

# Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables

Chunrong Ai<sup>a</sup>, Xiaohong Chen<sup>b,1</sup>

<sup>a</sup>*Department of Economics, University of Florida, Gainesville, FL 32611, USA*

<sup>b</sup>*Department of Economics, New York University, 269 Mercer Street, New York, NY 10003, USA*

First version: March 2003, Final revision: January 2006

---

## Abstract

Newey and Powell (2003) and Ai and Chen (2003) propose sieve minimum distance (SMD) estimation of both finite dimensional parameter ( $\theta$ ) and infinite dimensional parameter ( $h$ ) that are identified through a conditional moment restriction model, in which  $h$  could depend on endogenous variables. This paper modifies their SMD procedure to allow for different conditioning variables to be used in different equations, and derives the asymptotic properties when the model may be *misspecified*. Under low-level sufficient conditions, we show that: (i) the modified SMD estimators of both  $\theta$  and  $h$  converge to some pseudo-true values in probability; (ii) the SMD estimators of smooth functionals, including the  $\theta$  estimator and the average derivative estimator, are asymptotically normally distributed; and (iii) the estimators for the asymptotic covariances of the SMD estimators of smooth functionals are consistent and easy to compute. These results allow for asymptotically valid tests of various hypotheses on the smooth functionals regardless of whether the semiparametric model is correctly specified or not.

*JEL Classification:* C14; C22

*Keywords:* Misspecification; Sieve minimum distance; Conditional moment models with different conditioning sets; Nonparametric endogeneity; Weighted average derivatives

---

---

<sup>1</sup>Corresponding author. Tel.: +1 212 998 8970; Fax: +1 212 995 4186.

*E-mail addresses:* chunrong.ai@cba.ufl.edu (C. Ai), xiaohong.chen@nyu.edu (X. Chen).

# 1 Introduction

Newey and Powell (2003) and Ai and Chen (2003) propose sieve minimum distance (hereafter SMD) estimation of both the finite-dimensional parameter ( $\theta_o$ ) and the infinite-dimensional parameter ( $h_o$ ) that are identified through the conditional moment restriction model:  $E[\rho(Y, X; \theta_o, h_o(\cdot))|X] = 0$ , in which  $\rho(\cdot) \equiv (\rho_1(\cdot), \dots, \rho_J(\cdot))'$  is a  $J \times 1$  vector of mappings known up to the parameter  $\alpha_o = (\theta_o, h_o)$ , and the unknown functions  $h_o(\cdot)$  may depend on endogenous variables. Under some sufficient conditions, the consistency of the SMD estimators of  $(\theta_o, h_o)$  is proved in Newey and Powell (2003) and the asymptotic normality and the semiparametric efficiency of the SMD estimator of  $\theta_o$  are established in Ai and Chen (2003).

In this paper we modify their SMD procedure and extend their results in two directions. First, we allow different conditioning variables to be used in different equations. Second and perhaps more importantly, we derive the asymptotic properties of the modified SMD estimators when the conditional moment restriction model could be *misspecified*. Specifically, let  $\mathcal{A} = \Theta \times \mathcal{H}$ , where  $\Theta$  denotes a compact finite dimensional parameter space and  $\mathcal{H}$  an infinite dimensional parameter space. Let  $\alpha = (\theta, h) \in \mathcal{A}$  denote the unknown parameter. Let  $Z = (Y', X')' \in \mathcal{Z}$  denote all the random variables, and  $X_j \in \mathcal{X}_j$  denote the conditioning variables used in the  $j^{\text{th}}$  equation  $\rho_j(Z, \theta, h)$  for  $j = 1, \dots, J$ . Here  $X_j$  is either equal to a subset of  $X$  or a degenerate random variable; and if  $X_j$  is degenerate, the conditional expectation  $E\{\rho_j(Z, \theta, h)|X_j\}$  is the same as the unconditional expectation  $E\{\rho_j(Z, \theta, h)\}$ . If there are some  $\alpha_o = (\theta_o, h_o) \in \mathcal{A}$  such that  $E[\rho_j(Z, \alpha_o)|X_j] = 0$  for all  $j = 1, 2, \dots, J$ , we say the semiparametric conditional moment restriction model is correctly specified. For simplicity, in this paper we assume that, when the model is correctly specified, there is a unique  $\alpha_o = (\theta_o, h_o) \in \mathcal{A}$  satisfying the semiparametric conditional moment restriction:

$$E[\rho_j(Z, \alpha_o)|X_j] = 0, \quad j = 1, 2, \dots, J. \quad (1)$$

(We note that in the correctly specified case  $X_j$  are only required to be exogenous for the  $j$ -th equation but they could enter as endogenous variables to other equations). If

$$E\left[\sum_{j=1}^J \{E[\rho_j(Z, \alpha)|X_j]\}^2\right] > 0 \quad \text{for all } \alpha \in \mathcal{A},$$

we say the semiparametric conditional moment restriction model (1) is incorrectly specified. (This can happen when some of the moment functions  $\rho_j(Z, \alpha)$ ,  $j = 1, \dots, J$  are misspecified, i.e., the conditioning variables  $X_j$  are endogenous for the  $j^{\text{th}}$  equation.) Let  $m(X, \alpha) \equiv (m_1(X_1, \alpha), \dots, m_J(X_J, \alpha))'$  with  $m_j(X_j, \alpha) \equiv E\{\rho_j(Z, \alpha)|X_j\}$  and  $\Sigma(X)$  be a  $J \times J$ - positive definite weighting matrix. We assume that  $\alpha_* = (\theta_*, h_*) \in \mathcal{A}$  is the unique solution to  $\inf_{\alpha \in \mathcal{A}} E\{m(X, \alpha)' \Sigma(X)^{-1} m(X, \alpha)\}$ . Clearly  $m(X, \alpha_*) = 0$  if and only if the semiparametric conditional moment restriction model (1) is correctly specified, and in this case  $\alpha_* = \alpha_o$ .

In this paper we present a modified SMD estimator  $\hat{\alpha} = (\hat{\theta}, \hat{h})$  for  $\alpha_* = (\theta_*, h_*)$ , and derive the asymptotic properties of  $\hat{\alpha}$  without assuming the conditional moment restriction model (1) is correctly specified. Under low-level sufficient conditions, we show that: (i)  $\hat{\alpha}$  converges to the pseudo-true value  $\alpha_*$  in probability; (ii) the SMD estimators of smooth functionals of  $\alpha_*$ , including the estimators of  $\theta_*$  and of the average derivative of  $h_*$ , are asymptotically normally distributed; and (iii) the estimators for the asymptotic covariances of the SMD estimators of smooth functionals are consistent and easy to compute. These results allow us to perform asymptotically valid tests of various hypotheses on the smooth functionals of  $\alpha_*$  regardless of whether model (1) is correctly specified or not.

If the semiparametric conditional moment restriction (1) is satisfied, then  $\alpha_* = \alpha_o$  and our results in this case extend those of Newey and Powell (2003) and Ai and Chen (2003) from the model  $E[\rho(Z, \alpha_o)|X] = 0$  to the conditional moment restriction model with different conditioning variables. This extension is important for at least two reasons. First, if we interpret each  $\rho_j(Z, \alpha_o)$  as equation and  $X_j$  as the instrumental variables for that equation, then the model (1) is a system of equations with different instruments for different equations. There are many applications where different equations may require different set of instruments. The semiparametric hedonic price system where some explanatory variables in some equations are correlated with the errors in other equations is one such example (see e.g., Ekeland, Heckman and Nesheim (2004) and Heckman, Matzkin and Nesheim (2004)). The simultaneous equations model with measurement error in some exogenous variables, or some omitted variables correlated with what would otherwise be exogenous variables is another example (see e.g., Hausman (1977), Wooldridge (1996) and Lewbel (2005)). Semiparametric panel data models where some variables that are uncorrelated with the error in a given time period are correlated with the errors in previous periods is a third example (see e.g., Baltagi and Li (2003)). The triangular simultaneous equations system studied in Newey, Powell and Vella (1999), the panel data attrition with refreshment sample model studied in Bhattacharya (2005), and the dynamic panel sample selection model studied in Gayle and Viauroux (2005) also fit the general framework (1).<sup>2</sup>

The second reason that our extension is important is that the semiparametric conditional moment restriction model (1) provides a convenient framework for deriving the asymptotic distribution of the plug-in SMD estimator of a smooth functional defined via expectation, and for computing a consistent estimator of the asymptotic covariance of the plug-in estimator. (See Section 2 for further discussion). Newey (1984), Newey and McFadden (1994) and others present a general formula for computing the consistent asymptotic covariance matrix of the plug-in estimator in a parametric

---

<sup>2</sup>Although the semiparametric conditional moment restriction model (1) includes many applications, due to the lack of space, we shall not provide detailed studies of any specific applications in this paper. Interested readers could find more discussions on semiparametric dynamic panel data models from Ai and Chen (2005), in which we consider semiparametric efficient estimation of smooth functionals when the model (1) is correctly specified.

moment restriction framework. We extend their results to the semiparametric conditional moment setting (1), where the unknown functions  $h(\cdot)$  may depend on endogenous variables and where the model (1) may not be correctly specified.

The asymptotic properties of the extremum estimator of  $\theta$  for possibly misspecified parametric models have been widely studied in the literature; see e.g., White (1982, 1994), Hansen and Jagannathan (1997) and Hall and Inoue (2003). The asymptotic properties of the estimators of  $\alpha = (\theta, h)$  for possibly misspecified semi/non-parametric models, however, have not attracted much attention from researchers. A notable exception is Stone (1985), who considers estimation of additive regression model without imposing the correct specification of the conditional mean restriction  $E[Y|W_1, \dots, W_q] = \theta_o + \sum_{j=1}^q h_{oj}(W_j)$ . Instead Stone (1985) obtains convergence rates of his spline estimators of  $h_{*j}$ ,  $j = 1, \dots, q$  that are the best approximation to  $E[Y|W_1, \dots, W_q]$  in the mean squared error sense:

$$(\theta_*, h_{*1}, \dots, h_{*q}) = \arg_{\theta, E[h_j(W_j)] = 0, E[h_j(W_j)]^2 < \infty} \inf E \left[ \left\{ E[Y|W_1, \dots, W_q] - \theta - \sum_{j=1}^q h_j(W_j) \right\}^2 \right].$$

Our results apply to Stone's (1985) model as well as other models involving semi-nonparametric dimension-reduction specifications that are convenient but might not be correct. In particular, our general form of the asymptotic covariance matrix and its consistent estimator – which are valid whether or not the model (1) is correctly specified – permit researchers to conduct robust inference on the structural parameters of interest even when the first stage semi-nonparametric specifications are incorrect. For example, even when the propensity score function is estimated using a wrong additive (or single index or partially linear) regression specification, applied researchers can still use our results to consistently estimate the standard error for the plug-in estimator of the average treatment effect (or average treatment effect for the treated) parameter.

General theory on the  $\sqrt{n}$ -asymptotic normality of the plug-in estimators of smooth functionals has already been presented in many papers under various assumptions and for various models.<sup>3</sup> For example, Newey and McFadden (1994), Newey (1994), Andrews (1994) and Pakes and Olley (1995) establish their results under the assumption that there is a nonparametric estimator  $\hat{h}_n$  which converges to  $h_o$  at a rate of  $o_p(n^{-1/4})$  under the supremum norm. This convergence rate may not be obtainable when  $h_o$  depends on endogenous variables. Although the results of Shen (1997) and Chen and Shen (1998) do not require the convergence rate of  $o_p(n^{-1/4})$  under the supremum norm, they are applicable only to semiparametric models where  $h_o$  is estimated via the M-estimation method and hence rule out the important nonparametric Instrumental Variables (IV) regression example. The results of Chen, Linton and van Keilegom (2003) allow for nonparametric

---

<sup>3</sup>There are more papers on  $\sqrt{n}$ - asymptotic normality of particular smooth functionals in specific semiparametric regression and semiparametric MLE type of models. The earlier ones include Robinson (1988), Powell, Stock and Stoker (1989), Ichimura (1993) and Klein and Spady (1993). See Powell (1994) and Horowitz (1998) for reviews.

component to depend on endogenous variables but they do not provide any estimator for  $h_o$ ; without specifying how to compute  $\hat{h}_n$  all they could suggest is to bootstrap the limiting distribution for the plug-in estimator of a smooth functional. In a companion paper (Ai and Chen, 2005), we study the semiparametric efficient estimation of smooth functionals of  $\alpha_o$  that satisfies the conditional moment model (1). However, to the best of our knowledge, none of the published general theory papers have considered the estimation of  $\alpha = (\theta, h)$  and its smooth functionals when the conditional moment model (1) is misspecified.

The rest of the paper is organized as follows. Section 2 first introduces a modified SMD estimator for possibly misspecified semiparametric conditional moment model (1). It then describes how the plug-in estimator of any smooth functional can be re-interpreted as the SMD estimator of a larger model. We also present two illustrative yet non-trivial examples: a weighted average derivative estimate of a possibly misspecified nonparametric additive Least Squares (LS) regression, and a weighted average derivative estimate of a possibly misspecified nonparametric IV regression. Section 3 provides low-level sufficient conditions for consistency and convergence rate of the SMD estimator to the pseudo-true value  $\alpha_*$ ; the conditions are slightly weaker than those in Ai and Chen (2003). Section 4 derives the asymptotic normality of the SMD estimator of a smooth functional of  $\alpha_*$ . It also discusses the complication of the asymptotic covariance matrix due to the misspecification and the nonlinearity of the conditional moment model (1). We then provide primitive sufficient conditions for the root- $n$  normality of the average derivative estimates for the two examples introduced in Section 2. Section 5 provides a consistent estimator of the asymptotic variance of the SMD estimator of a smooth functional and Section 6 briefly concludes. All the mathematical proofs and technical lemmas are presented in the Appendix.

## 2 The Modified SMD Procedure and Examples

Let  $\{z_i = (y'_i, x'_i)'\}_{i=1}^n$  be a random sample from the distribution of  $Z = (Y', X')' \in \mathcal{Z} = \mathcal{Y} \times \mathcal{X}$ . Recall that  $\alpha_* \in \mathcal{A}$  is the pseudo-true value defined as  $\alpha_* = \arg \inf_{\alpha \in \mathcal{A}} E \{m(X, \alpha)' \Sigma(X)^{-1} m(X, \alpha)\}$ . When the semiparametric conditional moment restriction model (1) is correctly specified,  $m(X, \alpha_*) = 0$  and  $\alpha_* = \alpha_o$ , which depends only on the true data generating process (DGP). When the model (1) is misspecified,  $m(X, \alpha_*) \neq 0$  holds with positive probability, and in this case, the pseudo-true value  $\alpha_*$  depends on both the underlying true DGP and the choice of the weighting matrix  $\Sigma(X)$ .

### 2.1 The modified SMD procedure

We now describe the SMD procedure for estimating  $\alpha_*$ . The SMD procedure requires a consistent estimator of the conditional mean  $m_j(X_j, \alpha) = E\{\rho_j(Z, \alpha)|X_j\}$ ,  $j = 1, \dots, J$ . Any nonparametric LS regression estimator (such as kernel, local linear regression and sieve LS) can be used here; and for simplicity we present a series LS estimator of  $m_j(X_j, \alpha)$ . Let  $\{p_{j1}(X_j), p_{j2}(X_j), \dots\}$

denote a sequence of known basis functions that can approximate any square integrable real-valued function of  $X_j$  arbitrarily well. Familiar basis functions include splines, wavelets, Hermite polynomials, power series and Fourier series. Denote  $p_j^{k_{jn}}(X_j) = (p_{j1}(X_j), \dots, p_{jk_{jn}}(X_j))'$  and  $P_j = (p_j^{k_{jn}}(x_{j1}), \dots, p_j^{k_{jn}}(x_{jn}))'$ . The series LS estimator of  $m_j(X_j, \alpha)$  is given by:

$$\hat{m}_j(X_j, \alpha) = p_j^{k_{jn}}(X_j)'(P_j'P_j)^{-1} \sum_{i=1}^n p_j^{k_{jn}}(x_{ji})\rho_j(z_i, \alpha).$$

If  $X_j$  is a degenerate random variable then  $k_{jn} = 1$ ,  $p_j^{k_{jn}}(X_j) = 1$  and  $\hat{m}_j(X_j, \alpha) = \hat{m}_j(\alpha) \equiv \frac{1}{n} \sum_{i=1}^n \rho_j(z_i, \alpha)$ . Denote  $\hat{m}(X, \alpha) = (\hat{m}_1(X_1, \alpha), \dots, \hat{m}_J(X_J, \alpha))'$ . Let  $\hat{\Sigma}(X)$  denote a consistent estimator of the positive definite weighting matrix  $\Sigma(X)$ . The SMD estimator of  $\alpha_*$  is defined as:

$$\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n) = \arg \min_{\alpha \in \mathcal{A}_{k(n)}} \frac{1}{n} \sum_{i=1}^n \hat{m}(x_i, \alpha)' \{\hat{\Sigma}(x_i)\}^{-1} \hat{m}(x_i, \alpha) \quad \text{with } \mathcal{A}_{k(n)} \equiv \Theta \times \mathcal{H}_{k(n)},$$

where  $\{\mathcal{H}_k : k = 1, 2, \dots\}$  is a sequence of non-decreasing approximation spaces (sieves) such that  $\cup_{k=1}^{\infty} \mathcal{H}_k$  is dense in the infinite dimensional space  $\mathcal{H}$ . Sometimes we use the notation  $\mathcal{A}_n \equiv \Theta \times \mathcal{H}_n$  to mean  $\mathcal{A}_{k(n)} \equiv \Theta \times \mathcal{H}_{k(n)}$ . In economics applications, the sieve spaces are usually compact finite dimensional parameter spaces whose dimension (or complexity) increases with sample size. Popular sieves are linear sieves, also called series, which are linear combinations of finite number of known basis functions such as splines, wavelets, power series, Hermite polynomials and Fourier series.

When the semiparametric conditional moment restriction (1) is correctly specified, we can establish the convergence rate of  $\hat{\alpha}_n$  to the true value  $\alpha_o$  and the limiting distribution of  $\hat{\theta}_n$  by modifying the results in Ai and Chen (2003) slightly. However, when the model (1) is incorrectly specified, the convergence rate of  $\hat{\alpha}_n$  to the pseudo-true value  $\alpha_*$  and the limiting distribution of  $\hat{\theta}_n$  are generally affected by the convergence of  $\hat{\Sigma}(X)$  to  $\Sigma(X)$  in a complicated manner. This is the case even for the parametric (e.g.,  $\alpha = \theta$ ), misspecified and overidentified unconditional moment models; for such models Hall and Inoue (2003) show that the limiting distribution of  $\hat{\theta}_n$  depends on the asymptotic distribution of the estimated weighting matrix. In our more general misspecified semiparametric framework, the asymptotic distribution of the estimated weighting matrix is difficult to derive, hence its effect on the asymptotic distribution of the SMD estimator  $\hat{\theta}_n$  is hard to quantify. To keep our derivation simple, in this paper we restrict our attention to a known weighting matrix  $\Sigma(X)$ . Furthermore we assume  $\Sigma(X) = I$  (the identity weighting). Hence, the pseudo-true value  $\alpha_*$  is defined as

$$\alpha_* = \arg \inf_{\alpha \in \mathcal{A}} E \{m(X, \alpha)'m(X, \alpha)\}, \quad (2)$$

and its SMD estimator is given by

$$\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n) = \arg \min_{\alpha \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^n \hat{m}(x_i, \alpha)' \hat{m}(x_i, \alpha). \quad (3)$$

Let  $\mathcal{J}_{ex}$  consist of all the indices  $j \in \{1, \dots, J\}$  such that  $m_j(X_j, \alpha) - m_j(X_j, \alpha_*) = \rho_j(Z, \alpha) - \rho_j(Z, \alpha_*)$  for all  $\alpha \in \mathcal{A}$  with probability one. Let  $\mathcal{J}_{en} \equiv \{1, 2, \dots, J\} \setminus \mathcal{J}_{ex}$ . Denote  $\mathcal{J}_{1en}$  as a subset of  $\mathcal{J}_{en}$  consisting of all the  $j$ 's whose conditioning variable  $X_j$  are not degenerate. Then  $\mathcal{J}_{2en} \equiv \mathcal{J}_{en} \setminus \mathcal{J}_{1en}$  denote indices of those equations where  $X_j$  for  $j \in \mathcal{J}_{2en}$  are degenerate, and denote  $m_j(X_j, \alpha) = m_j(\alpha) \equiv E[\rho_j(Z, \alpha)]$  for  $j \in \mathcal{J}_{2en}$ . It is easy to show that the unique solution  $\alpha_*$  to (2) also satisfies

$$\alpha_* = \arg \inf_{\alpha \in \mathcal{A}} \left\{ \sum_{j \in \mathcal{J}_{ex}} E[\rho_j(Z, \alpha)^2] + \sum_{j \in \mathcal{J}_{1en}} E[m_j(X_j, \alpha)^2] + \sum_{j \in \mathcal{J}_{2en}} [m_j(\alpha)]^2 \right\}. \quad (4)$$

The groupings of the equations allow us to handle the sieve (nonlinear) LS regression ( $\mathcal{J}_{ex}$  group) and the SMD estimation ( $\mathcal{J}_{1en}$  and  $\mathcal{J}_{2en}$ ) in a unified framework. The modified SMD estimator is now defined as

$$\hat{\alpha}_n = \arg \min_{\alpha \in \mathcal{A}_n} \left\{ \sum_{j \in \mathcal{J}_{ex}} \frac{1}{n} \sum_{i=1}^n \rho_j(z_i, \alpha)^2 + \sum_{j \in \mathcal{J}_{1en}} \frac{1}{n} \sum_{i=1}^n \hat{m}_j(x_{ji}, \alpha)^2 + \sum_{j \in \mathcal{J}_{2en}} \hat{m}_j(\alpha)^2 \right\}. \quad (5)$$

This modified procedure is simpler because it does not require the estimation of  $m_j(X_j, \alpha) = E\{\rho_j(Z, \alpha) | X_j\}$  for  $j \in \mathcal{J}_{ex}$ , and hence the sufficient conditions for convergence of  $\hat{m}_j(X_j, \alpha)$  to  $m_j(X_j, \alpha)$ ,  $j \in \mathcal{J}_{ex}$ , are not needed. Moreover, in this paper we show that the modified SMD estimator (5) and the original SMD estimator (3) have the same large sample properties.

## 2.2 Plug-in estimation

We now illustrate how the plug-in estimation of a smooth functional defined via expectation can be reformulated as a special case of our general SMD estimation problem. Suppose that  $(\theta_{*1}, h_*)$  is the unique solution to

$$\inf_{\Theta_1 \times \mathcal{H}} E \left[ \sum_{j=1}^{J-1} \{E[\rho_j(Z, \theta_1, h) | X_j]\}^2 \right]. \quad (6)$$

Suppose that we are interested in estimating a smooth functional  $\theta_2$  defined implicitly as:

$$E[\rho^{\theta_2}(Z; \theta_{*1}, \theta_{*2}, h_*)] = 0, \quad (7)$$

where  $\frac{dE[\rho^{\theta_2}(Z; \theta_{*1}, \theta_2, h_*(\cdot))]}{d\theta_2} |_{\theta_{*2}}$  has full rank  $d_{\theta_2} = \dim(\theta_2) = \dim(\rho^{\theta_2})$ .

