta_*)}{\partial \theta'}\right)\sqrt{n}\left(\widehat{\theta} - \theta_*\right) + o_p(1)
\end{aligned} \tag{23}$$

Now, in his mind, he will think that

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{i=1}^{n} D(z_i)\frac{\partial h(x_i,\theta_*)}{\partial \theta'}\right)\sqrt{n}\left(\widehat{\theta} - \theta_*\right) \\
&= E\left[D(z_i)\frac{\partial h(x_i,\theta_*)}{\partial \theta'}\right]\sqrt{n}\left(\widehat{\theta} - \theta_*\right) + o_p(1) \\
&= E\left[D(z_i)\frac{\partial h(x_i,\theta_*)}{\partial \theta'}\right]\left(E\left[\frac{\partial h(x,\theta_*)}{\partial \theta}\frac{\partial h(x,\theta_*)}{\partial \theta'}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial h(x_i,\theta_*)}{\partial \theta}(y_i - h(x_i,\theta_*))\right) + o_p(1)
\end{aligned} \tag{24}$$

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

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

and the fitted value is equal to

$$\begin{aligned}
&\frac{\partial h(x_i,\theta_*)}{\partial \theta'}\left(E\left[\frac{\partial h(x,\theta_*)}{\partial \theta}\frac{\partial h(x,\theta_*)}{\partial \theta'}\right]\right)^{-1} E\left[\frac{\partial h(x_i,\theta_*)}{\partial \theta}D(z_i)\right] \\
&= E\left[D(z_i)\frac{\partial h(x_i,\theta_*)}{\partial \theta'}\right]\left(E\left[\frac{\partial h(x,\theta_*)}{\partial \theta}\frac{\partial h(x,\theta_*)}{\partial \theta'}\right]\right)^{-1}\frac{\partial h(x_i,\theta_*)}{\partial \theta}
\end{aligned} \tag{25}$$

So, let's write

$$E^{*}\left[D\left(z_{i}\right)\left|\frac{\partial h\left(x_{i},\theta_{*}\right)}{\partial\theta}\right.\right]=E\left[D\left(z_{i}\right)\frac{\partial h\left(x_{i},\theta_{*}\right)}{\partial\theta'}\right]\left(E\left[\frac{\partial h\left(x,\theta_{*}\right)}{\partial\theta}\frac{\partial h\left(x,\theta_{*}\right)}{\partial\theta'}\right]\right)^{-1}\frac{\partial h\left(x_{i},\theta_{*}\right)}{\partial\theta} \tag{26}$$

where E^{*} denotes the best linear predictor. Combining (24) – (26), we can then see the practitioner's thought process would lead to the expression

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}m\left(z_{i},\beta_{*},h\left(x_{i},\widehat{\theta}\right)\right)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m\left(z_{i},\beta_{*},h\left(x_{i},\theta_{*}\right)\right)+E^{*}\left[D\left(z_{i}\right)\left|\frac{\partial h\left(x_{i},\theta_{*}\right)}{\partial\theta'}\right.\right]\left(y_{i}-h\left(x_{i},\theta_{*}\right)\right)\right)+o_{p}\left(1\right)\end{aligned} \tag{27}$$

We now compare (21) with (27). It is easy to see that, except for $E\left[D\left(z\right)\right|x=x_{i}]$ in (21) and $E^{*}\left[D\left(z_{i}\right)\left|\frac{\partial h(x_{i},\theta_{*})}{\partial\theta}\right.\right]$ in (27), the formulae that the practitioner uses for asymptotic variance calculation are identical. Obviously, we need to ask the question when $E^{*}\left[D\left(z_{i}\right)\left|\frac{\partial h(x_{i},\theta_{*})}{\partial\theta}\right.\right]$ can be interpreted to be an approximation of $E\left[D\left(z\right)\right|x=x_{i}]$. This is easy. Suppose that $h\left(x_{i},\theta\right)=p^{K}\left(x_{i}\right)'\theta$. Then

$$\frac{\partial h\left(x_{i},\theta_{*}\right)}{\partial\theta}=p^{K}\left(x_{i}\right)$$

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

A consistent estimator for the "parametric" asymptotic variance is equal to

$$\left(\widehat{M}_{\beta}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(m\left(z_{i},\widehat{\beta},h\left(x_{i},\widehat{\theta}\right)\right)+\widehat{E}^{*}\left[D\left(z_{i}\right)\left|p^{K}\left(x_{i}\right)\right.\right]\left(y_{i}-h\left(x_{i},\widehat{\theta}\right)\right)\right)^{2}\left(\widehat{M}_{\beta}'\right)^{-1}. \tag{28}$$

Numerical equivalence when $\widehat{h}$ is a sieve estimator When will Newey's estimator (20) of the semiparametric asymptotic variance (19) be numerically identical to the practitioner's parametric variance estimator (28)? If we are to use a sieve estimator with basis $p^{K}\left(x_{i}\right)$ to compute $\widehat{h}(x_{i})=\widehat{E}\left[y\right|x=x_{i}]$ and $\widehat{E}\left[D\left(z\right)\right|x_{i}]$ in Newey's (20), it can be easily seen that

$$\begin{aligned}&\widehat{E}\left[D\left(z\right)\right|x_{i}]\left(y_{i}-\widehat{E}\left[y\right|x=x_{i}]\right)\\&=p^{K}\left(x_{i}\right)'\left(P'P\right)^{-1}\left(\sum_{i=1}^{n}p^{K}\left(x_{i}\right)D\left(z_{i}\right)\right)\left(y_{i}-p^{K}\left(x_{i}\right)'\left(P'P\right)^{-1}\left(\sum_{i=1}^{n}p^{K}\left(x_{i}\right)y_{i}\right)\right),\end{aligned} \tag{29}$$

which is numerically identical to (22). It follows that Newey's estimator (20) is numerically identical to (28) when a sieve least squares estimator is used for $\widehat{h}$ and $\widehat{E}\left[D\left(z\right)\right|x_{i}]$. (In fact, Murphy and Topel's (1985) estimator is identical to (28).)

B Discussion of Ai and Chen (2007)

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

$$E\left[y_i - h_*(x_i) | x_i \right] = 0, \quad E\left[m\left(z_i, \beta_*, h_*(x_i)\right) \right] = 0. \tag{30}$$

Ai and Chen' (2007) modified sieve minimum distance (SMD) estimator[20] for $\alpha_* = (\beta_*, h_*)$ boils down to

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

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

1. Compute w^* to solve

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

2. Calculate

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

   and

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

3. Calculate

$$V = \left(E\left[\Delta_{w^*}(z_i)' \Delta_{w^*}(z_i) \right] \right)^{-1} \operatorname{Var}\left(\Delta_{w^*}(z_i)' \rho(z_i, \alpha_*) \right) \left(E\left[\Delta_{w^*}(z_i)' \Delta_{w^*}(z_i) \right] \right)^{-1}. \tag{31}$$

For this simple example model (30), it can be shown that the solutions in the above Steps 1 - 3 are

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

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

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

and

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

where $D(z) = \left.\partial m(z, \beta_*, h(x))/\partial h\right|_{h=h_*}$. In particular, we see that Ai and Chen's asymptotic variance V is identical to Newey's (1994) asymptotic variance (19) for this example model (30). We note that analytic characterization of $w^*(\cdot)$ hence population asymptotic variance V is not always easy for general semiparametric models considered in Ai and Chen (2007). Their sieve estimator of the asymptotic variance V simply uses a sample counterpart of the population minimization problem to bypass such a difficulty.