Let  $(\hat{\theta}_{1n}, \hat{h}_n)$  denote the modified SMD estimator given by

$$(\hat{\theta}_{1n}, \hat{h}_n) : \arg \min_{\theta_1 \in \Theta_1, h \in \mathcal{H}_n} \left\{ \sum_{j \in \mathcal{J}_{ex}} \frac{1}{n} \sum_{i=1}^n \rho_j(z_i, \theta_1, h)^2 + \sum_{j \in \mathcal{J}_{en}} \frac{1}{n} \sum_{i=1}^n \hat{m}_j(x_{ji}, \theta_1, h)^2 \right\}.$$

We estimate  $\theta_2$  by the plug-in estimator  $\hat{\theta}_{2n}$ , which solves the moment equation:

$$\frac{1}{n} \sum_{i=1}^n \rho^{\theta_2}(z_i, \hat{\theta}_{1n}, \hat{\theta}_{2n}, \hat{h}_n) = 0.$$

Let  $\theta = (\theta'_1, \theta'_2)'$  and  $m^{\theta_2}(\theta, h) = E\{\rho^{\theta_2}(Z, \theta, h)\}$ . Let  $\hat{\theta} = (\hat{\theta}'_1, \hat{\theta}'_2)'$  and  $\hat{m}^{\theta_2}(\theta, h) = \frac{1}{n} \sum_{i=1}^n \rho^{\theta_2}(z_i, \theta, h)$ . Then,  $\hat{\alpha}_n = (\hat{\theta}_n, \hat{h}_n)$  is the solution to

$$\min_{\alpha \in \mathcal{A}_n} \left\{ \sum_{j \in \mathcal{J}_{ex}} \frac{1}{n} \sum_{i=1}^n \rho_j(z_i, \alpha)^2 + \sum_{j \in \mathcal{J}_{en}} \frac{1}{n} \sum_{i=1}^n \hat{m}_j(x_{ji}, \alpha)^2 + \hat{m}^{\theta_2}(\theta, h)' \hat{m}^{\theta_2}(\theta, h) \right\}.$$

Clearly, the plug-in estimator  $\hat{\theta}_{2n}$  is a simple component of the SMD estimator and its asymptotic distribution follows from applying our general results.

Notice that, in the plug-in estimation problem (6)-(7), the parameter  $(\theta_{*1}, h_*)$  may not satisfy the conditional moment restrictions  $m_j(X_j, \theta_{*1}, h_*) = 0$  for  $j = 1, \dots, J-1$ , but  $\theta_{*2}$  must satisfy the moment restriction  $m^{\theta_2}(\theta_{*1}, \theta_{*2}, h_*) = 0$ . This fact could sometimes simplify the asymptotic covariance of  $\hat{\theta}_{2n}$ ; see Section 4 for detailed discussions.

**Remark 2.1.** In the problem (6)-(7), if  $E\{\rho_j(Z, \theta_1, h) - \rho_j(Z, \theta_{*1}, h_*) | X_j\} = \rho_j(Z, \theta_1, h) - \rho_j(Z, \theta_{*1}, h_*)$  for all  $\theta_1 \in \Theta_1, h \in \mathcal{H}$  and for all  $j = 1, \dots, J-1$ , and if further  $\frac{d\rho^{\theta_2}(Z; \theta_{*1}, \theta_2, h_*(\cdot))}{d\theta_2} |_{\theta_{*2}}$  is a constant, then one could estimate  $\alpha_* = (\theta_{*1}, \theta_{*2}, h_*)$  by the simple sieve nonlinear LS estimator  $\hat{\alpha}_{snls} = (\hat{\theta}_{snls}, \hat{h}_{snls})$ , which solves

$$\min_{\alpha \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{J-1} \rho_j(z_i, \theta_1, h)^2 + \rho^{\theta_2}(z_i; \theta_1, \theta_2, h)' \rho^{\theta_2}(z_i; \theta_1, \theta_2, h) \right\}.$$

### 2.3 Examples

Before we present concrete examples, we introduce a space of smooth functions, called the Hölder space. For any  $1 \times d_x$  vector  $\mathbf{a} = (a_1, \dots, a_{d_x})$  of non-negative integers, we write  $|\mathbf{a}| = \sum_{k=1}^{d_x} a_k$ , and for any  $x = (x_1, \dots, x_{d_x})' \in \mathcal{X} \subseteq \mathcal{R}^{d_x}$ , we denote the  $|\mathbf{a}|$ -th derivative of a function  $g : \mathcal{X} \rightarrow \mathcal{R}$  as:

$$\nabla^{\mathbf{a}} g(x) = \frac{\partial^{|\mathbf{a}|}}{\partial x_1^{a_1} \dots \partial x_{d_x}^{a_{d_x}}} g(x).$$

For some  $\gamma > 0$ , let  $\underline{\gamma}$  be the largest integer smaller than  $\gamma$ , and let  $\Lambda^\gamma(\mathcal{X})$  denote the Hölder space of order  $\gamma$ , i.e., a space of functions  $g : \mathcal{X} \rightarrow \mathcal{R}$  which have up to  $\underline{\gamma}$ -th continuous derivatives, and the highest  $(\underline{\gamma}$ -th) derivatives are Hölder continuous with the Hölder exponent  $\gamma - \underline{\gamma} \in (0, 1]$ . Denote the supremum norm as  $\|g\|_\infty = \sup_x |g(x)|$ , and define the Hölder norm as:

$$\|g\|_{\Lambda^\gamma} = \max_{|\mathbf{a}| \leq \underline{\gamma}} \sup_x |\nabla^{\mathbf{a}} g(x)| + \max_{|\mathbf{a}| = \underline{\gamma}} \sup_{x \neq \bar{x}} \frac{|\nabla^{\mathbf{a}} g(x) - \nabla^{\mathbf{a}} g(\bar{x})|}{\sqrt{(x - \bar{x})'(x - \bar{x})}^{\gamma - \underline{\gamma}}} < \infty.$$

The Hölder space  $\Lambda^\gamma(\mathcal{X}) \equiv \{g : \mathcal{X} \rightarrow \mathcal{R} : \|g\|_{\Lambda^\gamma} < \infty\}$  is a Banach space under the norm  $\|\cdot\|_{\Lambda^\gamma}$ . It is known that the Hölder ball (with radius  $c$ )  $\Lambda_c^\gamma(\mathcal{X}) \equiv \{g \in \Lambda^\gamma(\mathcal{X}) : \|g\|_{\Lambda^\gamma} \leq c\}$  is not compact under the norm  $\|\cdot\|_{\Lambda^\gamma}$ ; but when  $\mathcal{X}$  is a bounded and connected set with Lipschitz continuous boundary, the Hölder ball  $\Lambda_c^\gamma(\mathcal{X})$  is compact under the norms  $\|\cdot\|_{\Lambda^{\gamma'}}$  ( $\gamma' \in (0, \gamma)$ ) and  $\|\cdot\|_\infty$ .

The Hölder space is a convenient space to describe classes of smooth functions, other commonly used smooth function classes include Sobolev and Besov spaces. Although our general theory does not require the pseudo-true functions  $h_*$  to be in a Hölder space, we shall make such a convenient assumption in the following two illustrative examples.

**Example 2.1:** (weighted average derivative of a possibly misspecified nonparametric additive LS regression):  $\rho_1(Z, \alpha) = Y - h_1(W_1) - h_2(W_2)$ ,  $\rho_2(Z, \alpha) = \theta - a(W_1)\nabla^s h_1(W_1)$ , where  $s \geq 1$  is a known finite integer, and  $a(\cdot)$  is a known non-negative weight function that goes to zero smoothly at the boundary of the support of  $W_1$ . For simplicity we assume that  $Y$  and  $W_l$  are scalar random variables, and that the density of  $W_l$  is continuous with support  $[b_{1l}, b_{2l}]$ ,  $l = 1, 2$ . Let  $Z = (Y, X_1)'$ ,  $X_1 = (W_1, W_2)'$ , and  $X_2$  be a degenerate random variable. Note that the pointwise derivative of  $\rho_1(Z, \alpha)$  with respect to  $h_1$  and  $h_2$  depends on  $X_1$  only. With  $\alpha = (\theta, h_1, h_2) \in A = \Theta \times \mathcal{H}^1 \times \mathcal{H}^2$ . The pseudo-true value is given by

$$\alpha_* = \arg \inf_{\theta, h_l \in \mathcal{H}^l, l=1,2} \left( E\{[Y - h_1(W_1) - h_2(W_2)]^2\} + E\{[\theta - a(W_1)\nabla^s h_1(W_1)]^2\} \right).$$

Clearly, this example model is correctly specified only when  $E[Y|W_1, W_2] = h_{*1}(W_1) + h_{*2}(W_2)$  holds with probability one. When  $E[Y|W_1, W_2] \neq h_{*1}(W_1) + h_{*2}(W_2)$  holds with positive probability, the example model is incorrectly specified. The following condition is sufficient for the existence of a unique  $\alpha_*$ :

**Condition 2.1.1.** (i)  $W_1$  is not a measurable function of  $W_2$ , and  $W_2$  is not a measurable function of  $W_1$ ; (ii)  $\mathcal{H}^1 = \Lambda^{\gamma_1}([b_{11}, b_{21}])$  with  $\gamma_1 > s \geq 1$ ,  $\mathcal{H}^2 = \{h_2 \in \Lambda^{\gamma_2}([b_{12}, b_{22}]) : h_2(w_{02}) = 0 \text{ for a known } w_{02} \in (b_{12}, b_{22})\}$  with  $\gamma_2 > 1/2$ , and  $\Theta$  is a compact interval containing  $\theta_* = E\{a(W_1)\nabla^s h_{*1}(W_1)\}$ ; (iii)  $E\{[a(W_1)]^2\} < \infty$ ; (iv)  $E[Y^2|X_1]$  is bounded.

Let  $q_j^{k_{h_j n}}(W_j) = (q_{j1}(W_j), \dots, q_{jk_{h_j n}}(W_j))'$  denote either the Fourier series or the spline series (of  $[\gamma_j] + 1$ -th order) on  $[b_{1j}, b_{2j}]$  with  $k_{h_j n}$  number of terms. Let  $\mathcal{H}_n^1 = \{h_1(w_1) = q_1^{k_{h_1 n}}(w_1)' \pi_1 : \|h_1\|_{\Lambda^{\gamma_1}} \leq c \log(k_{h_1 n})\}$  and  $\mathcal{H}_n^2 = \{h_j = q_j^{k_{h_j n}}(w_j)' \pi_j : h_j(w_{0j}) = 0, \|h_j\|_{\infty} \leq c \log(k_{h_j n})\}$ . Then  $\mathcal{H}_n = \mathcal{H}_n^1 \times \mathcal{H}_n^2$  is a sieve space for  $\mathcal{H} = \mathcal{H}^1 \times \mathcal{H}^2$ . Let  $\{z_i = (y_i, x_{1i}) = (y_i, w_{1i}, w_{2i}), i = 1, 2, \dots, n\}$  denote a random sample of observations. The modified SMD estimator is given by

$$\hat{\alpha}_n = \arg \min_{h_1 \in \mathcal{H}_n^1, h_2 \in \mathcal{H}_n^2, \theta \in \Theta} \left( \frac{1}{n} \sum_{i=1}^n [y_i - h_1(w_{1i}) - h_2(w_{2i})]^2 + [\theta - a(w_{1i})\nabla^s h_1(w_{1i})]^2 \right).$$

In Section 4 we show how the model misspecification affects the asymptotic variance of  $\hat{\theta}_n$ .

**Example 2.2:** (weighted average derivative of a possibly misspecified nonparametric IV regression):  $\rho_1(Z, \alpha) = Y_1 - h(Y_2)$ ,  $\rho_2(Z, \alpha) = \theta - a(Y_2)\nabla^s h(Y_2)$ , where  $s \geq 1$  is a known finite integer,  $a(\cdot)$  is a known non-negative weight function,  $Y_1, Y_2$  and  $X_1$  are scalar continuous random variables, the support of  $Y_2$  is  $\mathcal{R}$  and the support of  $X_1$  is  $[a, b]$ . Denote  $Z = (Y_1, Y_2, X_1)$  and  $X_2$  is degenerate.

Let  $m_1(X_1, \alpha) = E[Y_1 - h(Y_2)|X_1]$  and  $m_2(\alpha) = E[\theta - a(Y_2)\nabla^s h(Y_2)]$  with  $\alpha = (\theta, h) \in \mathcal{A} = \Theta \times \mathcal{H}$ . The pseudo-true value is given by

$$\alpha_* = \arg \inf_{\theta \in \Theta, h \in \mathcal{H}} \left( E[\{E[Y_1 - h(Y_2)|X_1]\}^2] + E\{[\theta - a(Y_2)\nabla^s h(Y_2)]^2\} \right).$$

This example model is correctly specified only when  $E[Y_1 - h_*(Y_2)|X_1] = 0$ ; and it is otherwise incorrectly specified. The following condition is sufficient for the existence of a unique  $\alpha_*$ :

**Condition 2.2.1.** (i)  $E\{h(Y_2)|X_1\} = 0$  if and only if  $h(Y_2) = 0$ ; (ii)  $\mathcal{H} = \Lambda_c^\gamma(\mathcal{R})$  with  $\gamma > s \geq 1$ ,  $\Theta$  is a compact interval containing  $\theta_* = E\{a(Y_2)\nabla^s h_*(Y_2)\}$ ; (iii)  $E\{[a(Y_2)]^2\} < \infty$ ; (iv)  $E[\{Y_1 - h_*(Y_2)\}^2|X_1]$  is bounded; (v)  $E[|Y_1|^4] < \infty$ ,  $E[\{1 + (Y_2)^2\}^\varsigma|X_1]$  is bounded for some  $\varsigma > \gamma$ .

To approximate the conditional mean function  $m_1(X_1, \alpha)$ , we shall use the series basis functions such as the cosine series or splines denoted by  $p_1^{k_{1n}}(X_1) = (p_{11}(X_1), \dots, p_{1k_{1n}}(X_1))'$ . The unknown function  $h(Y_2)$  is approximated by some other spline basis functions  $q^{k_{hn}}(Y_2) = (q_1(Y_2), \dots, q_{k_{hn}}(Y_2))'$ ; see Ai and Chen (2003). Let  $\mathcal{H}_n = \{h(y_2) = q^{k_{hn}}(y_2)' \pi : \max_{r \leq \underline{\gamma}} \sup_{y_2} |\nabla^r h(y_2)| \leq c\}$  be a sieve space for  $h$ . Obviously, we need  $k_{1n} \geq k_{hn}$  to estimate the unknown  $h_*$ . Let  $\{z_i = (y_{1i}, y_{2i}, x_{1i}), i = 1, 2, \dots, n\}$  denote a random sample of observations. The series LS estimator of  $m_1(X_1, \alpha)$  is given by:  $\hat{m}_1(X_1, h) = p_1^{k_{1n}}(X_1)'(P_1'P_1)^{-1} \sum_{i=1}^n p_1^{k_{1n}}(x_{1i})\{y_{1i} - h(y_{2i})\}$ . The proposed SMD estimator is

$$\hat{\alpha}_n = \arg \min_{h \in \mathcal{H}_n, \theta \in \Theta} \left( \frac{1}{n} \sum_{i=1}^n \hat{m}_1(x_{1i}, h)^2 + [\theta - a(y_{2i})\nabla^s h(y_{2i})]^2 \right).$$

Since the unknown  $h_*(\cdot)$  depends on the endogenous variable  $Y_2$ , the conditions to ensure  $\sqrt{n}$ -asymptotic normality of the plug-in estimator  $\hat{\theta}_n$  is more restrictive than those for the weighted average derivative estimator of a nonparametric LS regression; see Section 4 for details.

### 3 Consistency and Convergence Rate

We begin by introducing additional notation and definitions to aid the exposition. Let  $c$  denote a generic positive finite constant that may take specific value in specific context. Let  $\|\cdot\|_E$  denote the standard Euclidean norm, and let  $\|\cdot\|_s$  denote a pseudo metric (e.g., the supreme norm or the mean squared metric) on  $\mathcal{A} = \Theta \times \mathcal{H}$ . The following definitions are introduced in Ai and Chen (2003) and restated here.

**Definition 3.1:** A real-valued measurable function  $g(Z, \alpha)$  satisfies an *envelope condition* over  $\alpha \in \mathcal{A}_n$  if there exists a measurable function  $c_1(Z)$  with  $E\{c_1(Z)^4\} < \infty$  such that  $|g(Z, \alpha)| \leq c_1(Z)$  for all  $Z \in \mathcal{Z}$  and  $\alpha \in \mathcal{A}_n$ .

**Definition 3.2:** A real-valued measurable function  $g(Z, \alpha)$  is *Hölder continuous* in  $\alpha \in \mathcal{A}$  (or  $\mathcal{A}_n$ ) if there exist a constant  $\kappa \in (0, 1]$  and a measurable function  $c_2(Z)$  with  $E[c_2(Z)^2|X]$  bounded, such that  $|g(Z, \alpha_1) - g(Z, \alpha_2)| \leq c_2(Z)\|\alpha_1 - \alpha_2\|_s^\kappa$  for all  $Z \in \mathcal{Z}$ ,  $\alpha_1, \alpha_2 \in \mathcal{A}$  (or  $\mathcal{A}_n$ ).

Throughout the paper, let  $N(\delta, \mathcal{A}_{k(n)}, \|\cdot\|_s)$  denote the minimal number of radius  $\delta$  covering balls of  $\mathcal{A}_{k(n)} = \Theta \times \mathcal{H}_{k(n)}$  under the  $\|\cdot\|_s$  metric. Let  $k_{hn}$  denote the number of unknown sieve coefficients of  $h \in \mathcal{H}_{k(n)}$  and  $d_\theta$  the dimension of  $\theta \in \Theta$ . Let  $k_{ex}$  denote the total number of unknown parameters (including both  $\theta$  and sieve coefficients of  $h$ ) appeared in the equation group  $\mathcal{J}_{ex}$ , and let  $\dim(\mathcal{J}_{2en})$  denote the number of equations in  $\mathcal{J}_{2en}$ . For  $j = 1, \dots, J$ , let  $\mathcal{X}_j$  denote the support of  $X_j$  and  $d_{x_j}$  denote the dimension of  $X_j$ . If  $X_j$  is degenerate we denote  $\mathcal{X}_j = \{1\}$  and  $d_{x_j} = 1$ . We first provide mild sufficient conditions for consistency under the stronger metric  $\|\cdot\|_s$ .

**Assumption 3.1.** (i) The data  $\{z_i = (y'_i, x'_i)' : i = 1, 2, \dots, n\}$  are i.i.d.; (ii) for  $j \in \mathcal{J}_{1en}$ ,  $\mathcal{X}_j$  is compact with non-empty interior; (iii) for  $j \in \mathcal{J}_{1en}$ , the density of  $X_j$  is bounded and bounded away from zero.

**Assumption 3.2.** For  $j \in \mathcal{J}_{1en}$ , (i) the smallest and the largest eigenvalues of  $E\{p_j^{k_{jn}}(X_j)p_j^{k_{jn}}(X_j)'\}$  are bounded and bounded away from zero for all  $k_{jn}$ ; (ii) for any  $g \in \{m_j(\cdot, \alpha) : \alpha \in \mathcal{A}_{k(n)}\}$ , there exists  $p_j^{k_{jn}}(\cdot)'\pi$  such that  $E[\{g(X_j) - p_j^{k_{jn}}(X_j)'\pi\}^2] = o(1)$  uniformly over  $\alpha \in \mathcal{A}_{k(n)}$ .

**Assumption 3.3.** There is a pseudo metric  $\|\cdot\|_s$  on  $\mathcal{A}$  such that  $\|\alpha\|_s < \infty$  for all  $\alpha \in \mathcal{A}$ , and for all  $k \geq 1$ , (i)  $\mathcal{A}_k$  is compact under  $\|\cdot\|_s$ ; (ii)  $\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}$ , and for  $\alpha_* \in \mathcal{A}$  there exists  $\Pi_k \alpha_* \in \mathcal{A}_k$  with  $\|\Pi_k \alpha_* - \alpha_*\|_s = o(1)$ .

**Assumption 3.4.** There are a non-increasing positive function  $\delta(\cdot)$  with  $\liminf_k \delta(k) > 0$ , and a positive function  $g(\cdot)$  such that for all  $\varepsilon > 0$  and for all  $k \geq 1$ ,

$$\inf_{\alpha \in \mathcal{A}_k : \|\alpha - \alpha_*\|_s \geq \varepsilon} E[m(X, \alpha)'m(X, \alpha)] - E[m(X, \alpha_*)'m(X, \alpha_*)] \geq \delta(k)g(\varepsilon) > 0.$$

**Assumption 3.5.** For  $j = 1, \dots, J$ , (i)  $E[|\rho_j(Z, \alpha_*)|^2 | X_j]$  is bounded; (ii)  $\rho_j(Z_j, \alpha)$  is Hölder continuous in  $\alpha \in \mathcal{A}$ .

**Assumption 3.6.** (i) for  $j \in \mathcal{J}_{1en}$ ,  $k_{jn} \rightarrow \infty$ ,  $k_{jn}/n \rightarrow 0$ ; and  $\sum_{j \in \mathcal{J}_{1en}} k_{jn} + \dim(\mathcal{J}_{2en}) + k_{ex} \geq d_\theta + k_{hn}$ .

**Assumption 3.7.** (i)  $\ln[N(\varepsilon^{1/\kappa}, \mathcal{A}_{k(n)}, \|\cdot\|_s)] \times n^{-1} \rightarrow 0$ .

Assumptions 3.1(ii)(iii) and 3.2(i)(ii) are typical conditions imposed for series (or linear sieve) LS estimation of conditional mean functions  $m_j(X_j, \alpha)$  for  $j \in \mathcal{J}_{1en}$ . Assumptions 3.1(ii)(iii) require the regressors of the  $j$ -th equation to have bounded supports. These conditions are restrictive but not critical. Trimming can be used so that these conditions are no longer needed. For instance, if  $X_j$  has unbounded support or its density is zero on the boundary of the support, we can replace  $\rho_j(Z, \alpha)$  by  $\rho_j(Z, \alpha)1\{c_0 \leq X_j \leq c_1\}$  for some known constants  $c_0, c_1$  provided that the density of  $X_j$  is positive over  $c_0 \leq X_j \leq c_1$ . It is important to note that we cannot simply discard observations with large  $X_j$  values. Doing so might bias the proposed estimator since  $X_j$  may be endogenous in other equations. Assumptions 3.5(i)(ii) are typically imposed on the residual

function even in the literature about parametric nonlinear estimation. Assumptions 3.1(i)(ii)(iii), 3.2(i)(ii) and 3.5(i)(ii) are useful to establish the convergence of  $\widehat{m}_j(X_j, \alpha)$  to  $m_j(X_j, \alpha)$  uniformly over  $\alpha \in \mathcal{A}_n$  for  $j \in \mathcal{J}_{1en}$ . Assumption 3.6(i) requires that the number of moment conditions is at least as large as the number of unknown coefficients. Assumption 3.7(i) requires that the size of the sieve space  $\mathcal{A}_n$  does not grow too fast in terms of the covering number. For commonly used linear sieves  $\mathcal{A}_n$  such as power series, Fourier series, splines, and wavelet linear sieves, we have  $\ln[N(\delta, \mathcal{A}_n, \|\cdot\|_s)] = ck_{hn} \ln(\frac{1}{\delta})$  [see e.g. Chen and Shen (1998)], hence Assumption 3.7(i) is satisfied as long as  $k_{hn}/n \rightarrow 0$ . For commonly used nonlinear sieves  $\mathcal{A}_n$  such as neural network and ridgelet nonlinear sieves, we have  $\ln[N(\delta, \mathcal{A}_n, \|\cdot\|_s)] = ck_{hn} \ln(\frac{k_{hn}}{\delta})$  [see e.g. Chen and White (1999)], hence Assumption 3.7(i) is satisfied as long as  $k_{hn} \ln(k_{hn})/n \rightarrow 0$ . Assumption 3.3(i) requires that the sieve parameter space  $\mathcal{A}_n$  is compact under  $\|\cdot\|_s$ ; this assumption is weaker than that imposed in Ai and Chen (2003), who assume that the entire parameter space  $\mathcal{A}$  is compact under  $\|\cdot\|_s$ . Assumption 3.3(ii) is effectively the definition of the sieve space  $\mathcal{A}_n$ , which is typically satisfied when the size of the sieve space  $\mathcal{A}_n$  (as measured in terms of covering number  $N(\varepsilon, \mathcal{A}_n, \|\cdot\|_s)$  or  $k_{hn}$ ) grows with the sample size  $n$ .

Assumption 3.4 is an identification condition. It is implied by Assumption 3.5,  $\alpha_*$  being the unique minimizer of  $E[m(X, \alpha)'m(X, \alpha)]$  over  $\mathcal{A}$ , and  $\mathcal{A}$  being compact under  $\|\cdot\|_s$ . When  $h(\cdot)$  does not depend on endogenous variables, the condition  $\liminf_k \delta(k) > 0$  in Assumption 3.4 is typically satisfied even when  $\mathcal{A}$  is not compact (under  $\|\cdot\|_s$ ). However,  $\liminf_k \delta(k)$  could be 0 when  $h(\cdot)$  depends on endogenous variable and  $\mathcal{A}$  is not compact. When  $\liminf_k \delta(k) = 0$  we can still establish  $\|\widehat{\alpha}_n - \alpha_*\|_s = o_p(1)$  after strengthening Assumptions 3.2(ii), 3.3(ii) and 3.7(i) according to how fast  $\delta(k)$  goes to 0; see e.g. Chen (2005, Theorem 3.1, Remark 3.1) for details. The following result is a simple consequence of Theorem 3.1 in Chen (2005) and hence we omit its proof.

**Lemma 3.1.** *Let  $\widehat{\alpha}_n$  be the SMD estimator defined in (5). Under Assumptions 3.1, 3.2(i)(ii), 3.3(i)(ii), 3.4, 3.5(i)(ii), 3.6(i) and 3.7(i), we have  $\|\widehat{\alpha}_n - \alpha_*\|_s = o_p(1)$ .*

Given Lemma 3.1, we can now restrict our attention to a shrinking  $\|\cdot\|_s$ -neighborhood around  $\alpha_*$ . Let  $\mathcal{A}_{os} \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_*\|_s = o(1), \|\alpha\|_s \leq c\}$  and  $\mathcal{A}_{osn} \equiv \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_*\|_s = o(1), \|\alpha\|_s \leq c\}$ . We introduce a pseudo metric  $\|\cdot\|$  on  $\mathcal{A}_{os}$  that is generally weaker than the metric  $\|\cdot\|_s$  (i.e.,  $\|\alpha\| \leq c\|\alpha\|_s$ ), but it is useful when the nonparametric component  $h$  depends on endogenous variables. Let  $\mathcal{A}_{os}$  and  $\mathcal{A}_{osn}$  be convex parameter spaces, and define the pathwise derivatives at the direction  $[\alpha - \alpha_*]$  evaluated at  $\alpha_*$  by:

$$\begin{aligned} \frac{dm(X, \alpha_*)}{d\alpha}[\alpha - \alpha_*] &\equiv \left. \frac{dm(X, (1 - \tau)\alpha_* + \tau\alpha)}{d\tau} \right|_{\tau=0} \quad a.s. \ X; \\ \frac{d^2m(X, \alpha_*)}{d\alpha^2}[\alpha - \alpha_*, \alpha - \alpha_*] &\equiv \left. \frac{d^2m(X, (1 - \tau)\alpha_* + \tau\alpha)}{d\tau^2} \right|_{\tau=0} \quad a.s. \ X. \end{aligned}$$

For any  $\alpha_1, \alpha_2 \in \mathcal{A}_{os}$ , the metric  $\|\cdot\|$  is defined as

$$\|\alpha_1 - \alpha_2\|^2 \equiv E \left\{ \left\| \frac{dm(X, \alpha_*)}{d\alpha} [\alpha_1 - \alpha_2] \right\|_E^2 + \left( \frac{d^2m(X, \alpha_*)}{d\alpha^2} [\alpha_1 - \alpha_2, \alpha_1 - \alpha_2] \right)' m(X, \alpha_*) \right\}.$$

By construction, the metric  $\|\alpha_1 - \alpha_2\|^2$  is the second pathwise derivative of the population objective function:

$$\frac{d^2 E \{ m(X, \alpha_* + \tau(\alpha_1 - \alpha_2))' m(X, \alpha_* + \tau(\alpha_1 - \alpha_2)) \} / 2}{d\tau^2} \Big|_{\tau=0},$$

which must be non-negative since  $\alpha_*$  is the minimizer.

In general, the metric  $\|\alpha_1 - \alpha_2\|$  defined here differs from the norm,  $\sqrt{E \left\{ \left\| \frac{dm(X, \alpha_o)}{d\alpha} [\alpha_1 - \alpha_2] \right\|_E^2 \right\}}$ , introduced in Ai and Chen (2003). Note that the two metrics are identical if and only if

$$\sum_{j=1}^J E \left\{ \left( \frac{d^2 m_j(X_j, \alpha_*)}{d\alpha^2} [v, v] \right) m_j(X_j, \alpha_*) \right\} = 0 \quad \text{for all } v \in \mathcal{A}_{os},$$

which is satisfied if for all  $j = 1, \dots, J$ , either  $m_j(X_j, \alpha_*) = 0$  (the  $j$ -th conditional moment restriction is satisfied), or  $m_j(X_j, \alpha)$  is linear in  $\alpha$ .

For the purpose of establishing a rate of convergence under the  $\|\cdot\|$  metric, we can treat  $\mathcal{A}_{os}$  as the new parameter space and  $\mathcal{A}_{osn}$  as its sieve space. Denote  $N(\delta, \mathcal{A}_{osn}, \|\cdot\|_s)$  as the minimal number of radius  $\delta$  covering balls of  $\mathcal{A}_{osn}$  under the  $\|\cdot\|_s$  metric. For every  $j \in \mathcal{J}_{1en}$  let  $\xi_{jn} \equiv \sup_{X_j \in \mathcal{X}_j} \left\| p_j^{k_{jn}}(X_j) \right\|_E$ , which is nondecreasing in  $k_{jn}$ . The following conditions are similar to those imposed in Ai and Chen (2003), except that Assumptions 3.2(iii), 3.5(iii)(iv) and 3.7 are only required to be satisfied over the local sieve  $\mathcal{A}_{osn}$  (instead of the original sieve  $\mathcal{A}_n$ ).

**Assumption 3.2.** for  $j \in \mathcal{J}_{1en}$ , (iii) for any  $g(\cdot) \in \Lambda_c^{\gamma_j}(\mathcal{X}_j)$  with  $\gamma_j > d_{x_j}/2$ , there exists  $p_j^{k_{jn}}(\cdot)' \pi \in \Lambda_c^{\gamma_j}(\mathcal{X}_j)$  such that  $\sup_{X_j \in \mathcal{X}_j} |g(X_j) - p_j^{k_{jn}}(X_j)' \pi| = O(k_{jn}^{-\gamma_j/d_{x_j}})$  and  $n^{1/4} k_{jn}^{-\gamma_j/d_{x_j}} \rightarrow 0$ .

**Assumption 3.3.** (iii) There is a finite constant  $c > 0$  such that for all  $\alpha \in \mathcal{A}_{os}$  we have  $\|\alpha\| \leq c \|\alpha\|_s < \infty$ ; (iv) there is a constant  $\mu_0 > 0$  such that  $\|\Pi_{k(n)} \alpha_* - \alpha_*\| = O(k_{hn}^{-\mu_0})$  and  $n^{1/4} k_{hn}^{-\mu_0} \rightarrow 0$ .

**Assumption 3.5.** for  $j \in \mathcal{J}_{1en}$ , (iii)  $\rho_j(Z, \alpha)$  satisfies an envelope condition in  $\alpha \in \mathcal{A}_{osn}$ ; (iv)  $m_j(\cdot, \alpha) \in \Lambda_c^{\gamma_j}(\mathcal{X}_j)$  with  $\gamma_j > d_{x_j}/2$  uniformly in  $\alpha \in \mathcal{A}_{osn}$ .

**Assumption 3.6.** for all  $j \in \mathcal{J}_{1en}$ , (ii)  $\ln[N(\varepsilon^{1/\kappa}, \mathcal{A}_{osn}, \|\cdot\|_s)] \times \xi_{jn}^2 \times n^{-1/2} \rightarrow 0$ .

**Assumption 3.7.** (ii)  $\ln[N(\varepsilon^{1/\kappa}, \mathcal{A}_{osn}, \|\cdot\|_s)] \times n^{-1/2} \rightarrow 0$ .

**Assumption 3.8.** (i)  $\mathcal{A}_{os}$  is convex at  $\alpha_*$ ; (ii)  $\rho(Z, \alpha)$  is continuously twice pathwise differentiable with respect to  $\alpha \in \mathcal{A}_{os}$ ; (iii) there is a positive finite constant  $c$  such that for all  $\alpha \in \mathcal{A}_{osn}$ ,

$$c \|\alpha - \alpha_*\|^2 \leq \sum_{j \in \mathcal{J}_{ex}} E[\rho_j(Z, \alpha)^2 - \rho_j(Z, \alpha_*)^2] + \sum_{j \in \mathcal{J}_{en}} E[m_j(X_j, \alpha)^2 - m_j(X_j, \alpha_*)^2].$$

Assumptions 3.2(iii), 3.5(iii)(iv) and 3.6(ii) are sufficient conditions to establish convergence rate of  $\hat{m}_j(X_j, \alpha)$  to  $m_j(X_j, \alpha)$  uniformly over  $\alpha \in \mathcal{A}_{osn}$  for  $j \in \mathcal{J}_{1en}$ . These conditions are not

needed when  $\mathcal{J}_{1en}$  is an empty set. Assumptions 3.2(iii) imply that, for all  $\alpha \in \mathcal{A}_{osn}$ , the linear sieve  $p_j^{k_{jn}}(\cdot)' \pi$  can approximate any conditional mean function  $m_j(\cdot, \alpha)$  in the Hölder ball well. It is known that the method of sieves (or series) can allow for random variables that have discrete probability distributions. However, to make the presentation simple, in most part of the paper we implicitly assume that  $X_j$  has continuous density and satisfies Assumptions 3.1(ii)(iii). Then Assumption 3.2(iii) is satisfied by polynomial, B-spline, and Fourier and many other linear sieves. Assumption 3.5(iv) is satisfied if the conditional density of  $Z$  given  $X_j$  is sufficiently smooth with respect to  $X_j$ . Assumptions 3.5(ii)(iii) impose some typical restrictions on the residual function. Assumption 3.6(ii) can be verified after  $\xi_{jn}$  is computed. For example,  $\xi_{jn} = k_{jn}^{1/2}$  if  $p_j^{k_{jn}}(X_j)$  is a tensor-product B-spline basis of order  $[\gamma_j] + 1$  or a Fourier series sieve;  $\xi_{jn} = k_{jn}$  if  $p_j^{k_{jn}}(X_j)$  is a tensor-product polynomial power series sieve; see Newey (1997) for more examples. Define  $\tilde{L}_n(\alpha) \equiv \frac{1}{2n} \sum_{i=1}^n \ell(z_i, \alpha)$  with

$$\ell(z_i, \alpha) \equiv - \left( \sum_{j \in \mathcal{J}_{ex}} \rho_j(z_i, \alpha)^2 + \sum_{j \in \mathcal{J}_{en}} [2m_j(x_{ji}, \alpha) \rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha)^2] \right). \quad (8)$$

Note that  $\alpha_*$  also solves  $\sup_{\alpha \in \mathcal{A}} E\{\tilde{L}_n(\alpha)\}$ . In the Appendix we show that, given Lemma 3.1 and Assumptions 3.1, 3.2, 3.5 and 3.6, the SMD estimator  $\hat{\alpha}_n$  given in (5) also solves

$$\max_{\alpha \in \mathcal{A}_{osn}} \tilde{L}_n(\alpha) - o_p(n^{-1/2}).$$

Assumptions 3.3(iii), 3.7(ii) and 3.8 are sufficient conditions for the faster than  $n^{-1/4}$  convergence rate under the  $\|\cdot\|$  metric for the sieve M-estimator,  $\arg \max_{\alpha \in \mathcal{A}_{osn}} \tilde{L}_n(\alpha)$ , hence they are imposed even when the conditional mean functions  $m_j(X_j, \alpha)$  (for  $j \in \mathcal{J}_{1en}$ ) were known. Assumptions 3.3(i)(ii) imply that  $\Pi_n \alpha_* \in \mathcal{A}_{osn}$ . Assumption 3.3(iii) is on the approximation error rate (under the  $\|\cdot\|$  metric) of the sieve space  $\mathcal{A}_{osn}$  to the parameter space  $\mathcal{A}_{os}$ . This condition is satisfied if the parameter space is a Hölder ball, and the approximating functions are power series, Fourier series or B-splines. Assumption 3.7(ii) requires that the size of the sieve space  $\mathcal{A}_{osn}$  does not grow too fast in terms of the covering number. Recall that  $\mathcal{A}_{osn}$  is a small subset of the original sieve space  $\mathcal{A}_n$ . For commonly used linear sieves we have  $\ln[N(\varepsilon, \mathcal{A}_{osn}, \|\cdot\|_s)] \leq ck_{hn} \ln(\frac{1}{\varepsilon})$ , and for commonly used nonlinear sieves we have  $\ln[N(\varepsilon, \mathcal{A}_{osn}, \|\cdot\|_s)] \leq ck_{hn} \ln(\frac{k_{hn}}{\varepsilon})$ . Assumption 3.8 requires that the metric  $\|\cdot\|^2$  is well-defined and can locally approximate the population criterion difference. This condition is trivially satisfied when  $\rho_j$  is linear in  $\alpha$ . When  $\rho_j$  is nonlinear in  $\alpha$ , this condition is still satisfied in the neighborhood of  $\alpha_*$  defined by  $\|\alpha - \alpha_*\|_s = o(1)$ , as long as the third order term in the Taylor expansion of  $E\{m(X, \alpha_*(1 - \tau) + \tau\alpha)' m(X, \alpha_*(1 - \tau) + \tau\alpha)\}/2$  around  $\tau = 0$  is dominated by the second order term.

Assumptions 3.6(ii) and 3.7(ii) are respectively implied by Assumptions 3.6'(ii) and 3.7'(ii), which were used in Ai and Chen (2003):

**Assumption 3.6'**. for all  $j \in \mathcal{J}_{1en}$ , (ii)  $k_{hn} \times \ln n \times \xi_{jn}^2 \times n^{-1/2} \rightarrow 0$ .

**Assumption 3.7'**. (ii)  $\ln[N(\varepsilon^{1/\kappa}, \mathcal{A}_{osn}, \|\cdot\|_s)] \leq ck_{hn} \ln(\frac{k_{hn}}{\varepsilon})$  and  $k_{hn}(\ln n)n^{-1/2} \rightarrow 0$ .

**Theorem 3.1.** Let  $\hat{\alpha}_n$  be the SMD estimator defined in (5). Suppose Assumptions 3.1 - 3.8 hold. Then:  $\|\hat{\alpha}_n - \alpha_*\| = o_p(n^{-1/4})$ .

Assumptions 3.1 - 3.8 are low-level sufficient conditions and are easy to verify in specific applications once the pseudo norms are defined. For instance, for **Example 2.1**, the norms are

$$\|\alpha - \alpha_*\|^2 = E\{[\sum_{l=1}^2\{h_l(W_l) - h_{*l}(W_l)\}]^2\} + (\theta - \theta_* - E[a(W_1)\nabla^s\{h_1(W_1) - h_{*1}(W_1)\}])^2,$$

and  $\|\alpha - \alpha_*\|_s = |\theta - \theta_*| + \|\nabla^s\{h_1 - h_{*1}\}\|_\infty + \sum_{l=1}^2\|h_l - h_{*l}\|_\infty$ . For **Example 2.2**, the norms are

$$\|\alpha - \alpha_*\|^2 = E\left\{(E[h(Y_2) - h_*(Y_2)|X_1])^2\right\} + (\theta - \theta_* - E[a(Y_2)\nabla^s\{h(Y_2) - h_*(Y_2)\}])^2,$$

and  $\|\alpha - \alpha_*\|_s = |\theta - \theta_*| + \|\omega\nabla^s\{h_1 - h_{*1}\}\|_\infty + \|\omega\{h_1 - h_{*1}\}\|_\infty$  with  $\omega(y_2) = [1 + |y_2|]^{-\varsigma}$ . It is easy to show that Assumptions 3.1 - 3.8 of Theorem 3.1 are trivially satisfied by Condition 2.1.1 for Example 2.1 and by Condition 2.2.1 for Example 2.2.

## 4 Asymptotic Normality

We now derive the asymptotic distribution of the modified SMD estimator  $\hat{\theta}_n$ . The approach follows the one in Ai and Chen (2003) closely, except that the semiparametric conditional moment restriction (1) may not be satisfied. Define the pathwise derivatives as

$$\begin{aligned} \frac{dm(X, \alpha_*)}{dh}[h - h_*] &= \frac{dm(X, \theta_*, h_*(1 - \tau) + \tau h)}{d\tau}\Big|_{\tau=0}; \\ \frac{d^2m(X, \alpha_*)}{dh^2}[h - h_*, h - h_*] &= \frac{d^2m(X, \theta_*, h_*(1 - \tau) + \tau h)}{d\tau^2}\Big|_{\tau=0}; \\ \frac{d^2m(X, \alpha_*)}{\partial\theta dh}[h - h_*] &= \frac{d(\partial m(X, \theta_*, h_*(1 - \tau) + \tau h)/\partial\theta)}{d\tau}\Big|_{\tau=0}. \end{aligned}$$

Let  $\bar{\mathcal{V}}$  denote the closure of the linear span of  $\mathcal{A} - \{\alpha_*\} = \{\alpha - \alpha_* : \text{for all } \alpha \in \mathcal{A}\}$  under the metric  $\|\cdot\|$ . Then we can write  $\bar{\mathcal{V}} = \mathcal{R}^{d_\theta} \times \bar{\mathcal{W}}$  with  $\bar{\mathcal{W}} \equiv \bar{\mathcal{H}} - \{h_*\}$ . For each component  $\theta_l$  (of  $\theta$ ),  $l = 1, \dots, d_\theta$ , suppose that there exists a  $w_l^* \in \bar{\mathcal{W}}$  that solves:

$$w_l^* : \inf_{w_l \in \bar{\mathcal{W}}} E \left\{ \begin{aligned} &\left( \frac{\partial m(X, \alpha_*)}{\partial\theta_l} - \frac{dm(X, \alpha_*)}{dh}[w_l] \right)' \left( \frac{\partial m(X, \alpha_*)}{\partial\theta_l} - \frac{dm(X, \alpha_*)}{dh}[w_l] \right) \\ &+ \left( \frac{\partial^2 m(X, \alpha_*)}{\partial\theta_l^2} - 2 \frac{d^2 m(X, \alpha_*)}{\partial\theta_l dh} [w_l] + \frac{d^2 m(X, \alpha_*)}{dh^2} [w_l, w_l] \right)' m(X, \alpha_*) \end{aligned} \right\}.$$

Denote  $w^* = (w_1^*, \dots, w_{d_\theta}^*)$ ,

$$\begin{aligned}
\frac{dm(X, \alpha_*)}{dh}[w^*] &= \left( \frac{dm(X, \alpha_*)}{dh}[w_1^*], \dots, \frac{dm(X, \alpha_*)}{dh}[w_{d_\theta}^*] \right), \\
\frac{d^2m(X, \alpha_*)}{\partial\theta dh}[w^*] &= \left( \frac{d^2m(X, \alpha_*)}{\partial\theta dh}[w_1^*], \dots, \frac{d^2m(X, \alpha_*)}{\partial\theta dh}[w_{d_\theta}^*] \right); \\
\frac{d^2m(X, \alpha_*)}{dh^2}[w^*, w^*] &= \begin{pmatrix} \frac{d^2m(X, \alpha_*)}{dh^2}[w_1^*, w_1^*] & \dots & \frac{d^2m(X, \alpha_*)}{dh^2}[w_1^*, w_{d_\theta}^*] \\ \dots & \dots & \dots \\ \frac{d^2m(X, \alpha_*)}{dh^2}[w_{d_\theta}^*, w_1^*] & \dots & \frac{d^2m(X, \alpha_*)}{dh^2}[w_{d_\theta}^*, w_{d_\theta}^*] \end{pmatrix}.
\end{aligned}$$

Also denote

$$\begin{aligned}
D_{w^*}(X) &\equiv \frac{\partial m(X, \alpha_*)}{\partial\theta'} - \frac{dm(X, \alpha_*)}{dh}[w^*]; \quad D_{jw^*}(X) \equiv \frac{\partial m_j(X_j, \alpha_*)}{\partial\theta'} - \frac{dm_j(X_j, \alpha_*)}{dh}[w^*]; \\
V_{w^*}(X) &= \sum_{j=1}^J \left( \frac{\partial^2 m_j(X_j, \alpha_*)}{\partial\theta\partial\theta'} - 2 \frac{d^2 m_j(X_j, \alpha_*)}{\partial\theta dh}[w^*] + \frac{d^2 m_j(X_j, \alpha_*)}{dh^2}[w^*, w^*] \right) m_j(X_j, \alpha_*).
\end{aligned}$$

Suppose that  $E\{D_{w^*}(X)'D_{w^*}(X) + V_{w^*}(X)\}$  is nonsingular. For any fixed  $\lambda \neq 0$ , denote  $v^* \equiv (v_\theta^*, v_h^*)$  with

$$v_\theta^* = (E\{D_{w^*}(X)'D_{w^*}(X) + V_{w^*}(X)\})^{-1}\lambda \quad \text{and} \quad v_h^* = -w^* \times v_\theta^*.$$

We impose the following additional conditions for  $\sqrt{n}$ -asymptotic normality of  $\widehat{\theta}_n$ :

**Assumption 4.1.** (i)  $w^*$  exists (i.e.,  $w_l^* \in \overline{\mathcal{W}}$  for  $l = 1, \dots, d_\theta$ ) and  $E[D_{w^*}(X)'D_{w^*}(X) + V_{w^*}(X)]$  is positive-definite; (ii)  $\theta_* \in \text{int}(\Theta)$ .

Assumption 4.1 implies that  $\lambda'(\theta - \theta_*) = \langle v^*, \alpha - \alpha_* \rangle$  for all  $\alpha \in \mathcal{A}$ , where  $\langle \cdot, \cdot \rangle$  denotes the inner product induced by the norm  $\|\cdot\|$ .

**Assumption 4.2.** There is a  $v_n^* \equiv (v_\theta^*, -\Pi_n w^* \times v_\theta^*) \in \mathcal{A}_n - \{\alpha_*\}$  such that  $\|v_n^* - v^*\| \times \|\widehat{\alpha}_n - \alpha_*\| = o_p(n^{-1/2})$ .

Given Theorem 3.1 ( $\|\widehat{\alpha}_n - \alpha_*\| = o_p(n^{-1/4})$ ), Assumption 4.2 is implied by Assumption 4.2':

**Assumption 4.2'.** There is a  $v_n^* \equiv (v_\theta^*, -\Pi_n w^* \times v_\theta^*) \in \mathcal{A}_n - \{\alpha_*\}$  such that  $\|v_n^* - v^*\| = O(n^{-1/4})$ .

Denote  $\mathcal{N}_o \equiv \{\alpha \in \mathcal{A}_{os} : \|\alpha - \alpha_*\| = o(n^{-1/4})\}$  and  $\mathcal{N}_{on} \equiv \{\alpha \in \mathcal{A}_{osn} : \|\alpha - \alpha_*\| = o(n^{-1/4})\}$ . Define  $\frac{d\rho(Z, \alpha)}{d\alpha}[v_n^*]$  and  $\frac{d^2\rho(Z, \alpha)}{d\alpha^2}[v_n^*, v_n^*]$  analogously to  $\frac{dm(X, \alpha)}{d\alpha}[v_n^*]$  and  $\frac{d^2m(X, \alpha)}{d\alpha^2}[v_n^*, v_n^*]$  respectively.

**Assumption 4.3.** (i) For  $j = 1, \dots, J$ ,  $E\left(\left\{\frac{d\rho_j(Z, \alpha_*)}{d\alpha}[v_n^*]\right\}^2 | X_j\right)$  is bounded, and  $\frac{d\rho_j(Z, \alpha)}{d\alpha}[v_n^*]$  is Hölder continuous in  $\alpha \in \mathcal{N}_o$ ; (ii) for  $j = 1, \dots, J$ , there is a function  $c_5(Z)$  with  $E\{[c_5(Z)]^2\} < \infty$  such that  $\left|\frac{d^2\rho_j(Z, \alpha)}{d\alpha^2}[v_n^*, v_n^*]\right| \leq c_5(Z)$  for all  $\alpha \in \mathcal{N}_{on}$ ; (iii) for  $j \in \mathcal{J}_{1en}$ ,  $\frac{d\rho_j(Z, \alpha)}{d\alpha}[v_n^*]$  satisfies the envelope condition and  $\frac{dm_j(X_j, \alpha)}{d\alpha}[v_n^*]$  is in  $\Lambda_c^{\gamma_j}(\mathcal{X}_j)$ ,  $\gamma_j > d_{x_j}/2$ , for all  $\alpha \in \mathcal{N}_o$ .

**Assumption 4.4.** With  $\alpha(t) = \alpha_* + t(\alpha - \alpha_*)$ ,

$$\sup_{0 \leq t \leq 1} \left| \frac{d^2 E \left[ \left\{ \frac{dm(X, \alpha(t))}{d\alpha}[v_n^*] \right\}' m(X, \alpha(t)) \right]}{dt^2} \right| = o(n^{-1/2}) \text{ uniformly over } \alpha \in \mathcal{N}_{on}.$$

**Assumption 4.5.**  $\int_0^1 \sqrt{\ln[N(\varepsilon^{1/\kappa}, \mathcal{N}_{on}, \|\cdot\|_s)]} d\varepsilon < \infty$ .

**Assumption 4.6.**  $E \left( \left\{ \left\{ \frac{dm(x_i, \alpha_*)}{d\alpha} [v_n^* - v^*] \right\}' \rho(z_i, \alpha_*) + \left\{ \frac{d(\rho(z_i, \alpha_*) - m(x_i, \alpha_*))}{d\alpha} [v_n^* - v^*] \right\}' m(x_i, \alpha_*) \right\}^2 \right)$  goes to zero as  $\|v_n^* - v^*\|$  goes to zero.

Assumption 4.1(i) is critical for obtaining the  $\sqrt{n}$  convergence of  $\hat{\theta}$  to  $\theta_*$  and its asymptotic normality. There exist semiparametric models that do not satisfy Assumption 4.1(i). We notice that it is possible that  $\theta_*$  is uniquely identified but Assumption 4.1(i) is not satisfied. If this happens,  $\theta_*$  can still be consistently estimated but the best achievable convergence rate is slower than the  $\sqrt{n}$ -rate. In a sense, Assumption 4.1(i) gives a class of models in which it is possible to obtain the  $\sqrt{n}$ -consistency. Assumption 4.2 controls the approximation bias; it is satisfied if  $w^*$  belongs to some typical smooth function class (such as a Hölder, Sobolev or Besov space). This condition imposes additional smoothness requirement on a semiparametric model. It is possible that Assumption 4.1 is satisfied but Assumption 4.2 may not without additional smoothness restriction. Assumption 4.3 is similar to Assumption 3.5 except that it is imposed on the derivatives. Assumptions 4.3(i)(iii) are used to establish consistency with convergence rate of  $\frac{d\hat{m}_j(X_j, \hat{\alpha})}{d\alpha} [v_n^*]$  to  $\frac{dm_j(X_j, \alpha_*)}{d\alpha} [v_n^*]$  for  $j \in \mathcal{J}_{1en}$ . Assumption 4.4 is needed when  $\alpha$  enters  $\rho$  in a highly nonlinear manner. This condition is imposed to control the asymptotic bias when  $\alpha$  enters  $\rho$  nonlinearly. It is similar to the assumptions 4.4 - 4.5 of Ai and Chen (2003) in the sense that it requires, within a shrinking neighborhood, the third order term is bounded by the second order term. But it imposes a stronger restriction on the function  $m(X, \alpha)$  when  $m(X, \alpha_*) \neq 0$  with positive probability. Notice that when  $\rho$  is linear in  $\alpha$ , Assumptions 4.3 and 4.4 are trivially satisfied.

In the Appendix we show that the modified SMD estimator also maximizes  $\frac{1}{n} \sum_{i=1}^n \ell(z_i, \alpha)$  over  $\mathcal{A}_{osn}$ , where  $\ell(z_i, \alpha)$  is given in (8). Notice that

$$\frac{-1}{2} \frac{d\ell(z_i, \alpha)}{d\alpha} [v_n^*] = \left\{ \frac{dm(x_i, \alpha)}{d\alpha} [v_n^*] \right\}' \rho(z_i, \alpha) + \left\{ \frac{d(\rho(z_i, \alpha) - m(x_i, \alpha))}{d\alpha} [v_n^*] \right\}' m(x_i, \alpha).$$

Under Assumptions 3.5 and 4.3, Assumption 4.5 is a sufficient condition for

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{d\ell(z_i, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{d\ell(z_i, \alpha_*)}{d\alpha} [v_n^*] \right) - E \left( \frac{d\ell(z_i, \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{d\ell(z_i, \alpha_*)}{d\alpha} [v_n^*] \right) = o_p(n^{-1/2})$$

uniformly over  $\tilde{\alpha} \in \mathcal{N}_{on}$ . Therefore, Assumption 4.5 can be replaced by any other sufficient conditions for this stochastic equicontinuity condition. In applications, Assumption 4.5 is typically implied by Assumption 3.7(ii). Assumption 4.6 ensures that

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{d\ell(z_i, \alpha_*)}{d\alpha} [v_n^*] - \frac{d\ell(z_i, \alpha_*)}{d\alpha} [v^*] \right) = o_p(n^{-1/2}).$$

Notice that, when  $\left\{ \frac{d(\rho(z_i, \alpha_*) - m(x_i, \alpha_*))}{d\alpha} [v_n^* - v^*] \right\}' m(x_i, \alpha_*) = 0$ , which can happen if  $m(X, \alpha_*) = 0$ , Assumption 4.6 is implied by Assumptions 4.2 and 3.5(i). Thus, Assumption 4.6 is not needed when the semiparametric conditional moment model (1) is correctly specified.

**Remark 4.1** (i)  $E\{V_{w^*}(X)\} = 0$  if for all  $j = 1, \dots, J$ , either  $m_j(X_j, \alpha_*) = 0$  (the  $j$ -th conditional moment restriction is satisfied), or  $m_j(X_j, \alpha)$  is linear in  $\alpha$ .

(ii) When  $E\{V_{w^*}(X)\} = 0$ , the Riesz representer  $v^*$  (or  $w^*$ ) is the same as the one defined in Ai and Chen (2003) under correct specification of the conditional moment restriction (1). In this case Assumption 4.2' becomes:

$$\|v_n^* - v^*\|^2 = v_\theta^{*\prime} E \left\{ \left( \frac{dm(X, \alpha_*)}{dh} [w^* - \Pi_n w^*] \right)' \left( \frac{dm(X, \alpha_*)}{dh} [w^* - \Pi_n w^*] \right) \right\} v_\theta^* = O(n^{-1/2}).$$

Denote

$$\Omega_* \equiv Cov \left\{ \left[ \frac{\partial \rho(Z, \alpha_*)}{\partial \theta'} - \frac{d\rho(Z, \alpha_*)}{dh} [w^*] - D_{w^*}(X) \right]' m(X, \alpha_*) + D_{w^*}(X)' \rho(Z, \alpha_*) \right\}.$$

The following result is proved in the Appendix.

**Theorem 4.1.** Under Assumptions 3.1 - 3.8 and 4.1 - 4.6,  $\sqrt{n}(\hat{\theta}_n - \theta_*) \implies N(0, V^{-1})$  where

$$V^{-1} \equiv (E\{D_{w^*}(X)' D_{w^*}(X) + V_{w^*}(X)\})^{-1} \Omega_* (E\{D_{w^*}(X)' D_{w^*}(X) + V_{w^*}(X)\})^{-1}. \quad (9)$$

When the conditional moment restriction (1) is satisfied (i.e.,  $m(X, \alpha_*) = 0$  and  $\alpha_* = \alpha_o$ ), we have  $V_{w^*}(X) = 0$  and  $\Omega_* = Var\{D_{w^*}(X)' \rho(Z, \alpha_o)\}$ , and the asymptotic covariance  $V^{-1}$  in Theorem 4.1 becomes

$$V^{-1} = (E\{D_{w^*}(X)' D_{w^*}(X)\})^{-1} Var\{D_{w^*}(X)' \rho(Z, \alpha_o)\} (E\{D_{w^*}(X)' D_{w^*}(X)\})^{-1} \quad (10)$$

which is the asymptotic covariance derived in Ai and Chen (2003, Theorem 4.1) for the SMD estimator with identity weighting matrix. When the conditional moment model (1) is not satisfied, the asymptotic covariance  $V^{-1}$  of  $\hat{\theta}$  is generally different from the asymptotic covariance (10).

**Remark 4.2:** (i) When the conditional moment restriction (1) is not satisfied (e.g.  $m(X, \alpha_*) \neq 0$  and  $\alpha_* \neq \alpha_o$ ), there are still cases where  $E\{V_{w^*}(X)\} = 0$  and  $\Omega_* = Var\{D_{w^*}(X)' \rho(Z, \alpha_*)\}$ . In these cases, the asymptotic covariance  $V^{-1}$  in Theorem 4.1 simplifies to

$$V^{-1} = (E\{D_{w^*}(X)' D_{w^*}(X)\})^{-1} Var\{D_{w^*}(X)' \rho(Z, \alpha_*)\} (E\{D_{w^*}(X)' D_{w^*}(X)\})^{-1}. \quad (11)$$

Remark 4.1 discusses cases where  $E\{V_{w^*}(X)\} = 0$  holds. Note that  $\Omega_* = Var\{D_{w^*}(X)' \rho(Z, \alpha_*)\}$  if

$$\sum_{j=1}^J \left[ \frac{\partial \rho_j(Z, \alpha_*)}{\partial \theta'} - \frac{d\rho_j(Z, \alpha_*)}{dh} [w^*] - E \left\{ \frac{\partial \rho_j(Z, \alpha_*)}{\partial \theta'} - \frac{d\rho_j(Z, \alpha_*)}{dh} [w^*] | X_j \right\} \right]' m_j(X_j, \alpha_*) = 0,$$

which is satisfied if for all  $j = 1, \dots, J$ , either  $m_j(X_j, \alpha_*) = 0$  or  $E\left\{ \frac{d\rho_j(Z, \alpha_*)}{d\alpha} [v] | X_j \right\} = \frac{d\rho_j(Z, \alpha_*)}{d\alpha} [v]$ .

(ii) For the plug-in sieve LS problem in Remark 2.1, we have  $\Omega_* = Var\{D_{w^*}(X)' \rho(Z, \alpha_*)\}$ . If  $\rho_j(Z, \alpha)$  is linear in  $\alpha$  for  $j = 1, \dots, J - 1$ , then we have  $E\{V_{w^*}(X)\} = 0$ . Therefore for the

special plug-in sieve *linear* LS problem, the asymptotic variance  $V^{-1}$  of  $\hat{\theta}$  has the form of (11). However, for the plug-in sieve *nonlinear* LS problem, the model misspecification ( $E\{\rho_j(Z, \alpha_*)|X_j\} \neq 0, j = 1, \dots, J-1$ ) and nonlinearity ( $\rho_j(Z, \alpha)$  is nonlinear in  $\alpha$  for  $j = 1, \dots, J-1$ ) together imply  $E\{V_{w^*}(X)\} \neq 0$ ; in this case the asymptotic variance  $V^{-1}$  of  $\hat{\theta}$  is more complicated than (11).

(iii) Even if the asymptotic covariance  $V^{-1}$  could take the simplified form of (11), it may still differ from the one in (10) under correct specification. This is because  $Var\{D_{w^*}(X)' \rho(Z, \alpha_*)\}$  for a misspecified model may differ from the expression  $Var\{D_{w^*}(X)' \rho(Z, \alpha_o)\}$  for a correctly specified model; the difference is due to the presence of some correlation terms under misspecification. See the example below.

#### 4.1 Possibly misspecified nonparametric additive LS regression

We now apply Theorem 4.1 to **Example 2.1**. Recall that  $\rho_1(Z, \alpha) = Y - h_1(W_1) - h_2(W_2)$ ,  $\rho_2(Z, \alpha) = \theta - a(W_1)\nabla^s h_1(W_1)$ ,  $m_1(X_1, \alpha) = E[Y|X_1] - h_1(W_1) - h_2(W_2)$  and  $m_2(\alpha) = \theta - E\{a(W_1)\nabla^s h_1(W_1)\}$  where  $X_1 = (W_1, W_2)'$  and  $X_2$  is degenerate. It is easy to show  $\Omega_* = Var\{D_{w^*}(X)' \rho(Z, \alpha_*)\}$  and  $V_{w^*}(X) = 0$ . To apply Theorem 4.1 it suffices to verify Assumptions 4.1 and 4.2' where  $w^* \in \overline{W}$  solves the following minimization problem:

$$\inf_{w \in \overline{W}} E\{D_w(X)' D_w(X)\} = \inf_{w \in \overline{W}} \left\{ E[\{w^1(W_1) + w^2(W_2)\}^2] + \left[1 + E\{a(W_1)\nabla^s w^1(W_1)\}\right]^2 \right\},$$

where  $\overline{W} = \{(w^1, w^2) : E[\{w^j(W_j)\}^2] < \infty, j = 1, 2; [E\{a(W_1)\nabla^s w^1(W_1)\}]^2 < \infty\}$ . By calculus variation,  $w^{*j}(W_j), j = 1, 2$  solve:

$$E\left\{\left\{\sum_{j=1}^2 w^{*j}(W_j)\right\} \delta_1(W_1)\right\} + \left(1 + E\left\{a(W_1)\nabla^s w^{*1}(W_1)\right\}\right) E\left\{a(W_1)\nabla^s \delta_1(W_1)\right\} = 0, \quad (12)$$

$$E\left\{\left\{\sum_{j=1}^2 w^{*j}(W_j)\right\} \delta_2(W_2)\right\} = 0, \quad (13)$$

for any measurable function  $(\delta_1, \delta_2) \in \overline{W}$ . Then

$$\begin{aligned} E\{D_{w^*}(X)' D_{w^*}(X)\} &= E[\{\sum_{j=1}^2 w^{*j}(W_j)\}^2] + \left[1 + E\{a(W_1)\nabla^s w^{*1}(W_1)\}\right]^2 \\ &= 1 + E\{a(W_1)\nabla^s w^{*1}(W_1)\}. \end{aligned}$$

Let  $f_j(W_j)$  be the density of  $W_j$  for  $j = 1, 2$  and  $f(W_1, W_2)$  be the joint density of  $(W_1, W_2)$ . Denote  $l^{(s)}(W_1) \equiv \frac{\nabla^s [a(W_1)f_1(W_1)]}{f_1(W_1)}$ . We impose the following assumption:

**Condition 2.1.2:** (i) The joint density  $f(W_1, W_2)$  of  $(W_1, W_2)$  is Hölder continuous with exponent greater than 1; (ii)  $\int \left[\frac{f(w_1, w_2)}{f_1(w_1)f_2(w_2)}\right]^2 dw_1 dw_2 < \infty$ ; (iii)  $[a(W_1)f_1(W_1)]$  is  $s$ -times continuously differentiable and is zero on the boundary of the support of  $W_1$ , (iv)  $E[\{l^{(s)}(W_1)\}^2] < \infty$ .

Condition 2.1.2(iii)(iv) and integration by parts yield

$$E\{D_{w^*}(X)' D_{w^*}(X)\} = E[\{\sum_{j=1}^2 w^{*j}(W_j)\}^2] + \left[1 + (-1)^s E\{l^{(s)}(W_1)w^{*1}(W_1)\}\right]^2.$$

First, we verify Assumption 4.1. Since  $E\{D_w(X)'D_w(X)\}$  is continuous and convex in  $w \in \overline{W}$  and  $\overline{W}$  is a closed linear space, the minimizer  $w^{*j}(W_j)$ ,  $j = 1, 2$  exists. Moreover  $E\{D_{w^*}(X)'D_{w^*}(X)\} > 0$ . This is because  $E\{D_{w^*}(X)'D_{w^*}(X)\} = 0$  if and only if  $E[\{\sum_{j=1}^2 w^{*j}(W_j)\}^2] = 0$  and  $1 + (-1)^s E[l^{(s)}(W_1)w^{*1}(W_1)] = 0$ , which could happen only when  $w^{*2}(W_2) = -w^{*1}(W_1)$  and  $w^{*1}(W_1) \neq 0$ , which is impossible since  $(W_1, W_2)$  has well-defined multivariate density that is not degenerate.

Next we verify Assumption 4.2'. By Remark 4.1(ii), since

$$\begin{aligned} & E \left\{ \left( \frac{dm(X, \alpha_*)}{dh} [w^* - \Pi_n w^*] \right)' \left( \frac{dm(X, \alpha_*)}{dh} [w^* - \Pi_n w^*] \right) \right\} \\ & \leq O \left( \sum_{j=1}^2 E[\{w^{*j}(W_j) - \Pi_n w^{*j}(W_j)\}^2] + E[\{l^{(s)}(W_1)\}^2] E[\{w^{*1}(W_1) - \Pi_n w^{*1}(W_1)\}^2] \right), \end{aligned}$$

Assumption 4.2' is satisfied provided  $\max_{j=1,2} \{E[\{w^{*j}(W_j) - \Pi_n w^{*j}(W_j)\}^2]\} = O(n^{-1/2})$ , which is satisfied if the solution  $w^{*j}(W_j)$ ,  $j = 1, 2$  is Hölder continuous with exponent greater than  $1/2$ . Equations (12)-(13) and integration by parts imply that  $w^{*j}(W_j)$ ,  $j = 1, 2$  solve

$$w^{*1}(W_1) + E[w^{*2}(W_2)|W_1] + (-1)^s \left[ 1 + (-1)^s E\{l^{(s)}(W_1)w^{*1}(W_1)\} \right] l^{(s)}(W_1) = 0, \quad (14)$$

$$w^{*2}(W_2) + E[w^{*1}(W_1)|W_2] = 0. \quad (15)$$

Let  $T$  be the conditional expectation operator of  $W_1$  given  $W_2$  (i.e.,  $Th_1 \equiv E[h_1(W_1)|W_2]$  for any measurable function  $h_1$  with  $E\{[h_1(W_1)]^2\} < \infty$ ), and  $T^*$  be the adjoint of  $T$  (i.e.,  $T^*h_2 \equiv E[h_2(W_2)|W_1]$  for any measurable function  $h_2$  with  $E\{[h_2(W_2)]^2\} < \infty$ ). Then  $(I - T^*T)^{-1}$  is a bounded operator, and (14)-(15) yield:  $w^{*2}(W_2) = -Tw^{*1}$  and

$$w^{*1}(W_1) = (-1)^{s+1} (I - T^*T)^{-1} l^{(s)}(W_1) \left( 1 + E\{(I - T^*T)^{-1} [l^{(s)}(W_1)]^2\} \right)^{-1}.$$

Condition 2.1.2 imply that  $w^{*1}(W_1)$  and  $w^{*2}(W_2)$  are smooth enough to satisfy Assumption 4.2'. Note that  $Var\{D_{w^*}(X)' \rho(Z, \alpha_*)\} =$

$$Var \left( \sum_{j=1}^2 w^{*j}(W_j) [Y - \sum_{j=1}^2 h_{*j}(W_j)] + [1 + E\{a(W_1) \nabla^s w^{*1}(W_1)\}] [\theta_* - a(W_1) \nabla^s h_{*1}(W_1)] \right).$$

Applying Theorem 4.1, we have  $\sqrt{n}(\hat{\theta}_n - \theta_*) \implies N(0, V^{-1})$  with  $V^{-1}$  given in (11), where

$$V^{-1} = Var \left( \frac{\{\sum_{j=1}^2 w^{*j}(W_j)\} [Y - \sum_{j=1}^2 h_{*j}(W_j)]}{1 + E\{a(W_1) \nabla^s w^{*1}(W_1)\}} + [\theta_* - a(W_1) \nabla^s h_{*1}(W_1)] \right). \quad (16)$$

We note that under correct specification  $E[Y - \sum_{j=1}^2 h_{oj}(W_j)|X_1] = 0$  and  $\alpha_* = \alpha_o$ , we have

$$V^{-1} = Var \left( \frac{\{\sum_{j=1}^2 w^{*j}(W_j)\} [Y - \sum_{j=1}^2 h_{*j}(W_j)]}{1 + E\{a(W_1) \nabla^s w^{*1}(W_1)\}} \right) + Var(\theta_* - a(W_1) \nabla^s h_{*1}(W_1)). \quad (17)$$

Under misspecification  $E[Y - \sum_{j=1}^2 h_{*j}(W_j)|X_1] \neq 0$  and we have non-zero correlation term:

$$E \left( [\theta_* - a(W_1) \nabla^s h_{*1}(W_1)] \{\sum_{j=1}^2 w^{*j}(W_j)\} [Y - \sum_{j=1}^2 h_{*j}(W_j)] \right) \neq 0. \quad (18)$$

The asymptotic variance of  $\widehat{\theta}$  under misspecification equals to the variance (17) plus some non-zero correlation term, where the non-zero correlation term arises from the model misspecification.

**Remark 4.3.** Consider the nonparametric LS regression with possibly omitted variable problem:  $h_{*1} = \arg \inf_{h_1 \in \mathcal{H}_1} E\{[E\{Y|X_1\} - h_1(W_1)]^2\}$ , where  $W_1$  is a subset of  $X_1$ . The correlation term in (18) is now zero even if  $E\{Y|X_1\} \neq h_{*1}(W_1)$ , and the asymptotic variance  $V^{-1}$  of  $\widehat{\theta}$  is

$$V^{-1} = Var \left( \frac{w^{*1}(W_1)[Y - h_{*1}(W_1)]}{1 + E\{a(W_1)\nabla^s w^{*1}(W_1)\}} \right) + Var(\theta_* - a(W_1)\nabla^s h_{*1}(W_1)). \quad (19)$$

Furthermore we can solve  $w^{*1}$  explicitly as:

$$w^{*1}(W_1) = \frac{(-1)^{s+1}l^{(s)}(W_1)}{1 + E\{[l^{(s)}(W_1)]^2\}}, \text{ thus } E\{D_{w^*}(X)'D_{w^*}(X)\} = \frac{1}{1 + E\{[l^{(s)}(W_1)]^2\}}.$$

Substituting  $w^{*1}(W_1)$  into (19) we obtain the asymptotic variance of  $\widehat{\theta}$ :

$$V^{-1} = E \left[ \left( \frac{\nabla^s[a(W_1)f_1(W_1)]}{f_1(W_1)} \right)^2 Var\{Y - h_{*1}(W_1)|X_1\} \right] + E \left[ \{\theta_* - a(W_1)\nabla^s h_{*1}(W_1)\}^2 \right].$$

The asymptotic variance for the special case of  $s = 1$  and  $W_1 = X_1$  coincides with the semiparametric efficient variance of the weighted average derivative estimator for  $\theta_o = E[a(W_1)\nabla h_{o1}(W_1)]$  (with  $h_{o1} = E[Y|W_1]$ ) derived in Newey and Stoker (1993, equation (3.8)).

## 4.2 Possibly misspecified nonparametric IV regression

Next, we apply Theorem 4.1 to **Example 2.2**. Recall that  $\rho_1(Z, \alpha) = Y_1 - h(Y_2)$ ,  $\rho_2(Z, \alpha) = \theta - a(Y_2)\nabla^s h(Y_2)$ ,  $m_1(X_1, \alpha) = E[Y_1 - h(Y_2)|X_1]$  and  $m_2(\alpha) = \theta - E\{a(Y_2)\nabla^s h(Y_2)\}$  since  $X_2$  is degenerate. Because the model is linear, Assumptions 4.3 - 4.4 are trivially satisfied and  $V_{w^*}(X) = 0$ . To apply Theorem 4.1 we need to verify Assumptions 4.1 and 4.2' where  $w^* \in \overline{W}$  solves the following minimization problem:

$$\inf_{w \in \overline{W}} E\{D_w(X)'D_w(X)\} = \inf_{w \in \overline{W}} \left\{ E[(E\{w(Y_2)|X_1\})^2] + (1 + E\{a(Y_2)\nabla^s w(Y_2)\})^2 \right\},$$

where  $\overline{W} = \{w : E[(E\{w(Y_2)|X_1\})^2] + (E\{a(Y_2)\nabla^s w(Y_2)\})^2 < \infty\}$ . By calculus variation,  $w^*(Y_2)$  solves

$$E[E\{w^*(Y_2)|X_1\}E\{\delta(Y_2)|X_1\}] + (1 + E\{a(Y_2)\nabla^s w^*(Y_2)\}) E\{a(Y_2)\nabla^s \delta(Y_2)\} = 0, \quad (20)$$

for all measurable functions  $\delta \in \overline{W}$ .

Let  $f(X_1, Y_2)$  denote the joint density of  $(X_1, Y_2)$ ,  $f_1(X_1)$  and  $f_2(Y_2)$  denote the marginal densities of  $X_1$  and  $Y_2$  respectively. Denote  $l^{(s)}(Y_2) \equiv \frac{\nabla^s[a(Y_2)f_2(Y_2)]}{f_2(Y_2)}$ . Without loss of generality, assume that  $p_1(X_1) = (p_{11}(X_1), p_{12}(X_1), \dots)$  are orthonormal basis functions satisfying:

$$E\{p_{1j}(X_1)^2\} = 1 \text{ for all } j \text{ and } E\{p_{1j}(X_1)p_{1k}(X_1)\} = 0 \text{ for all } j \neq k,$$

and that  $q(Y_2) = (q_1(Y_2), q_2(Y_2), \dots)$  are orthonormal basis functions satisfying

$$E\{q_j(Y_2)^2\} = 1 \text{ for all } j \text{ and } E\{q_j(Y_2)q_k(Y_2)\} = 0 \text{ for all } j \neq k.$$

Suppose that  $E\{q_j(Y_2)|X_1\} = p_{1j}(X_1)\rho_j$  where  $\rho_j$  denotes the  $j$ -th singular value. Suppose that  $l^{(s)}(Y_2)$  has the following series expansion

$$l^{(s)}(Y_2) = \sum_{j=1}^{\infty} \gamma_j q_j(Y_2), \text{ with coefficients satisfying } \sum_{j=1}^{\infty} \gamma_j^2 < \infty.$$

In addition to Condition 2.2.1, we impose the following assumption:

**Condition 2.2.2:** (i) for all  $j \geq 1$ ,  $\rho_j > 0$  and  $\sum_{j=1}^{\infty} \rho_j^2 < \infty$ ; (ii)  $[a(Y_2)f_2(Y_2)]$  is  $s$ -times continuously differentiable and is zero on the boundary of the support of  $Y_2$ , (iii)  $\sum_{j=1}^{\infty} \rho_j^{-2}\gamma_j^2 < \infty$ , (iv)

$$\sqrt{n} \sum_{j=k_{hn}}^{\infty} \rho_j^{-2}\gamma_j^2 < \infty, \text{ (v) } \sum_{j=1}^{\infty} \rho_j^{-4}\gamma_j^2 < \infty.$$

Under Conditions 2.2.1 and 2.2.2, we can show that

$$w^*(Y_2) = \sum_{j=1}^{\infty} \omega_j^* q_j(Y_2) \text{ with } \omega_j^* = (-1)^{s+1} \frac{\gamma_j}{\rho_j^2} \left[ 1 + \sum_{k=1}^{\infty} \frac{\gamma_k^2}{\rho_k^2} \right]^{-1} \text{ for all } j \geq 1 \quad (21)$$

solves the problem (20).<sup>4</sup> Furthermore,

$$E\{D_{w^*}(X)'D_{w^*}(X)\} = 1 + (-1)^s \sum_{j=1}^{\infty} \gamma_j \omega_j^* = \left[ 1 + \sum_{j=1}^{\infty} \frac{\gamma_j^2}{\rho_j^2} \right]^{-1}.$$

Thus, Assumption 4.1(i) is satisfied by Condition 2.2.2(iii). (Note that when  $Y_2$  is endogenous,  $w^* \in \overline{\mathcal{W}}$  is strictly weaker than the requirement of  $E\{[w^*(Y_2)]^2\} = \sum_{j=1}^{\infty} \omega_j^{*2} = \left( \sum_{j=1}^{\infty} \frac{\gamma_j^2}{\rho_j^4} \right) \left[ 1 + \sum_{k=1}^{\infty} \frac{\gamma_k^2}{\rho_k^2} \right]^{-2} < \infty$ . Nevertheless, we impose the stronger condition 2.2.2(v)  $\sum_{j=1}^{\infty} \rho_j^{-4}\gamma_j^2 < \infty$  so that it is easier to verify assumptions for Theorem 4.1.)

Next we verify Assumption 4.2'. By Remark 4.1(ii), since

$$\begin{aligned} & E \left\{ \left( \frac{dm(X, \alpha_*)}{dh} [w^* - \Pi_n w^*] \right)' \left( \frac{dm(X, \alpha_*)}{dh} [w^* - \Pi_n w^*] \right) \right\} \\ &= [E\{w^*(Y_2) - \Pi_n w^*(Y_2)|X_1\}]^2 + \left( E\{l^{(s)}(Y_2)[w^*(Y_2) - \Pi_n w^*(Y_2)]\} \right)^2 \\ &= \left\{ \sum_{j=k_{hn}}^{\infty} \frac{\gamma_j^2}{\rho_j^2} + \left( \sum_{j=k_{hn}}^{\infty} \frac{\gamma_j^2}{\rho_j^2} \right)^2 \right\} \left[ 1 + \sum_{k=1}^{\infty} \frac{\gamma_k^2}{\rho_k^2} \right]^{-2}, \end{aligned}$$

Assumption 4.2' is satisfied by Condition 2.2.2(iv).

<sup>4</sup>We are indebted to Whitney Newey who generously provides some insightful calculation that inspires this solution.

Applying Theorem 4.1, we have  $\sqrt{n}(\hat{\theta}_n - \theta_*) \implies N(0, V^{-1})$  with  $V^{-1} = [1 + E\{a(Y_2)\nabla^s w^*(Y_2)\}]^{-2}\Omega_*$ , where  $\Omega_*$  takes a complex form due to misspecification and endogeneity:

$$\Omega_* = Var \left\{ \begin{array}{c} [w^*(Y_2) - E\{w^*(Y_2)|X_1\}][E\{Y_1 - h_*(Y_2)|X_1\}] \\ + E\{w^*(Y_2)|X_1\}[Y_1 - h_*(Y_2)] \\ + [1 + E\{a(Y_2)\nabla^s w^*(Y_2)\}][\theta_* - a(Y_2)\nabla^s h_*(Y_2)] \end{array} \right\}.$$

Under correct specification  $E\{Y_1 - h_*(Y_2)|X_1\} = 0$  we have

$$\Omega_* = Var \{E\{w^*(Y_2)|X_1\}[Y_1 - h_*(Y_2)] + [1 + E\{a(Y_2)\nabla^s w^*(Y_2)\}][\theta_* - a(Y_2)\nabla^s h_*(Y_2)]\}.$$

Conditions 2.2.2(iii), (iv) and (v) impose smoothness restrictions on  $l^{(s)}(Y_2)$ . They may not be satisfied in some applications. If Condition 2.2.2(iii) is not satisfied, then we can not find a  $w^*(Y_2)$  with finite  $\|\cdot\|$ -norm such that  $E\{D_{w^*}(X)'D_{w^*}(X)\} > 0$ ; in this case,  $\theta_*$  can not be estimated at the  $\sqrt{n}$ -rate. The question is whether there exist some interesting models where Conditions 2.2.2(iii), (iv) and (v) are satisfied. To answer this question, we consider the special case  $l^{(s)}(Y_2) \equiv \nabla\{\log f_2(Y_2)\}$ . In this case, note that, if  $Y_2$  is normally distributed and  $q(Y_2)$  is power series,  $l^{(1)}(Y_2)$  is linear in  $Y_2$  and  $\gamma_k = 0$  for any  $k > 1$ . Thus, normally distributed regressor satisfies Conditions 2.2.2(iii)-(v) trivially. It is easy to show that the exponentially distributed regressor also satisfies this condition. Indeed, if  $f_2(Y_2) = const.\exp(t(Y_2))$  where  $t(Y_2)$  is a finite order polynomial, this condition is satisfied. If  $Y_2$  has a distribution that is not in this exponential family, we notice that  $\gamma_j^2$  is determined by the smoothness of  $l^{(s)}(Y_2)$ . This example demonstrates that it is not entirely impossible to obtain the root- $n$  consistent estimator.

Unlike Example 2.1, in Example 2.2 even when the model is correctly specified in the sense of  $E\{Y_1 - h_*(Y_2)|X_1\} = 0$ , due to nonparametric endogeneity, the asymptotic variance  $V^{-1}$  can not be simplified to:

$$V^{-1} = Var \left( \frac{E\{w^*(Y_2)|X_1\}[Y_1 - h_*(Y_2)]}{1 + E\{a(Y_2)\nabla^s w^*(Y_2)\}} \right) + Var(\theta_* - a(Y_2)\nabla^s h_*(Y_2)).$$

See Ai and Chen (2005) for semiparametric efficient estimation of (weighted) average derivatives of the nonparametric IV regression model.

## 5 Covariance Estimator

To estimate the covariance matrix  $V^{-1}$ , we estimate each of its components consistently. First, we estimate  $w^* = (w_1^*, \dots, w_{d_\theta}^*)$ . For  $l = 1, \dots, d_\theta$ , we estimate  $w_l^*$  by  $\hat{w}_l^*$ , which is the solution to the minimization problem:

$$\min_{w_l \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left\{ \begin{array}{c} \left( \frac{\partial \hat{m}(x_i, \hat{\alpha}_n)}{\partial \theta_l} - \frac{d \hat{m}(x_i, \hat{\alpha}_n)}{dh} [w_l] \right)' \left( \frac{\partial \hat{m}(x_i, \hat{\alpha}_n)}{\partial \theta_l} - \frac{d \hat{m}(x_i, \hat{\alpha}_n)}{dh} [w_l] \right) + \\ \left( \frac{\partial^2 \hat{m}(x_i, \hat{\alpha}_n)}{\partial \theta_l^2} - 2 \frac{d^2 \hat{m}(x_i, \hat{\alpha}_n)}{\partial \theta_l dh} [w_l] + \frac{d^2 \hat{m}(x_i, \hat{\alpha}_n)}{dh^2} [w_l, w_l] \right)' \hat{m}(x_i, \hat{\alpha}_n) \end{array} \right\}.$$

Notice that here we use the same sieve space  $\mathcal{H}_n$  to estimate  $w_l^*$ . This is for the purpose of simplifying notations only, and in practice many other finite-dimensional linear sieve spaces  $\mathcal{W}_n$  can be used to compute a consistent estimator for  $w_l^*$ . Denote  $\hat{w}^* = (\hat{w}_1^*, \dots, \hat{w}_{d_\theta}^*)$ . Then  $D_{w^*}(X)$  and  $V_{w^*}(X)$  are estimated respectively by

$$\begin{aligned}\hat{D}_{\hat{w}^*}(X) &= \frac{\partial \hat{m}(X, \hat{\alpha}_n)}{\partial \theta'} - \frac{d \hat{m}(X, \hat{\alpha}_n)}{dh} [\hat{w}^*]; \\ \hat{V}_{\hat{w}^*}(X) &= \sum_{j=1}^J \left( \frac{\partial^2 \hat{m}_j(X_j, \hat{\alpha}_n)}{\partial \theta \partial \theta'} - 2 \frac{d^2 \hat{m}_j(X_j, \hat{\alpha}_n)}{\partial \theta dh} [\hat{w}^*] + \frac{d^2 \hat{m}_j(X_j, \hat{\alpha}_n)}{dh^2} [\hat{w}^*, \hat{w}^*] \right)' \hat{m}_j(X_j, \hat{\alpha}_n).\end{aligned}$$

Next,  $\Omega_*$  is estimated by  $\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i \hat{\varepsilon}_i'$ , with

$$\hat{\varepsilon}_i = \left[ \frac{\partial \rho(z_i, \hat{\alpha}_n)}{\partial \theta'} - \frac{d \rho(z_i, \hat{\alpha}_n)}{dh} [\hat{w}^*] - \hat{D}_{\hat{w}^*}(x_i) \right]' \hat{m}(x_i, \hat{\alpha}_n) + \hat{D}_{\hat{w}^*}(x_i)' \rho(z_i, \hat{\alpha}_n).$$

The estimator of  $V^{-1}$  is

$$\hat{V}^{-1} \equiv \left( \frac{1}{n} \sum_{i=1}^n \{ \hat{D}_{\hat{w}^*}(x_i)' \hat{D}_{\hat{w}^*}(x_i) + \hat{V}_{\hat{w}^*}(x_i) \} \right)^{-1} \hat{\Omega} \left( \frac{1}{n} \sum_{i=1}^n \{ \hat{D}_{\hat{w}^*}(x_i)' \hat{D}_{\hat{w}^*}(x_i) + \hat{V}_{\hat{w}^*}(x_i) \} \right)^{-1}.$$

The above expressions are in compact forms. We can rewrite them in more detailed formats corresponding to the modified SMD procedure (5). First,  $\hat{w}^* = (\hat{w}_1^*, \dots, \hat{w}_{d_\theta}^*)$  is computed as the minimizer of

$$\min_{w_l \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left\{ \begin{aligned} & \sum_{j \in \mathcal{J}_{ex}} \left( \frac{\partial \rho_j(z_i, \hat{\alpha}_n)}{\partial \theta_l} - \frac{d \rho_j(z_i, \hat{\alpha}_n)}{dh} [w_l] \right)^2 + \\ & \sum_{j \in \mathcal{J}_{1en}} \left( \frac{\partial \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{\partial \theta_l} - \frac{d \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{dh} [w_l] \right)^2 + \sum_{j \in \mathcal{J}_{2en}} \left( \frac{\partial \hat{m}_j(\hat{\alpha}_n)}{\partial \theta_l} - \frac{d \hat{m}_j(\hat{\alpha}_n)}{dh} [w_l] \right)^2 + \\ & \sum_{j \in \mathcal{J}_{ex}} \left( \frac{\partial^2 \rho_j(z_i, \hat{\alpha}_n)}{\partial \theta_l^2} - 2 \frac{d^2 \rho_j(z_i, \hat{\alpha}_n)}{\partial \theta_l dh} [w_l] + \frac{d^2 \rho_j(z_i, \hat{\alpha}_n)}{dh^2} [w_l, w_l] \right) \rho_j(z_i, \hat{\alpha}_n) + \\ & \sum_{j \in \mathcal{J}_{1en}} \left( \frac{\partial^2 \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{\partial \theta_l^2} - 2 \frac{d^2 \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{\partial \theta_l dh} [w_l] + \frac{d^2 \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{dh^2} [w_l, w_l] \right) \hat{m}_j(x_{ji}, \hat{\alpha}_n) + \\ & \sum_{j \in \mathcal{J}_{2en}} \left( \frac{\partial^2 \hat{m}_j(\hat{\alpha}_n)}{\partial \theta_l^2} - 2 \frac{d^2 \hat{m}_j(\hat{\alpha}_n)}{\partial \theta_l dh} [w_l] + \frac{d^2 \hat{m}_j(\hat{\alpha}_n)}{dh^2} [w_l, w_l] \right) \hat{m}_j(\hat{\alpha}_n). \end{aligned} \right\}.$$

Second,  $E\{D_{w^*}(X)' D_{w^*}(X) + V_{w^*}(X)\}$  is estimated by  $\frac{1}{n} \sum_{i=1}^n \{ \hat{D}_{\hat{w}^*}(x_i)' \hat{D}_{\hat{w}^*}(x_i) + \hat{V}_{\hat{w}^*}(x_i) \} =$

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j \in \mathcal{J}_{ex}} \{ \hat{D}_{j\hat{w}^*}^{ex}(z_i) \}' \hat{D}_{j\hat{w}^*}^{ex}(z_i) + \sum_{j \in \mathcal{J}_{1en}} \{ \hat{D}_{j\hat{w}^*}^{1en}(x_{ji}) \}' \hat{D}_{j\hat{w}^*}^{1en}(x_{ji}) + \sum_{j \in \mathcal{J}_{2en}} \{ \hat{D}_{j\hat{w}^*}^{2en} \}' \hat{D}_{j\hat{w}^*}^{2en} \right\} \\ & + \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j \in \mathcal{J}_{ex}} \hat{V}_{j\hat{w}^*}^{ex}(z_i) + \sum_{j \in \mathcal{J}_{1en}} \hat{V}_{j\hat{w}^*}^{1en}(x_{ji}) + \sum_{j \in \mathcal{J}_{2en}} \hat{V}_{j\hat{w}^*}^{2en} \right\}, \end{aligned}$$

where for all  $j \in \mathcal{J}_{1en}$ ,

$$\begin{aligned}\hat{D}_{j\hat{w}^*}^{1en}(x_{ji}) &\equiv \frac{\partial \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{\partial \theta'} - \frac{d \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{dh} [\hat{w}^*], \\ \hat{V}_{j\hat{w}^*}^{1en}(x_{ji}) &\equiv \left( \frac{\partial^2 \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{\partial \theta \partial \theta'} - 2 \frac{d^2 \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{\partial \theta dh} [\hat{w}^*] + \frac{d^2 \hat{m}_j(x_{ji}, \hat{\alpha}_n)}{dh^2} [\hat{w}^*, \hat{w}^*] \right) \hat{m}_j(x_{ji}, \hat{\alpha}_n); \end{aligned}$$

$\widehat{D}_{j\widehat{w}^*}^{ex}(z_i)$ ,  $j \in \mathcal{J}_{ex}$ , and  $\widehat{D}_{j\widehat{w}^*}^{2en}$ ,  $j \in \mathcal{J}_{2en}$  are defined in the same way as  $\widehat{D}_{j\widehat{w}^*}^{1en}(x_{ji})$ , with  $\widehat{m}_j(x_{ji}, \widehat{\alpha}_n)$  replaced by  $\rho_j(z_i, \widehat{\alpha}_n)$ ,  $j \in \mathcal{J}_{ex}$ , and  $\widehat{m}_j(\widehat{\alpha}_n)$ ,  $j \in \mathcal{J}_{2en}$  respectively. Likewise,  $\widehat{V}_{j\widehat{w}^*}^{ex}(z_i)$ ,  $j \in \mathcal{J}_{ex}$ , and  $\widehat{V}_{j\widehat{w}^*}^{2en}$ ,  $j \in \mathcal{J}_{2en}$  are defined in the same way as  $\widehat{V}_{j\widehat{w}^*}^{1en}(x_{ji})$ , with  $\widehat{m}_j(x_{ji}, \widehat{\alpha}_n)$  replaced by  $\rho_j(z_i, \widehat{\alpha}_n)$ ,  $j \in \mathcal{J}_{ex}$ , and  $\widehat{m}_j(\widehat{\alpha}_n)$ ,  $j \in \mathcal{J}_{2en}$  respectively.

Finally,  $\Omega_*$  is estimated by  $\widehat{\Omega} = \frac{1}{n} \sum_{i=1}^n \widehat{\varepsilon}_i \widehat{\varepsilon}_i'$  with

$$\begin{aligned} \widehat{\varepsilon}_i &= \widehat{\varepsilon}_i^{ex} + \widehat{\varepsilon}_i^{1en} + \widehat{\varepsilon}_i^{2en}, \quad \text{where} \quad \widehat{\varepsilon}_i^{ex} \equiv \sum_{j \in \mathcal{J}_{ex}} \{\widehat{D}_{j\widehat{w}^*}^{ex}(z_i)\}' \rho_j(z_i, \widehat{\alpha}_n), \\ \widehat{\varepsilon}_i^{1en} &\equiv \sum_{j \in \mathcal{J}_{1en}} \left\{ \left\{ \frac{\partial \rho_j(z_i, \widehat{\alpha}_n)}{\partial \theta'} - \frac{d\rho_j(z_i, \widehat{\alpha}_n)}{dh} [\widehat{w}^*] - \widehat{D}_{j\widehat{w}^*}^{1en}(x_{ji})' \widehat{m}_j(x_{ji}, \widehat{\alpha}_n) + \{\widehat{D}_{j\widehat{w}^*}^{1en}(x_{ji})\}' \rho_j(z_i, \widehat{\alpha}_n) \right\}, \right. \\ \widehat{\varepsilon}_i^{2en} &\equiv \sum_{j \in \mathcal{J}_{2en}} \left\{ \left\{ \frac{\partial \rho_j(z_i, \widehat{\alpha}_n)}{\partial \theta'} - \frac{d\rho_j(z_i, \widehat{\alpha}_n)}{dh} [\widehat{w}^*] - \widehat{D}_{j\widehat{w}^*}^{2en} \}' \widehat{m}_j(\widehat{\alpha}_n) + \{\widehat{D}_{j\widehat{w}^*}^{2en}\}' \rho_j(z_i, \widehat{\alpha}_n) \right\}. \end{aligned}$$

In the Appendix, we show that the following additional conditions are sufficient for  $\widehat{V}^{-1}$  to be a consistent estimator of  $V^{-1}$ .

**Assumption 5.1.** For all  $j$  and each component  $\theta_l$ ,  $l = 1, \dots, d_\theta$ ,  $\frac{d\rho_j(Z, \alpha)}{d\theta_l} - \frac{d\rho_j(Z, \alpha)}{dh} [w_l]$  satisfies an envelope condition and is Hölder continuous in  $\alpha \in \mathcal{N}_o$  and  $w_l \in \{v \in \overline{\mathcal{W}} : \|v\|_s \leq c < \infty\}$ .

**Assumption 5.2.** For all  $j$  and each component  $\theta_l$ ,  $l = 1, \dots, d_\theta$ ,  $\frac{\partial^2 \rho_j(Z, \alpha)}{\partial \theta_l^2} - 2 \frac{d^2 \rho_j(Z, \alpha)}{\partial \theta_l dh} [w_l] + \frac{d^2 \rho_j(Z, \alpha)}{dh^2} [w_l, w_l]$  satisfies an envelope condition and is Hölder continuous in  $\alpha \in \mathcal{N}_o$  and  $w_l \in \{v \in \overline{\mathcal{W}} : \|v\|_s \leq c < \infty\}$ .

**Theorem 5.1.** Under Assumptions 3.1 - 3.8, 4.1 - 4.6, and 5.1 - 5.2, we have:  $\widehat{V}^{-1} = V^{-1} + o_p(1)$ .

## 6 Conclusion

In this paper, we propose a modified SMD estimation method for a general class of conditional moment restriction models in which different equations may require different conditioning variables. We derive the asymptotic results of the modified SMD estimator without imposing the correct specification of the conditional moment restrictions. Under mild and low-level sufficient conditions, we show that the SMD estimator converges in probability to some pseudo-true value that minimizes the population objective function and that the SMD estimator for any smooth functional is  $\sqrt{n}$ -asymptotically normally distributed. We also provide a simple consistent covariance estimator for the SMD estimate of any smooth functional. These results allow researchers to conduct asymptotically valid inferences on the smooth functionals regardless of whether the semiparametric conditional model is correctly specified or not. As illustration, we apply our general theory to two non-trivial yet popular examples: a weighted average derivative estimate of a possibly misspecified nonparametric additive Least Squares (LS) regression, and a weighted average derivative estimate of a possibly misspecified nonparametric IV regression.

We are currently working on several closely related projects. The first project considers model selection tests when all the competing semiparametric conditional moment models (1) could be misspecified. The second project relaxes the pointwise Hölder continuity assumption of  $\rho_j(Z, \alpha)$  in  $\alpha$ . This can be done by modifying our current proof using the results in Chen, Linton and van Keilegom (2003). The third project extends the results of this paper from i.i.d. data to stationary beta-mixing time series data; such an extension is needed when we study possibly misspecified semiparametric asset pricing and financial time series models. In fact, in their estimation of semiparametric ARCH( $\infty$ ) models, Linton and Mammen (2005) discuss a class of *weak* form ARCH( $\infty$ ) models that could be misspecified. Finally, we plan to investigate the use of nonparametric bootstrap to provide an asymptotically valid confidence region for the SMD estimate of any smooth functional of  $\alpha_*$ .<sup>5</sup> Recently Nishiyama and Robinson (2005) establish the bootstrap refinement of the average derivative estimate for the nonparametric LS regression model. It is worthwhile to see how bootstrap procedure performs when the semiparametric conditional moment models (1) could be misspecified.

## Mathematical Appendix

Recall that  $\mathcal{A}_n$  denotes a sieve approximation of  $\mathcal{A}$ ,  $N(\varepsilon, \mathcal{A}_n, \|\cdot\|_s)$  denotes the minimal number of  $\varepsilon$ -radius covering balls of  $\mathcal{A}_n$  under the metric  $\|\cdot\|_s$ . In the following lemma,  $\varepsilon(Z, \alpha) : \mathcal{Z} \times \mathcal{A} \rightarrow \mathcal{R}$  denotes a generic measurable function of the data  $Z \in \mathcal{Z}$  and the parameter  $\alpha \in \mathcal{A}$  and satisfies  $E\{\varepsilon(Z, \alpha)|X\} = 0$  for all  $X$  and all  $\alpha$ . Let  $\{Z_1, \dots, Z_n\}$  denote an i.i.d. sample. Let  $g_i(X_1, \dots, X_n, \alpha)$  denote some function satisfying for all  $\{X_1, \dots, X_n\}$ :

$$\sup_{\alpha \in \mathcal{A}_n, 1 \leq i \leq n} |g_i(X_1, \dots, X_n, \alpha)| = O_p(\delta_n);$$

$$\sup_{\alpha, \alpha' \in \mathcal{A}_n, 1 \leq i \leq n} |g_i(X_1, \dots, X_n, \alpha) - g_i(X_1, \dots, X_n, \alpha')| = O_p(\|\alpha - \alpha'\|_s^\kappa).$$

The following lemma is a modification of Lemma A.1 of Ai and Chen (2003).

**Lemma A.1:** *Suppose that the followings are satisfied:*

(i) *there exist a constant  $c_{1n}$  and a measurable function  $c_1(Z) : \mathcal{Z} \rightarrow [0, \infty)$  with  $E[c_1(Z)^p] < \infty$  for some  $p \geq 2$  such that  $|\varepsilon(Z, \alpha)| \leq c_{1n}c_1(Z)$  for all  $\alpha \in \mathcal{A}_n$  and  $Z \in \mathcal{Z}$ ;*

(ii) *there exist a constant  $\kappa \in (0, 1]$  and a measurable function  $c_2(Z) : \mathcal{Z} \rightarrow [0, \infty)$  with  $E[c_2(Z)] < \infty$  such that  $|\varepsilon(Z, \alpha_1) - \varepsilon(Z, \alpha_2)| \leq c_{2n}c_2(Z) \|\alpha_1 - \alpha_2\|_s^\kappa$  holds for all  $Z \in \mathcal{Z}$  and  $\alpha_1, \alpha_2 \in \mathcal{A}_n$ ;*

---

<sup>5</sup>We thank Oliver Linton for suggesting the bootstrap alternative.

(iii) Let  $\delta_{1n} = o(1)$  and  $\delta_{1n} = o(\delta_n)$  be such that

$$\frac{n\delta_{1n}^2}{\ln\left(N(\{\min\{\frac{\delta_{1n}}{c_{2n}\delta_n}, \frac{\delta_{1n}}{c_{1n}}\}\}^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_s)\max\{c_{1n}^2\delta_n^2, (c_{1n}\delta_n)^{1+2/p}\delta_{1n}^{1-(2/p)}\}\right)} \rightarrow +\infty.$$

Then:  $\frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha) \varepsilon(Z_i, \alpha) = o_p(\delta_{1n})$  uniformly over  $\alpha \in \mathcal{A}_n$ .

When applying Lemma A.1 in this Appendix, we typically have  $p = 2$ ,  $c_{1n} = 1$ ,  $c_{2n} = 1$  and either **(a)**  $\delta_n = O(1)$ ,  $\delta_{1n} = o(1)$  and condition (iii)  $\ln\left(N(\{\delta_{1n}\}^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_s)\right) \times n^{-1} \rightarrow 0$ , or **(b)**  $\delta_n = O(1)$ ,  $\delta_{1n} = n^{-1/4}$  and condition (iii)  $\ln\left(N(\{\delta_{1n}\}^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_s)\right) \times n^{-1/2} \rightarrow 0$ , or **(c)**  $\delta_n = o(n^{-1/4})$ ,  $\delta_{1n} = n^{-1/2}$  and condition (iii)  $\ln\left(N(\{\delta_{1n}\}^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_s)\right) \times n^{-1/2} \rightarrow 0$ .

**Proof. (Lemma A.1)** Let  $c$  denote a generic constant which may have different values in different expressions. For any  $\alpha, \alpha' \in \mathcal{A}_n$ , we write

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha) \varepsilon(Z_i, \alpha) - \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha') \varepsilon(Z_i, \alpha') \right| \\ & \leq \frac{1}{n} \sum_{i=1}^n |g_i(X_1, \dots, X_n, \alpha)| \times |\varepsilon(Z_i, \alpha) - \varepsilon(Z_i, \alpha')| + \\ & \quad \frac{1}{n} \sum_{i=1}^n |g_i(X_1, \dots, X_n, \alpha) - g_i(X_1, \dots, X_n, \alpha')| \times |\varepsilon(Z_i, \alpha')| \\ & \leq O(c_{2n}\delta_n) \|\alpha - \alpha'\|_s^\kappa \frac{1}{n} \sum_{i=1}^n c_2(Z_i) + c_{1n} O_p(\|\alpha - \alpha'\|_s^\kappa) \frac{1}{n} \sum_{i=1}^n c_1(Z_i) \end{aligned}$$

by conditions (i), (ii) and the condition on  $g_i$ . Notice that  $\frac{1}{n} \sum_{i=1}^n c_2(Z_i) = O_p(1)$  and that  $\frac{1}{n} \sum_{i=1}^n c_1(Z_i) = O_p(1)$ . There exists a constant  $c$  such that:

$$P\left(\sup_{\alpha, \alpha' \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha) \varepsilon(Z_i, \alpha) - \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha') \varepsilon(Z_i, \alpha') \right| > c(c_{2n}\delta_n + c_{1n}) \|\alpha - \alpha'\|_s^\kappa\right) < \eta$$

for sufficiently large  $n$  and any small  $\eta$ .

For any small  $\epsilon$ , partition  $\mathcal{A}_n$  into  $b_n$  mutually exclusive subsets  $\mathcal{A}_{nm}$  for  $m = 1, 2, \dots, b_n$ , where  $\alpha, \alpha' \in \mathcal{A}_{nm}$  satisfy  $\|\alpha - \alpha'\|_s^\kappa \leq \epsilon \times \min\{\frac{\delta_{1n}}{c_{2n}\delta_n}, \frac{\delta_{1n}}{c_{1n}}\}$ . Then with probability approaching one,

$$\left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha) \varepsilon(Z_i, \alpha) - \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha') \varepsilon(Z_i, \alpha') \right| \leq \epsilon \delta_{1n}.$$

Let  $\alpha^m$  denote a fixed point in  $\mathcal{A}_{nm}$ . For any  $\alpha \in \mathcal{A}_n$ , there exists a  $m \in \{1, \dots, b_n\}$  such that  $\|\alpha - \alpha^m\|_s^\kappa \leq \epsilon \times \min\{\frac{\delta_{1n}}{c_{2n}\delta_n}, \frac{\delta_{1n}}{c_{1n}}\}$ . Then, with probability approaching one,

$$\sup_{\alpha \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha) \varepsilon(Z_i, \alpha) \right| \leq \epsilon \delta_{1n} + \max_m \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon(Z_i, \alpha^m) \right|.$$

Hence

$$\begin{aligned} & P \left( \sup_{\alpha \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha) \varepsilon(Z_i, \alpha) \right| > 2\epsilon\delta_{1n} \right) \\ & < \eta + P \left( \max_m \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon(Z_i, \alpha^m) \right| > \epsilon\delta_{1n} \right). \end{aligned}$$

For some constant  $c > 0$ , let  $M_n = \left( \frac{c\delta_n c_{1n}}{\delta_{1n} \epsilon \eta} \right)^{2/p}$ . Define  $d_{in} = 1 \{c_1(Z_i) \leq M_n\}$ . Define  $\varepsilon_1(Z_i, \alpha) = d_{in} \varepsilon(Z_i, \alpha)$  and  $\varepsilon_2(Z_i, \alpha) = (1 - d_{in}) \varepsilon(Z_i, \alpha)$ . It follows that

$$\begin{aligned} & P \left( \max_m \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon(Z_i, \alpha^m) \right| > \epsilon\delta_{1n} \right) \\ & \leq P \left( \max_m \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_1(Z_i, \alpha^m) \right| > \epsilon\delta_{1n} \right) \\ & \quad + P \left( \max_m \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_2(Z_i, \alpha^m) \right| > \epsilon\delta_{1n} \right) \equiv P_1 + P_2. \end{aligned}$$

Applying the Markov inequality yields

$$\begin{aligned} P_2 & \leq \frac{E \left[ \max_m \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_2(Z_i, \alpha^m) \right| \right]}{\epsilon\delta_{1n}} \leq \delta_n c_{1n} \frac{E \left[ \frac{1}{n} \sum_{i=1}^n (1 - d_{in}) c_1(Z_i) \right]}{\epsilon\delta_{1n}} \\ & \leq \delta_n c_{1n} \frac{\sqrt{E[(1 - d_{in})]} \sqrt{E[c_1(Z_i)^2]}}{\epsilon\delta_{1n}} \leq \delta_n c_{1n} \frac{1}{M_n^{p/2} \epsilon\delta_{1n}} \leq \eta. \end{aligned}$$

Some calculations yield

$$\sigma_m^2 \equiv n \times E \left\{ \left[ \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_1(Z_i, \alpha^m) \right]^2 \right\} = O(c_{1n}^2 \delta_n^2).$$

and  $|g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_1(Z_i, \alpha^m)| \leq \delta_n c_{1n} M_n$ . Note that

$$\begin{aligned} & P \left( \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_1(Z_i, \alpha^m) \right| > \epsilon\delta_{1n} \right) = \\ & E \left[ P \left( \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_1(Z_i, \alpha^m) \right| > \epsilon\delta_{1n} \mid X_1, \dots, X_n \right) \right]. \end{aligned}$$

Applying the Bernstein inequality for independent processes, we obtain:

$$P \left( \left| \frac{1}{n} \sum_{i=1}^n g_i(X_1, \dots, X_n, \alpha^m) \varepsilon_1(Z_i, \alpha^m) \right| > \epsilon\delta_{1n} \right) \leq 2 \exp \left( - \frac{n\epsilon^2 \delta_{1n}^2}{4[cc_{1n}^2 \delta_n^2 + \epsilon\delta_{1n} \delta_n c_{1n} M_n]} \right).$$

Hence,

$$P_1 < 2b_n \exp \left( - \frac{n\epsilon^2 \delta_{1n}^2}{4[cc_{1n}^2 \delta_n^2 + \epsilon\delta_{1n} \delta_n c_{1n} M_n]} \right),$$

which is arbitrarily small if

$$\frac{n\epsilon^2 \delta_{1n}^2}{4[cc_{1n}^2 \delta_n^2 + \epsilon\delta_{1n} \delta_n c_{1n} M_n]} - \ln(b_n) = \ln(b_n) \left\{ \frac{n\epsilon^2 \delta_{1n}^2}{\ln(b_n) \times 4[cc_{1n}^2 \delta_n^2 + \epsilon\delta_{1n} \delta_n c_{1n} M_n]} - 1 \right\}$$

is a big positive number. Notice that

$$b_n = O\left(N(\{\min\{\frac{\delta_{1n}}{c_{2n}\delta_n}, \frac{\delta_{1n}}{c_{1n}}\}\}^{1/\kappa}, \mathcal{A}_n, \|\cdot\|_s)\right) \quad \text{and} \quad M_n = \left(\frac{c\delta_n c_{1n}}{\delta_{1n}\epsilon\eta}\right)^{2/p}.$$

Substituting for  $b_n$  and  $M_n$ , we obtain that  $P_1$  is arbitrarily small when condition (iii) holds. ■

**Proof. (Theorem 3.1):** Denote

$$\begin{aligned} \widehat{L}_n(\alpha) &\equiv \frac{-1}{2n} \sum_{i=1}^n \left( \sum_{j \in \mathcal{J}_{ex}} \rho_j(z_i, \alpha)^2 + \sum_{j \in \mathcal{J}_{1en}} \widehat{m}_j(x_{ji}, \alpha)^2 + \sum_{j \in \mathcal{J}_{2en}} \widehat{m}_j(\alpha)^2 \right); \\ L_n(\alpha) &\equiv \frac{-1}{2n} \sum_{i=1}^n \left( \sum_{j \in \mathcal{J}_{ex}} \rho_j(z_i, \alpha)^2 + \sum_{j \in \mathcal{J}_{1en}} m_j(x_{ji}, \alpha)^2 + \sum_{j \in \mathcal{J}_{2en}} m_j(\alpha)^2 \right). \end{aligned}$$

Given Lemma 3.1  $\|\widehat{\alpha}_n - \alpha_*\|_s = o_p(1)$ , we can now restrict our attention to the parameter space  $\mathcal{A}_{os} = \{\alpha \in \mathcal{A} : \|\alpha - \alpha_*\|_s = o_p(1), \|\alpha\|_s \leq c\}$  and its sieve space  $\mathcal{A}_{osn} = \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_*\|_s = o_p(1), \|\alpha\|_s \leq c\}$ . Under Assumptions 3.1, 3.2, 3.5, 3.6 and 3.7, Corollary A.1(i) of Ai and Chen (2003) is still applicable with their  $\mathcal{A}_n$  replaced by our  $\mathcal{A}_{osn}$ , hence we obtain:

$$\frac{1}{n} \sum_{i=1}^n (\widehat{m}_j(x_{ji}, \alpha) - m_j(x_{ji}, \alpha))^2 = o_p(n^{-1/2}) \text{ uniformly over } \alpha \in \mathcal{A}_{osn} \text{ for } j \in \mathcal{J}_{1en}. \quad (22)$$

By Assumption 3.5(iv), uniformly over  $\alpha \in \mathcal{A}_{osn}$  and for all  $j \in \mathcal{J}_{1en}$ ,  $m_j(x_j, \alpha)$  is bounded in  $x_j$ .

Since for all  $\alpha \in \mathcal{A}_{osn}$ , Assumption 3.5(ii) implies

$$|\rho_j(z_i, \alpha)| \leq |\rho_j(z_i, \alpha) - \rho_j(z_i, \alpha_*)| + |\rho_j(z_i, \alpha_*)| \leq c \times c_2(z_i) + |\rho_j(z_i, \alpha_*)| \text{ for all } j.$$

Under Assumptions 3.1(i), 3.5(i)(ii) and 3.7, we can apply Lemma A.1(b) ( $\delta_n = O(1)$ ,  $\delta_{1n} = n^{-1/4}$ ) and obtain:

$$\frac{1}{n} \sum_{i=1}^n \rho_j(z_i, \alpha) - E\{\rho_j(Z, \alpha)\} = o_p(n^{-1/4}) \text{ uniformly over } \alpha \in \mathcal{A}_{osn} \text{ for } j \in \mathcal{J}_{2en}, \quad (23)$$

and  $m_j(\alpha) = E\{\rho_j(Z, \alpha)\}$  is bounded uniformly in  $\alpha \in \mathcal{A}_{osn}$  for  $j \in \mathcal{J}_{2en}$ .

Applying (22) and (23), we have

$$\begin{aligned} &\widehat{L}_n(\alpha) - L_n(\alpha) \\ &= \frac{-1}{n} \sum_{i=1}^n \left( \sum_{j \in \mathcal{J}_{1en}} m_j(x_{ji}, \alpha) [\widehat{m}_j(x_{ji}, \alpha) - m_j(x_{ji}, \alpha)] + \sum_{j \in \mathcal{J}_{2en}} m_j(\alpha) [\widehat{m}_j(\alpha) - m_j(\alpha)] \right) \\ &\quad + o_p(n^{-1/2}) \text{ uniformly over } \alpha \text{ in } \mathcal{A}_{osn}. \end{aligned}$$

For  $j \in \mathcal{J}_{1en}$ , let  $\widetilde{m}_j(x_{ji}, \alpha)$  denote the fitted value of regressing  $m_j(x_{ji}, \alpha)$  on  $p_j^{k_{jn}}(x_{ji})$ ,  $i = 1, 2, \dots, n$ . Note that  $\widehat{m}_j(x_{ji}, \alpha)$  are the fitted values of regressing  $\rho_j(z_i, \alpha)$ , not  $m_j(x_{ji}, \alpha)$ , on

$p_j^{k_{jn}}(x_{ji}), i = 1, 2, \dots, n$ . Hence for all  $j \in \mathcal{J}_{1en}$ , we have  $\frac{1}{n} \sum_{i=1}^n \tilde{m}_j(x_{ji}, \alpha)[m_j(x_{ji}, \alpha) - \tilde{m}_j(x_{ji}, \alpha)] = 0$  uniformly over  $\alpha \in \mathcal{A}_{osn}$ , and

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n m_j(x_{ji}, \alpha)[\hat{m}_j(x_{ji}, \alpha) - m_j(x_{ji}, \alpha)] \\
&= \frac{1}{n} \sum_{i=1}^n m_j(x_{ji}, \alpha)[\hat{m}_j(x_{ji}, \alpha) - \tilde{m}_j(x_{ji}, \alpha)] + \frac{1}{n} \sum_{i=1}^n m_j(x_{ji}, \alpha)[\tilde{m}_j(x_{ji}, \alpha) - m_j(x_{ji}, \alpha)] \\
&= \frac{1}{n} \sum_{i=1}^n \tilde{m}_j(x_{ji}, \alpha)[\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha)] - \frac{1}{n} \sum_{i=1}^n [m_j(x_{ji}, \alpha) - \tilde{m}_j(x_{ji}, \alpha)]^2 \\
&= \frac{1}{n} \sum_{i=1}^n \tilde{m}_j(x_{ji}, \alpha)[\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha)] + o_p(n^{-1/2}) \text{ uniformly over } \alpha \in \mathcal{A}_{osn},
\end{aligned}$$

where the last equality is due to Assumptions 3.2 (iii) and 3.5(iv), the approximation errors of  $m_j(X_j, \alpha)$  by the basis functions  $p_j^{k_{jn}}(X_j)$  is  $O(k_{jn}^{-\gamma_j/d_{x_j}}) = o(n^{-1/4})$ . By Lemma A.1(c) ( $\delta_n = o(n^{-1/4})$ ,  $\delta_{1n} = n^{-1/2}$ ) we have for  $j \in \mathcal{J}_{1en}$ ,

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \tilde{m}_j(x_{ji}, \alpha)[\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha)] \\
&= \frac{1}{n} \sum_{i=1}^n m_j(x_{ji}, \alpha)[\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha)] + o_p(n^{-1/2}) \text{ uniformly over } \alpha \in \mathcal{A}_{osn}.
\end{aligned}$$

Hence

$$\begin{aligned}
& \hat{L}_n(\alpha) - L_n(\alpha) \\
&= \frac{-1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{J}_{1en}} m_j(x_{ji}, \alpha)[\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha)] - \frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{J}_{2en}} m_j(\alpha)[\rho_j(z_i, \alpha) - m_j(\alpha)] \\
&\quad + o_p(n^{-1/2}) \text{ uniformly over } \alpha \text{ in } \mathcal{A}_{osn}.
\end{aligned}$$

Recall that  $\tilde{L}_n(\alpha) \equiv \frac{1}{2n} \sum_{i=1}^n \ell(z_i, \alpha)$  with  $\ell(z_i, \alpha)$  defined in (8). Then

$$\begin{aligned}
\hat{L}_n(\alpha) &= L_n(\alpha) - \frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{J}_{1en}} m_j(x_{ji}, \alpha)[\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha)] \\
&\quad - \frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{J}_{2en}} m_j(\alpha)[\rho_j(z_i, \alpha) - m_j(\alpha)] + o_p(n^{-1/2}) \\
&= \tilde{L}_n(\alpha) + o_p(n^{-1/2}) \text{ uniformly over } \alpha \text{ in } \mathcal{A}_{osn}.
\end{aligned}$$

Similarly we have  $\hat{L}_n(\alpha_*) - \tilde{L}_n(\alpha_*) = o_p(n^{-1/2})$ . Hence

$$\hat{L}_n(\alpha) - \hat{L}_n(\alpha_*) - \{\tilde{L}_n(\alpha) - \tilde{L}_n(\alpha_*)\} = o_p(n^{-1/2}) \text{ uniformly over } \alpha \text{ in } \mathcal{A}_{osn},$$

and  $\hat{\alpha}_n$  is the approximate maximizer of  $\tilde{L}_n(\alpha)$  over  $\alpha$  in  $\mathcal{A}_{osn}$ :

$$\{\tilde{L}_n(\hat{\alpha}_n) - \tilde{L}_n(\alpha_*)\} \geq \max_{\alpha \in \mathcal{A}_{osn}} \{\tilde{L}_n(\alpha) - \tilde{L}_n(\alpha_*)\} - \eta_n \quad \text{with } \eta_n = o_p(n^{-1/2}).$$

Note that  $\alpha_* = \arg \sup_{\alpha \in \mathcal{A}} E\{\tilde{L}_n(\alpha)\}$ . Under Assumption 3.8(iii),  $E\{\tilde{L}_n(\alpha_*) - \tilde{L}_n(\alpha)\} \geq c\|\alpha - \alpha_*\|^2$  for all  $\alpha \in \mathcal{A}_{osn}$ . Also, for any  $\alpha, \alpha' \in \mathcal{A}_{osn}$ ,

$$\begin{aligned} & \ell(z_i, \alpha) - \ell(z_i, \alpha') \\ &= \sum_{j \in \mathcal{J}_{ex}} \{\rho_j(z_i, \alpha') - \rho_j(z_i, \alpha)\} \{\rho_j(z_i, \alpha') + \rho_j(z_i, \alpha)\} \\ & \quad + 2 \sum_{j \in \mathcal{J}_{en}} [m_j(x_{ji}, \alpha) \{\rho_j(z_i, \alpha') - \rho_j(z_i, \alpha)\} + \{m_j(x_{ji}, \alpha') - m_j(x_{ji}, \alpha)\} \rho_j(z_i, \alpha')] \\ & \quad - \sum_{j \in \mathcal{J}_{en}} \{m_j(x_{ji}, \alpha') - m_j(x_{ji}, \alpha)\} \{m_j(x_{ji}, \alpha') + m_j(x_{ji}, \alpha)\}. \end{aligned}$$

Recall that for  $j \in \mathcal{J}_{ex}$  we have  $\rho_j(z_i, \alpha) - \rho_j(z_i, \alpha_*) = m_j(x_{ji}, \alpha) - (x_{ji}, \alpha_*)$ , which is a measurable function of  $x_{ji}$  only. Under Assumptions 3.5(ii)(iii), we have for all  $\alpha \in \mathcal{A}_{osn}$ ,

$$\begin{aligned} |\rho_j(z_i, \alpha)| &\leq |\rho_j(z_i, \alpha) - \rho_j(z_i, \alpha_*)| + |\rho_j(z_i, \alpha_*)| \leq c \times c_2(x_{ji}) + |\rho_j(z_i, \alpha_*)| \quad \text{for } j \in \mathcal{J}_{ex}, \\ |\rho_j(z_i, \alpha)| &\leq \min\{c_1(z_i), c \times c_2(z_i) + |\rho_j(z_i, \alpha_*)|\} \quad \text{for } j \in \mathcal{J}_{1en}, \\ |\rho_j(z_i, \alpha)| &\leq c \times c_2(z_i) + |\rho_j(z_i, \alpha_*)| \quad \text{for } j \in \mathcal{J}_{2en}. \end{aligned}$$

Thus, under Assumptions 3.5(i)(ii)(iii), there is a function  $b(z_i)$  with  $E\{[b(z_i)]^2\} < \infty$  such that for any  $\alpha, \alpha' \in \mathcal{A}_{osn}$ ,

$$\begin{aligned} & |\ell(z_i, \alpha) - \ell(z_i, \alpha')| \\ &\leq \sum_{j \in \mathcal{J}_{ex}} |\rho_j(z_i, \alpha') - \rho_j(z_i, \alpha)| \times [|\rho_j(z_i, \alpha')| + |\rho_j(z_i, \alpha)|] \\ & \quad + 2 \sum_{j \in \mathcal{J}_{en}} [|m_j(x_{ji}, \alpha)| |\rho_j(z_i, \alpha') - \rho_j(z_i, \alpha)| + |m_j(x_{ji}, \alpha') - m_j(x_{ji}, \alpha)| |\rho_j(z_i, \alpha')|] \\ & \quad + \sum_{j \in \mathcal{J}_{en}} |m_j(x_{ji}, \alpha') - m_j(x_{ji}, \alpha)| \times [|m_j(x_{ji}, \alpha')| + |m_j(x_{ji}, \alpha)|] \\ &\leq b(z_i) \times \|\alpha - \alpha'\|_s^\kappa. \end{aligned}$$

Let  $\mathcal{F}_n = \{\ell(z_i, \alpha) - \ell(z_i, \alpha_*) : \alpha \in \mathcal{A}_{osn}\}$ . Then  $N_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_{L_2(P)}) \leq N(\{\varepsilon\}^{1/\kappa}, \mathcal{A}_{osn}, \|\cdot\|_s)$ , where  $N_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_{L_2(P)})$  denotes the minimal number of  $\varepsilon$ -radius covering brackets of  $\mathcal{F}_n$  under the mean square metric  $\|\cdot\|_{L_2(P)}$ . The rest of the proof of Theorem 3.1 follows from applying Theorem 1 of Chen and Shen (1998) (or the simpler i.i.d. version of theorem 3.1 in Chen, 2005) to  $\tilde{L}_n(\alpha)$  over  $\alpha$  in  $\mathcal{A}_{osn}$ . ■

**Proof. (Theorem 4.1)** Recall that the neighborhood  $\mathcal{N}_{on} = \{\alpha \in \mathcal{A}_{osn} : \|\alpha - \alpha_*\| = o(n^{-1/4})\}$ . Let  $\varepsilon_n > 0$  be at the order of  $o(n^{-1/2})$ . With  $v^*$  given in Section 4, denote  $u^* = v^*$  and  $u_n^* = v_n^*$ . Denote  $\alpha(t) = \hat{\alpha} + t\varepsilon_n u_n^*$ . By Assumption 3.8(ii),  $\hat{L}_n(\alpha(t))$  is twice continuously differentiable with respect to  $t$ . By definition of  $\hat{\alpha} = \hat{\alpha}_n$  and taking a second order Taylor expansion of  $\hat{L}_n(\alpha(t))$  around  $t = 0$ , we have

$$\begin{aligned} 0 &\leq \hat{L}_n(\hat{\alpha}) - \hat{L}_n(\hat{\alpha} + \varepsilon_n u_n^*) = \hat{L}_n(\alpha(0)) - \hat{L}_n(\alpha(1)) \\ &= -\frac{d\hat{L}_n(\alpha(t))}{dt}\Big|_{t=0} - \frac{1}{2} \frac{d^2\hat{L}_n(\alpha(t))}{dt^2}\Big|_{t=s} \quad \text{for some } s \in [0, 1]. \end{aligned}$$

For  $j \in \mathcal{J}_{1en}$ , denote  $\frac{d\widehat{m}_j(X_j, \alpha(\tau))}{d\alpha}[\varepsilon_n u_n^*] = \frac{d\widehat{m}_j(X_j, \alpha(t))}{dt}|_{t=\tau}$  and  $\frac{d^2\widehat{m}_j(X_j, \alpha(s))}{d\alpha d\alpha}[\varepsilon_n u_n^*, \varepsilon_n u_n^*] = \frac{d^2\widehat{m}_j(X_j, \alpha(t))}{dt^2}|_{t=s}$ . Define  $\frac{d\rho_j(Z, \alpha(\tau))}{d\alpha}[\varepsilon_n u_n^*]$  and  $\frac{d^2\rho_j(Z, \alpha(\tau))}{d\alpha d\alpha}[\varepsilon_n u_n^*, \varepsilon_n u_n^*]$  for  $j \in \mathcal{J}_{ex}$ , and  $\frac{d\widehat{m}_j(\alpha(\tau))}{d\alpha}[\varepsilon_n u_n^*]$  and  $\frac{d^2\widehat{m}_j(\alpha(s))}{d\alpha d\alpha}[\varepsilon_n u_n^*, \varepsilon_n u_n^*]$  for  $j \in \mathcal{J}_{2en}$  analogously. Hence

$$\begin{aligned}
0 \leq & \sum_{j \in \mathcal{J}_{ex}} \left( \begin{aligned} & \frac{1}{n} \sum_{i=1}^n \frac{d\rho_j(z_i, \widehat{\alpha})}{d\alpha}[\varepsilon_n u_n^*] \times \rho_j(z_i, \widehat{\alpha}) \\ & + \frac{1}{2n} \sum_{i=1}^n \frac{d^2\rho_j(z_i, \alpha(s))}{d\alpha d\alpha}[\varepsilon_n u_n^*, \varepsilon_n u_n^*] \times \rho_j(z_i, \alpha(s)) \\ & + \frac{1}{2n} \sum_{i=1}^n \frac{d\rho_j(z_i, \alpha(s))}{d\alpha}[\varepsilon_n u_n^*] \times \frac{d\rho_j(z_i, \alpha(s))}{d\alpha}[\varepsilon_n u_n^*] \end{aligned} \right) + \\
& \sum_{j \in \mathcal{J}_{1en}} \left( \begin{aligned} & \frac{1}{n} \sum_{i=1}^n \frac{d\widehat{m}_j(x_{ji}, \widehat{\alpha})}{d\alpha}[\varepsilon_n u_n^*] \times \widehat{m}_j(x_{ji}, \widehat{\alpha}) \\ & + \frac{1}{2n} \sum_{i=1}^n \frac{d^2\widehat{m}_j(x_{ji}, \alpha(s))}{d\alpha d\alpha}[\varepsilon_n u_n^*, \varepsilon_n u_n^*] \times \widehat{m}_j(x_{ji}, \alpha(s)) \\ & + \frac{1}{2n} \sum_{i=1}^n \frac{d\widehat{m}_j(x_{ji}, \alpha(s))}{d\alpha}[\varepsilon_n u_n^*] \times \frac{d\widehat{m}_j(x_{ji}, \alpha(s))}{d\alpha}[\varepsilon_n u_n^*] \end{aligned} \right) + \\
& \sum_{j \in \mathcal{J}_{2en}} \left( \begin{aligned} & \frac{d\widehat{m}_j(\widehat{\alpha})}{d\alpha}[\varepsilon_n u_n^*] \times \widehat{m}_j(\widehat{\alpha}) \\ & + \frac{1}{2} \frac{d^2\widehat{m}_j(\alpha(s))}{d\alpha d\alpha}[\varepsilon_n u_n^*, \varepsilon_n u_n^*] \times \widehat{m}_j(\alpha(s)) \\ & + \frac{1}{2} \frac{d\widehat{m}_j(\alpha(s))}{d\alpha}[\varepsilon_n u_n^*] \times \frac{d\widehat{m}_j(\alpha(s))}{d\alpha}[\varepsilon_n u_n^*] \end{aligned} \right)
\end{aligned}$$

where  $\alpha(s) = \widehat{\alpha} + s\varepsilon_n u_n^* \equiv \widetilde{\alpha} \in \mathcal{N}_{on}$ . Applying Lemma A.1 of Ai an Chen (2003) (also see the proof of their Corollary C.2), under Assumption 3.1 - 3.3, 3.5 - 3.8 and 4.3, we have uniformly over  $\alpha(s) \in \mathcal{N}_{on}$ , for  $j \in \mathcal{J}_{1en}$ ,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \left\{ \frac{d^2\widehat{m}_j(x_{ji}, \alpha(s))}{d\alpha d\alpha} [u_n^*, u_n^*] \right\} \widehat{m}_j(x_{ji}, \alpha(s)) &= O_p(1); \\
\frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\widehat{m}_j(x_{ji}, \alpha(s))}{d\alpha} [u_n^*] \right\} \frac{d\widehat{m}_j(x_{ji}, \alpha(s))}{d\alpha} [u_n^*] &= O_p(1).
\end{aligned}$$

Since for all  $j$  and for any  $\alpha \in \mathcal{N}_{on}$ ,

$$\left| \frac{d\rho_j(z_i, \alpha)}{d\alpha} [v_n^*] \right| \leq \left| \frac{d\rho_j(z_i, \alpha)}{d\alpha} [v_n^*] - \frac{d\rho_j(z_i, \alpha_*)}{d\alpha} [v_n^*] \right| + \left| \frac{d\rho_j(z_i, \alpha_*)}{d\alpha} [v_n^*] \right|,$$

under Assumption 4.3(i), we have  $E \left( \left\{ \sup_{\alpha \in \mathcal{N}_{on}} \left| \frac{d\rho_j(z_i, \alpha)}{d\alpha} [v_n^*] \right| \right\}^2 \right) < \infty$  for all  $j$ . Thus Assumptions 3.5(i)(ii), 3.8(ii) and 4.3(i)(ii) imply for  $j \in \mathcal{J}_{2en}$ ,

$$\left\{ \frac{d^2\widehat{m}_j(\alpha(s))}{d\alpha d\alpha} [u_n^*, u_n^*] \right\} \widehat{m}_j(\alpha(s)) = O_p(1), \quad \left( \frac{d\widehat{m}_j(\alpha(s))}{d\alpha} [u_n^*] \right)^2 = O_p(1),$$

and for  $j \in \mathcal{J}_{ex}$ :

$$\frac{1}{n} \sum_{i=1}^n \left\{ \frac{d^2\rho_j(z_i, \alpha(s))}{d\alpha d\alpha} [u_n^*, u_n^*] \right\} \rho_j(z_i, \alpha(s)) = O_p(1), \quad \frac{1}{n} \sum_{i=1}^n \left( \frac{d\rho_j(z_i, \alpha(s))}{d\alpha} [u_n^*] \right)^2 = O_p(1).$$

Hence, uniformly over  $\alpha(s) \in \mathcal{N}_{on}$ ,

$$\begin{aligned}
0 \leq & \frac{\varepsilon_n}{n} \sum_{j \in \mathcal{J}_{ex}} \sum_{i=1}^n \left\{ \frac{d\rho_j(z_i, \widehat{\alpha})}{d\alpha} [u_n^*] \right\} \rho_j(z_i, \widehat{\alpha}) + \\
& \frac{\varepsilon_n}{n} \sum_{j \in \mathcal{J}_{1en}} \sum_{i=1}^n \left\{ \frac{d\widehat{m}_j(x_{ji}, \widehat{\alpha})}{d\alpha} [u_n^*] \right\} \widehat{m}_j(x_{ji}, \widehat{\alpha}) + \varepsilon_n \sum_{j \in \mathcal{J}_{2en}} \left\{ \frac{d\widehat{m}_j(\widehat{\alpha})}{d\alpha} [u_n^*] \right\} \widehat{m}_j(\widehat{\alpha}) + O_p(\varepsilon_n^2).
\end{aligned}$$

Repeating the above reasoning with  $u^* = -v^*$  and noting that  $\varepsilon_n = o(n^{-1/2}) > 0$ , we obtain

$$\begin{aligned} o_p(n^{-1/2}) &= \sum_{j \in \mathcal{J}_{ex}} \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\rho_j(z_i, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \hat{\alpha}) + \\ &\quad \sum_{j \in \mathcal{J}_{1en}} \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\hat{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \hat{m}_j(x_{ji}, \hat{\alpha}) + \sum_{j \in \mathcal{J}_{2en}} \left\{ \frac{d\hat{m}_j(\hat{\alpha})}{d\alpha} [v_n^*] \right\} \hat{m}_j(\hat{\alpha}). \end{aligned}$$

Consider the second term on the right hand side. Applying Corollary A1(i) and C1(i) of Ai and Chen (2003), we obtain for  $j \in \mathcal{J}_{1en}$

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\hat{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \hat{m}_j(x_{ji}, \hat{\alpha}) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \hat{m}_j(x_{ji}, \hat{\alpha}) + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\hat{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] - \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) \\ &\quad + o_p(n^{-1/2}). \end{aligned}$$

Since the approximation error of  $m_j(X_j, \alpha)$  and  $\frac{dm_j(X_j, \alpha)}{d\alpha} [v_n^*]$  are  $o(n^{-1/4})$  by Assumptions 3.2(iii), 3.5(iv) and 4.3(iii), we have uniformly over  $\alpha \in \mathcal{N}_{on}$ ,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \alpha)}{d\alpha} [v_n^*] \right\} (\tilde{m}_j(x_{ji}, \alpha) - m_j(x_{ji}, \alpha)) \\ &= \frac{1}{n} \sum_{i=1}^n \left( \frac{dm_j(x_{ji}, \alpha)}{d\alpha} [v_n^*] - \frac{d\tilde{m}_j(x_{ji}, \alpha)}{d\alpha} [v_n^*] \right)' (\tilde{m}_j(x_{ji}, \alpha) - m_j(x_{ji}, \alpha)) = o_p(n^{-1/2}), \end{aligned}$$

where the first equality follows from the fact that  $m_j(x_{ji}, \alpha) - \tilde{m}_j(x_{ji}, \alpha)$  is the LS regression residual, and the second equality follows from applying the approximation error. Hence

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \hat{m}_j(x_{ji}, \hat{\alpha}) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \{ \hat{m}_j(x_{ji}, \hat{\alpha}) - \tilde{m}_j(x_{ji}, \hat{\alpha}) \} + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \tilde{m}_j(x_{ji}, \hat{\alpha}) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\tilde{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \{ \rho_j(z_i, \hat{\alpha}) - m_j(x_{ji}, \hat{\alpha}) \} + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \{ \rho_j(z_i, \hat{\alpha}) - m_j(x_{ji}, \hat{\alpha}) \} + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) + o_p(n^{-1/2}), \end{aligned}$$

where the last equality follows from applying Lemma A.1(c) ( $\delta_n = o(n^{-1/4})$ ,  $\delta_{1n} = n^{-1/2}$ ). Similarly,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\hat{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] - \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\hat{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] - \frac{d\tilde{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\tilde{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] - \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \{ m_j(x_{ji}, \hat{\alpha}) - \tilde{m}_j(x_{ji}, \hat{\alpha}) \} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\rho_j(z_i, \hat{\alpha})}{d\alpha} [v_n^*] - \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \tilde{m}_j(x_{ji}, \hat{\alpha}) + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\rho_j(z_i, \hat{\alpha})}{d\alpha} [v_n^*] - \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) + o_p(n^{-1/2}),
\end{aligned}$$

where the first equation is due to  $\frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\tilde{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] - \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \tilde{m}_j(x_{ji}, \hat{\alpha}) = 0$  and the last equation is due to Lemma A.1(c). Combing both parts of the results, we have for all  $j \in \mathcal{J}_{1en}$  :

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \left\{ \frac{d\hat{m}_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \hat{m}_j(x_{ji}, \hat{\alpha}) \\
&= \frac{1}{n} \sum_{i=1}^n \left( \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \hat{\alpha}) + \left\{ \frac{d(\rho_j(z_i, \hat{\alpha}) - m_j(x_{ji}, \hat{\alpha}))}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) \right) + o_p(n^{-1/2}).
\end{aligned}$$

Using the same reasoning, we obtain for all  $j \in \mathcal{J}_{2en}$  :

$$\begin{aligned}
&\left\{ \frac{d\hat{m}_j(\hat{\alpha})}{d\alpha} [v_n^*] \right\} \hat{m}_j(\hat{\alpha}) \\
&= \frac{1}{n} \sum_{i=1}^n \left( \left\{ \frac{dm_j(\hat{\alpha})}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \hat{\alpha}) + \left\{ \frac{d(\rho_j(z_i, \hat{\alpha}) - m_j(\hat{\alpha}))}{d\alpha} [v_n^*] \right\} m_j(\hat{\alpha}) \right) + o_p(n^{-1/2}).
\end{aligned}$$

Therefore we have

$$\frac{1}{n} \sum_{i=1}^n \left( \begin{aligned} &\sum_{j \in \mathcal{J}_{ex}} \left\{ \frac{d\rho_j(z_i, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \hat{\alpha}) + \\ &\sum_{j \in \mathcal{J}_{1en}} \left( \left\{ \frac{dm_j(x_{ji}, \hat{\alpha})}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \hat{\alpha}) + \left\{ \frac{d(\rho_j(z_i, \hat{\alpha}) - m_j(x_{ji}, \hat{\alpha}))}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \hat{\alpha}) \right) + \\ &\sum_{j \in \mathcal{J}_{2en}} \left( \left\{ \frac{dm_j(\hat{\alpha})}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \hat{\alpha}) + \left\{ \frac{d(\rho_j(z_i, \hat{\alpha}) - m_j(\hat{\alpha}))}{d\alpha} [v_n^*] \right\} m_j(\hat{\alpha}) \right) \end{aligned} \right) = o_p(n^{-1/2}),$$

which can be rewritten in a compact form:

$$\begin{aligned}
o_p(n^{-1/2}) &= \frac{1}{n} \sum_{i=1}^n \left( \left\{ \frac{dm(x_i, \hat{\alpha})}{d\alpha} [v_n^*] \right\}' \rho(z_i, \hat{\alpha}) + \left\{ \frac{d(\rho(z_i, \hat{\alpha}) - m(x_i, \hat{\alpha}))}{d\alpha} [v_n^*] \right\}' m(x_i, \hat{\alpha}) \right) \\
&= \frac{-1}{2n} \sum_{i=1}^n \frac{d\ell(z_i, \hat{\alpha})}{d\alpha} [v_n^*].
\end{aligned}$$

Notice that under Assumptions 3.5 and 4.3, for all  $j = 1, \dots, J$ ,

$$\begin{aligned}
&\left| \left\{ \frac{dm_j(x_{ji}, \alpha)}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \alpha) - \left\{ \frac{dm_j(x_{ji}, \alpha_*)}{d\alpha} [v_n^*] \right\} \rho_j(z_i, \alpha_*) \right| \\
&\leq \left| \frac{dm_j(x_{ji}, \alpha)}{d\alpha} [v_n^*] - \frac{dm_j(x_{ji}, \alpha_*)}{d\alpha} [v_n^*] \right| |\rho_j(z_i, \alpha)| + \left| \frac{dm_j(x_{ji}, \alpha_*)}{d\alpha} [v_n^*] \right| |\rho_j(z_i, \alpha) - \rho_j(z_i, \alpha_*)| \\
&\leq b(z_i) \|\alpha - \alpha_*\|_s^\kappa \quad \text{for some } E\{[b(z_i)]^2\} < \infty,
\end{aligned}$$

and for all  $j \in \mathcal{J}_{en}$ ,

$$\begin{aligned}
&\left| \left\{ \frac{d(\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha))}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \alpha) - \left\{ \frac{d(\rho_j(z_i, \alpha_*) - m_j(x_{ji}, \alpha_*))}{d\alpha} [v_n^*] \right\} m_j(x_{ji}, \alpha_*) \right| \\
&\leq \left| \frac{d(\rho_j(z_i, \alpha) - m_j(x_{ji}, \alpha))}{d\alpha} [v_n^*] - \frac{d(\rho_j(z_i, \alpha_*) - m_j(x_{ji}, \alpha_*))}{d\alpha} [v_n^*] \right| |m_j(x_{ji}, \alpha)| \\
&\quad + \left| \frac{d(\rho_j(z_i, \alpha_*) - m_j(x_{ji}, \alpha_*))}{d\alpha} [v_n^*] \right| |m_j(x_{ji}, \alpha) - m_j(x_{ji}, \alpha_*)| \\
&\leq b'(z_i) \|\alpha - \alpha_*\|_s^\kappa \quad \text{for some } E\{[b'(z_i)]^2\} < \infty.
\end{aligned}$$

Let  $\mathcal{F}_n = \left\{ \frac{d\ell(z_i, \alpha)}{d\alpha} [v_n^*] - \frac{d\ell(z_i, \alpha_*)}{d\alpha} [v_n^*] : \alpha \in \mathcal{N}_{on} \right\}$ . Then  $N_{\square}(\varepsilon, \mathcal{F}_n, \|\cdot\|_{L_2(P)}) \leq N(\{c\varepsilon\}^{1/\kappa}, \mathcal{N}_{on}, \|\cdot\|_s)$ . Under Assumption 4.5, we can apply Lemma 1 of Chen, Linton and van Keilegom (2003), and obtain:

$$\begin{aligned} o_p(n^{-1/2}) &= \frac{1}{n} \sum_{i=1}^n \left( \left\{ \frac{dm(x_i, \hat{\alpha})}{d\alpha} [v_n^*] \right\}' \rho(z_i, \hat{\alpha}) + \left\{ \frac{d(\rho(z_i, \hat{\alpha}) - m(x_i, \hat{\alpha}))}{d\alpha} [v_n^*] \right\}' m(x_i, \hat{\alpha}) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left( \left\{ \frac{dm(x_i, \alpha_*)}{d\alpha} [v_n^*] \right\}' \rho(z_i, \alpha_*) + \left\{ \frac{d(\rho(z_i, \alpha_*) - m(x_i, \alpha_*))}{d\alpha} [v_n^*] \right\}' m(x_i, \alpha_*) \right) \\ &\quad + E \left( \left\{ \frac{dm(x_i, \hat{\alpha})}{d\alpha} [v_n^*] \right\}' \rho(z_i, \hat{\alpha}) - \left\{ \frac{dm(x_i, \alpha_*)}{d\alpha} [v_n^*] \right\}' \rho(z_i, \alpha_*) \right) + o_p(n^{-1/2}). \end{aligned}$$

By the definition of the norm, we have

$$\langle v_n^*, \hat{\alpha} - \alpha_* \rangle \equiv E \left\{ \left\{ \frac{dm(X, \alpha_*)}{d\alpha} [v_n^*] \right\}' \frac{dm(X, \alpha_*)}{d\alpha} [\hat{\alpha} - \alpha_*] + \left( \frac{d^2 m(X, \alpha_*)}{d\alpha^2} [v_n^*, \hat{\alpha} - \alpha_*] \right)' m(X, \alpha_*) \right\}.$$

With  $\alpha(t) = \alpha_* + t(\hat{\alpha} - \alpha_*)$ , a Taylor expansion around  $t = 0$  gives

$$\begin{aligned} &E \left( \left\{ \frac{dm(x_i, \hat{\alpha})}{d\alpha} [v_n^*] \right\}' \rho(z_i, \hat{\alpha}) - \left\{ \frac{dm(x_i, \alpha_*)}{d\alpha} [v_n^*] \right\}' \rho(z_i, \alpha_*) \right) \\ &= E \left( \left\{ \frac{dm(x_i, \hat{\alpha})}{d\alpha} [v_n^*] \right\}' m(x_i, \hat{\alpha}) - \left\{ \frac{dm(x_i, \alpha_*)}{d\alpha} [v_n^*] \right\}' m(x_i, \alpha_*) \right) \\ &= \langle v_n^*, \hat{\alpha} - \alpha_* \rangle + \frac{d^2 E \left[ \left\{ \frac{dm(X, \alpha(\tilde{t}))}{d\alpha} [v_n^*] \right\}' m(X, \alpha(\tilde{t})) \right]}{dt^2} \\ &= \langle v_n^*, \hat{\alpha} - \alpha_* \rangle + o(n^{-1/2}) = \langle v^*, \hat{\alpha} - \alpha_* \rangle + o_p(n^{-1/2}), \end{aligned}$$

where  $\tilde{t}$  is between zero and one, and the third equality is due to Assumption 4.4 and the last equality is due to Assumption 4.2. Hence, we obtain

$$\begin{aligned} &\sqrt{n} \langle v^*, \hat{\alpha} - \alpha_* \rangle \\ &= \frac{-1}{\sqrt{n}} \sum_{i=1}^n \left( \left\{ \frac{dm(x_i, \alpha_*)}{d\alpha} [v_n^*] \right\}' \rho(z_i, \alpha_*) + \left\{ \frac{d\{\rho(z_i, \alpha_*) - m(x_i, \alpha_*)\}}{d\alpha} [v_n^*] \right\}' m(x_i, \alpha_*) \right) + o_p(1) \\ &= \frac{-1}{\sqrt{n}} \sum_{i=1}^n \left( \left\{ \frac{dm(x_i, \alpha_*)}{d\alpha} [v_n^*] \right\}' \rho(z_i, \alpha_*) + \left\{ \frac{d\{\rho(z_i, \alpha_*) - m(x_i, \alpha_*)\}}{d\alpha} [v_n^*] \right\}' m(x_i, \alpha_*) \right) + o_p(1), \end{aligned}$$

where the last equality is due to Assumptions 4.6 and 4.2. Since  $\langle v^*, \hat{\alpha} - \alpha_* \rangle = \lambda'(\hat{\theta} - \theta_*)$  for any fixed  $\lambda \in \mathcal{R}^{d_\theta}$  with  $|\lambda| \neq 0$ , we obtain Theorem 4.1 by applying a standard CLT for i.i.d. data. ■

**Proof. (Theorem 5.1):** First we show that uniformly over  $w_l \in \mathcal{H}_n$ , the following (24) - (29) hold:

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \left( \frac{\partial \rho_j(z_i, \hat{\alpha}_n)}{\partial \theta_l} - \frac{d \rho_j(z_i, \hat{\alpha}_n)}{dh} [w_l] \right)^2 \tag{24} \\ &= E \left\{ \left( \frac{\partial \rho_j(Z, \alpha_*)}{\partial \theta_l} - \frac{d \rho_j(Z, \alpha_*)}{dh} [w_l] \right)^2 \right\} + o_p(1) \quad \text{for } j \in \mathcal{J}_{ex}; \end{aligned}$$

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial \widehat{m}_j(x_{ji}, \widehat{\alpha}_n)}{\partial \theta_l} - \frac{d \widehat{m}_j(x_{ji}, \widehat{\alpha}_n)}{dh} [w_l] \right)^2 \\ &= E \left\{ \left( \frac{\partial m_j(X_j, \alpha_*)}{\partial \theta_l} - \frac{d m_j(X_j, \alpha_*)}{dh} [w_l] \right)^2 \right\} + o_p(1) \quad \text{for } j \in \mathcal{J}_{1en}; \end{aligned} \quad (25)$$

$$\left( \frac{\partial \widehat{m}_j(\widehat{\alpha}_n)}{\partial \theta_l} - \frac{d \widehat{m}_j(\widehat{\alpha}_n)}{dh} [w_l] \right)^2 = \left( \frac{\partial m_j(\alpha_*)}{\partial \theta_l} - \frac{d m_j(\alpha_*)}{dh} [w_l] \right)^2 + o_p(1) \quad \text{for } j \in \mathcal{J}_{2en}. \quad (26)$$

Similarly, for all  $j \in \mathcal{J}_{ex}$ ,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial^2 \rho_j(z_i, \widehat{\alpha}_n)}{\partial \theta_l^2} - 2 \frac{d^2 \rho_j(z_i, \widehat{\alpha}_n)}{\partial \theta_l dh} [w_l] + \frac{d^2 \rho_j(z_i, \widehat{\alpha}_n)}{dh^2} [w_l, w_l] \right) \rho_j(z_i, \widehat{\alpha}_n) \\ &= E \left\{ \left( \frac{\partial^2 \rho_j(Z, \alpha_*)}{\partial \theta_l^2} - 2 \frac{d^2 \rho_j(Z, \alpha_*)}{\partial \theta_l dh} [w_l] + \frac{d^2 \rho_j(Z, \alpha_*)}{dh^2} [w_l, w_l] \right) \rho_j(Z, \alpha_*) \right\} + o_p(1); \end{aligned} \quad (27)$$

for all  $j \in \mathcal{J}_{1en}$ ,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial^2 \widehat{m}_j(x_{ji}, \widehat{\alpha}_n)}{\partial \theta_l^2} - 2 \frac{d^2 \widehat{m}_j(x_{ji}, \widehat{\alpha}_n)}{\partial \theta_l dh} [w_l] + \frac{d^2 \widehat{m}_j(x_{ji}, \widehat{\alpha}_n)}{dh^2} [w_l, w_l] \right) \widehat{m}_j(x_{ji}, \widehat{\alpha}_n) \\ &= E \left\{ \left( \frac{\partial^2 m_j(X_j, \alpha_*)}{\partial \theta_l^2} - 2 \frac{d^2 m_j(X_j, \alpha_*)}{\partial \theta_l dh} [w_l] + \frac{d^2 m_j(X_j, \alpha_*)}{dh^2} [w_l, w_l] \right) m_j(X_j, \alpha_*) \right\} + o_p(1); \end{aligned} \quad (28)$$

and for all  $j \in \mathcal{J}_{2en}$ ,

$$\begin{aligned} & \left( \frac{\partial^2 \widehat{m}_j(\widehat{\alpha}_n)}{\partial \theta_l^2} - 2 \frac{d^2 \widehat{m}_j(\widehat{\alpha}_n)}{\partial \theta_l dh} [w_l] + \frac{d^2 \widehat{m}_j(\widehat{\alpha}_n)}{dh^2} [w_l, w_l] \right) \widehat{m}_j(\widehat{\alpha}_n) \\ &= \left( \frac{\partial^2 m_j(\alpha_*)}{\partial \theta_l^2} - 2 \frac{d^2 m_j(\alpha_*)}{\partial \theta_l dh} [w_l] + \frac{d^2 m_j(\alpha_*)}{dh^2} [w_l, w_l] \right) m_j(\alpha_*) + o_p(1). \end{aligned} \quad (29)$$

Recall that for all  $j \in \mathcal{J}_{1en}$ ,

$$\begin{aligned} & \frac{\partial \widehat{m}_j(X_j, \alpha)}{\partial \theta_l} - \frac{d \widehat{m}_j(X_j, \alpha)}{dh} [w_l] \\ &= p_j^{k_{jn}}(X_j)' (P_j' P_j)^{-1} \sum_{i=1}^n p_j^{k_{jn}}(x_{ji}) \left\{ \frac{\partial \rho_j(z_i, \alpha)}{\partial \theta_l} - \frac{d \rho_j(z_i, \alpha)}{dh} [w_l] \right\}, \\ & \frac{\partial^2 \widehat{m}_j(X_j, \alpha)}{\partial \theta_l^2} - 2 \frac{d^2 \widehat{m}_j(X_j, \alpha)}{\partial \theta_l dh} [w_l] + \frac{d^2 \widehat{m}_j(X_j, \alpha)}{dh^2} [w_l, w_l] \\ &= p_j^{k_{jn}}(X_j)' (P_j' P_j)^{-1} \sum_{i=1}^n p_j^{k_{jn}}(x_{ji}) \left\{ \frac{\partial^2 \rho_j(z_i, \alpha)}{\partial \theta_l^2} - 2 \frac{d^2 \rho_j(z_i, \alpha)}{\partial \theta_l dh} [w_l] + \frac{d^2 \rho_j(z_i, \alpha)}{dh^2} [w_l, w_l] \right\}. \end{aligned}$$

By Assumption 5.1 and applying Lemma A.1 of Ai and Chen (2003), we obtain that for all  $j \in \mathcal{J}_{1en}$ , uniformly over  $w_l \in \mathcal{H}_n$ ,

$$\frac{\partial \widehat{m}_j(X_j, \widehat{\alpha}_n)}{\partial \theta_l} - \frac{d \widehat{m}_j(X_j, \widehat{\alpha}_n)}{dh} [w_l] = \frac{\partial m_j(X_j, \widehat{\alpha}_n)}{\partial \theta_l} - \frac{d m_j(X_j, \widehat{\alpha}_n)}{dh} [w_l] + o_p(1).$$

Assumption 5.1 also implies that for all  $j \in \mathcal{J}_{1en}$ , uniformly over  $w_l \in \mathcal{H}_n$ ,

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{\partial m_j(X_j, \widehat{\alpha}_n)}{\partial \theta_l} - \frac{d m_j(X_j, \widehat{\alpha}_n)}{dh} [w_l] \right)^2 = \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial m_j(X_j, \alpha_*)}{\partial \theta_l} - \frac{d m_j(X_j, \alpha_*)}{dh} [w_l] \right)^2 + o_p(1).$$

Applying Lemma A.1 of Ai and Chen (2003), we obtain for all  $j \in \mathcal{J}_{1en}$  and uniformly over  $w_l \in \mathcal{H}_n$ ,

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{\partial m_j(X_j, \alpha_*)}{\partial \theta_l} - \frac{dm_j(X_j, \alpha_*)}{dh} [w_l] \right)^2 = E \left( \frac{\partial m_j(X_j, \alpha_*)}{\partial \theta_l} - \frac{dm_j(X_j, \alpha_*)}{dh} [w_l] \right)^2 + o_p(1).$$

This proves (25). Using the same reasonings, we can show (24) and (26)-(29), where we use Assumption 5.2 for (27)-(29).

Since  $\mathcal{H}_n$  is dense in  $\overline{\mathcal{W}}$ , and is compact under  $\|\cdot\|_s$ . It follows that  $\|\widehat{w}_l^* - w_l^*\|_s = o_p(1)$ .

The consistency of  $\widehat{w}_l^*$  and results (24)-(29) yields

$$\frac{1}{n} \sum_{i=1}^n \{ \widehat{D}_{\widehat{w}^*}(x_i)' \widehat{D}_{\widehat{w}^*}(x_i) + \widehat{V}_{\widehat{w}^*}(x_i) \} = E \{ D_{w^*}(X)' D_{w^*}(X) + V_{w^*}(X) \} + o_p(1).$$

Next, we show  $\widehat{\Omega} = \Omega_* + o_p(1)$ . The results above and the Hölder continuity of  $\rho_j(Z, \alpha)$  yields

$$\begin{aligned} \widehat{\Omega} &= \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varepsilon_i' + o_p(1), \quad \text{with } \varepsilon_i = \varepsilon_i^{ex} + \varepsilon_i^{1en} + \varepsilon_i^{2en}, \\ \varepsilon_i^{ex} &= \sum_{j \in \mathcal{J}_{ex}} \{ D_{jw^*}^{ex}(z_i) \}' \rho_j(z_i, \alpha_*), \\ \varepsilon_i^{1en} &\equiv \sum_{j \in \mathcal{J}_{1en}} \left\{ \left\{ \frac{\partial \rho_j(z_i, \alpha_*)}{\partial \theta'} - \frac{d\rho_j(z_i, \alpha_*)}{dh} [w^*] - \widehat{D}_{jw^*}^{1en}(x_{ji}) \right\}' m_j(x_{ji}, \alpha_*) + \{ D_{jw^*}^{1en}(x_{ji}) \}' \rho_j(z_i, \alpha_*) \right\} \\ \varepsilon_i^{2en} &\equiv \sum_{j \in \mathcal{J}_{2en}} \left\{ \left\{ \frac{\partial \rho_j(z_i, \alpha_*)}{\partial \theta'} - \frac{d\rho_j(z_i, \alpha_*)}{dh} [w^*] - D_{jw^*}^{2en} \right\}' m_j(\alpha_*) + \{ D_{jw^*}^{2en} \}' \rho_j(z_i, \alpha_*) \right\}. \end{aligned}$$

By a standard weak law of large numbers for i.i.d. data we have  $\frac{1}{n} \sum_{i=1}^n \varepsilon_i \varepsilon_i' = \Omega_* + o_p(1)$ , hence  $\widehat{\Omega} = \Omega_* + o_p(1)$ . The theorem now follows immediately. ■

## Acknowledgments

We thank the guest co-editor and three anonymous referees whose comments have greatly improved the presentation of the paper. We thank Whitney Newey for suggesting this topic and for insightful discussions on the average derivative IV estimator. We also thank Oliver Linton, Arthur Lewbel and Jim Powell for useful discussions and seminar participants at Yale, CEMMAP/London and University of Mannheim for comments. Ai's research is supported by the 2003 summer grant from the Warrington Business School at University of Florida. Chen's research is supported by the National Science Foundation and the C.V. Starr Center at New York University. All remaining errors are the responsibility of the authors.

## References

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

- Ai, C. and X. Chen, 2005, Efficient Estimation of Sequential Moment Restrictions Containing Unknown Functions, mimeo, University of Florida and New York University.
- Andrews, D., 1994, Asymptotics for Semi-parametric Econometric Models via Stochastic Equicontinuity, *Econometrica*, 62, 43-72.
- Baltagi, B. and Q. Li, 2003, On instrumental Variable Estimation of Semiparametric Dynamic Panel Data Models, *Economics Letters*, 76, 1-9.
- Bhattacharya, D., 2005, Inference in Panel Data Models Under Attrition on Unobservables, Using Auxiliary Information, mimeo.
- Chen, X. and X. Shen, 1998, Sieve Extremum Estimates for Weakly Dependent Data, *Econometrica*, 66, 289-314.
- Chen, X., 2005, Large Sample Sieve Estimation of Semi-Nonparametric Models, in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6. North-Holland, Amsterdam.
- Chen, X. and H. White, 1999, Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators, *IEEE Tran. Information Theory*, 45, 682-691.
- Chen, X., O. Linton and I. van Keilegom, 2003, Estimation of Semiparametric Models when the Criterion Function is not Smooth, *Econometrica*, 71, 1591-1608.
- Ekeland, I., J. Heckman and L. Nesheim, 2004, Identification and Estimation of Hedonic Models, *The Journal of Political Economy*, 112(S1): S60-S109.
- Gayle, G., and C. Viauroux, 2005, Root-N Consistent Semiparametric Estimators of a Dynamic Panel Sample Selection Model, mimeo, Carnegie Mellon University and University of Cincinnati.
- Hall, A. and A. Inoue, 2003, The Large Sample Behavior of the Generalized Method of Moments Estimator in Misspecified Models, *Journal of Econometrics*, 114, 361-394.
- Hansen, L.P. and R. Jagannathan, 1997, Assessing Specification Errors in Stochastic Discount Factor Models, *Journal of Finance*, 52, 557-590.
- Hausman, J., 1977, Errors-in-variables in Simultaneous Equations Models, *Journal of Econometrics*, 5, 389-401.
- Heckman, J., R. Matzkin and L. Nesheim, 2004, Simulation and Estimation of Hedonic Models, forthcoming in *Frontiers of Applied General Equilibrium Modeling: Essays in Honour of Herbert Scarf*.
- Horowitz, J., 1998, *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.
- Ichimura, H., 1993, Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models, *Journal of Econometrics*, 58, 71-120.
- Klein, R. and R. Spady, 1993, An Efficient Semiparametric Estimator for Binary Response Models, *Econometrica*, 61, 387-421.
- Lewbel, A., 2005, Identification of Endogenous Heteroskedastic Models, mimeo, Boston College.
- Linton, O. and E. Mammen, 2005, Estimating Semiparametric ARCH( $\infty$ ) Models by Kernel

- Smoothing Methods, *Econometrica*, 73, 771-836.
- Newey, W.K., 1984, A Method of Moments Interpretation of Sequential Estimators, *Economics Letters*, 14, 201-206.
- Newey, W.K., 1994, The Asymptotic Variance of Semiparametric Estimators, *Econometrica*, 62, 1349-1382.
- Newey, W.K., 1997, Convergence Rates and Asymptotic Normality for Series Estimators, *Journal of Econometrics*, 79, 147-168.
- Newey, W.K. and D. McFadden, 1994, Large sample estimation and hypothesis testing, in R. Engle and D. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.
- Newey, W.K. and J. Powell, 2003, Instrumental Variable Estimation of Nonparametric Models, *Econometrica*, 71, 1565-1578.
- Newey, W.K., J.L. Powell and F. Vella, 1999, Nonparametric Estimation of Triangular Simultaneous Equations Models, *Econometrica*, 67, 565-603.
- Newey, W.K. and T. Stoker, 1993, Efficiency of Weighted Average Derivative Estimators and Index Models, *Econometrica*, 61, 1199-1223.
- Nishiyama, Y. and P. Robinson, 2005, The Bootstrap and the Edgeworth Correction for Semiparametric Averaged Derivatives, *Econometrica*, 73, 903-948.
- Pakes, A. and S. Olley, 1995, A Limit Theorem for A Smooth Class of Semiparametric Estimators, *Journal of Econometrics*, 65, 295-332.
- Powell, J., 1994, Estimation of Semiparametric Models, in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.
- Powell, J., J. Stock, and T. Stoker, 1989, Semiparametric Estimation of Index Coefficients, *Econometrica* 57, 1403-1430.
- Robinson, P., 1988, Root-N-Consistent Semiparametric Regression, *Econometrica* 56, 931-954.
- Shen, X., 1997, On Methods of Sieves and Penalization, *The Annals of Statistics* 25, 2555-2591.
- Stone, C.J., 1985, Additive regression and other nonparametric models, *The Annals of Statistics*, 13, 689-705.
- Van der Vaart, A. and J. Wellner, 1996, *Weak Convergence and Empirical Processes: with Applications to Statistics*. New York: Springer-Verlag.
- White, H., 1982, Maximum likelihood estimation of misspecified models, *Econometrica*, 50, 143-161.
- White, H., 1994. *Estimation, Inference and Specification Analysis*. Cambridge University Press, Cambridge, UK.
- Wooldridge, J., 1996, Estimating Systems of Equations with Different Instruments for Different Equations, *Journal of Econometrics*, 74, 387 - 405.