E¢cient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions¤

Chunrong Ai
Department of Economics
University of Florida

Xiaohong Chen
Department of Economics
London School of Economics

August 1999

Abstract

We propose an estimation method for models of conditional moment restrictions that contain …nite dimensional unknown parameters (μ) as well as unknown functions (h), and that the unknowns can interact with endogenous variables. Our proposal is to approximate h with sieves and then to estimate μ and the sieve parameters jointly by applying the method of generalized minimum distance. We show that: (1) the sieve estimator of h is consistent under certain metric, and may attain the usual nonparametric convergence rates; (2) the estimator of μ is $\sqrt{n}$ consistent and asymptotically normally distributed; (3) the estimator for the asymptotic variance of the μ estimator is consistent and easy to compute; and (4) the optimally weighted minimum distance estimator of μ attains the semiparametric e¢ciency bound. These results extend the work of Hansen (1982) and Newey (1990, 1993) to a semiparametric conditional moment restrictions setting.

¤We acknowledge helpful comments from Richard Blundell, Lars Hansen, Han Hong, Bo Honore, Roger Klein, Oliver Linton, Whitney Newey, Jim Powell, Frank Vella and Seminar participants at Texas Austin, Rochester, Chicago, Duke, Rutgers, Princeton, UCL, Carol III at Madrid, and the 1999 North American Winter Meetings of the Econometric Society. Ai's research is supported by the Warrington College of Business Administration Summer Grants for 1998 and 1999 at the University of Florida.

1. Introduction

A general framework for analyzing data $(Y; X)$ is to presume that the data satisfy some conditional moment restrictions:

$$E[\psi(Z; \theta_o)|X] = 0, \tag{1.1}$$

where $Z \equiv (Y', X_z')'$, $X_z \subset X$, θ_o is a vector of unknown parameters, $\psi(\cdot;\cdot)$ is a vector of known functions, and $E[\psi(Z; \theta_o)|X]$ is the conditional expectation of $\psi(Z; \theta_o)$ given X, where the conditional distribution of Y given X is unspecified. A common method for estimating the parameters of interest, θ_o, is the generalized method of moments (GMM) estimation proposed by Hansen (1982). The GMM converts the conditional moment restrictions (1.1) into the unconditional moment restrictions:

$$E[\psi(Z; \theta_o)W] = 0; \tag{1.2}$$

where W are instruments (e.g. X and nonlinear measurable transformations of X), and estimates θ_o by solving the sample analog of (1.2). Under some sufficient conditions, Hansen (1982) showed that the GMM estimator of θ_o is $\sqrt{n}$ consistent and asymptotically normally distributed, where n is the sample size. Moreover, he established that the GMM estimator can attain the efficiency bound of the unconditional moment restrictions model (1.2). Chamberlain (1987) derived the efficiency bound of the conditional moment restrictions model (1.1). He showed that the GMM estimator attains his bound if W are the optimal instruments[1]. Robinson (1987, 1991) and Newey (1990, 1993) proposed to estimate the optimal instruments nonparametrically for their respective models and showed that their estimators attain Chamberlain's (1987) bound.

In this paper, we extend all those results to the case where θ_o may contain unknown functions. Specifically, we consider the model:

$$E[\psi(Z; \mu_o; h_o(\cdot))|X] = 0, \tag{1.3}$$

where $\theta_o = (\mu_o; h_o)$ are the parameters of interest, $\mu_o \in \Lambda$ is a vector of unknown parameters, and $h_o(\cdot) = (h_{o1}(\cdot); \cdots; h_{oq}(\cdot))' \in H$ is a vector of unknown functions. The dot "$\cdot$" in h_{oj} denotes the generic argument of $h_{oj}(\cdot)$ which may vary with applications and the subscript j. In some applications, for instance, the argument of h_{oj} is a subset of Z, or some known functions of Z up to unknown μ_o, while in other applications the argument includes other unknown functions $h_{oj'}$. The

[1] The optimal instruments are often unknown functions of X.

argument may even include unobserved variables. In general, the argument of h_{oj} can be written as $\pm_j(Z;\circledR_o;U)$, where $\pm_j(\cent;\cent;\cent)$ is a vector of known functions and U are unobserved variables. But for exposition we use "$\cent$" to denote the argument.

The extension to model (1.3) is needed because economic theories seldom produce exact functional forms and misspeci…cations in functional forms may lead to inconsistent parameter estimates. By specifying the model partially (i.e. including h_o as part of the unknown parameters), we can at least alleviate the inconsistency problem. In fact, many studies have already investigated some particular versions of (1.3). For example, Robinson (1988) and Fan, Härdle and Mammen (1998) studied the partially additive form $E[Y ¡ X_0^{\prime}\mu_o ¡ \sum_{j=1}^{q} h_{oj}(X_j)|X] = 0$, where the argument X_j of h_{oj} is a subset of conditioning variables $X = (X_0^{\prime};:::;X_q^{\prime})^{\prime}$. Powell, Stock and Stoker (1989) and Ichimura (1993) considered the index form $E[Y ¡ X_1^{\prime}\mu_{o1} ¡ h_o(X_2^{\prime}\mu_{o2})|X_1;X_2] = 0$, where the argument $X_2^{\prime}\mu_{o2}$ of h_o is a known function of conditioning variables X up to unknown μ_o. One can easily modify Newey and Powell's (1989) nonparametric instrumental variables regression model into: $E[Y_1 ¡ X_1^{\prime}\mu_o ¡ h_o(Y_2)|X_1;X_2] = 0$, where the argument Y_2 of h_o is a subset of the endogenous variables $Y = (Y_1;Y_2)^{\prime}$. One can also write down a version of transformation model with unknown link: $E[h_o(Y) ¡ X_1^{\prime}\mu_o|X_1;X_2] = 0$; where h_o is monotonic with some normalization say $h_o(0) = 1$. Chamberlain (1992) considered a particular version of (1.3): $E[½(Z;\mu_o;h_{o1}(\pm_1(Z));\cent\cent\cent;h_{oq}(\pm_q(Z)))|X] = 0$, where the arguments $\pm_j(Z)$ of h_{oj} for $j = 1;:::;q$ are completely known functions of Z. Our next example will be a modi…ed version of the semiparametric Engel curve model of Blundell, Browning and Crawford (2000): $½(Z;\circledR_o) = (½_1(Z;\circledR_o);:::;½_r(Z;\circledR_o))$; $½_l(Z;\circledR_o) = Y_l ¡ h_{ol}(X_2 ¡ h_{o;r+1}(X_1^{\prime}\mu_{o1})) ¡ X_1^{\prime}\mu_{o2l}$ and $E[½_l(Z;\circledR_o)|X_1;X_2] = 0$ for $l = 1;:::;r$, where the argument $X_2 ¡ h_{o;r+1}(X_1^{\prime}\mu_{o1})$ of h_{ol} contains another unknown function $h_{o;r+1}$. Additional examples can be found in Powell (1994) and Horowitz (1998).

Under the assumption that model (1.3) identi…es the parameters of interest $\circledR_o$,[2] we propose an estimator of $\circledR_o$ that has desirable large sample properties. Our approach is to apply the method of generalized minimum distance. Heuristically, if the functional forms of $h_o(\cent)$ and the conditional distribution of Y given X,

[2] This paper mainly tries to provide root-n e¢cient estimators for μ_o rather than pointwise estimation of h_o. In some cases such as the semiparametric IV example and the transformation with unknown link example, the identi…cation of h_o is in certain weak metric sense rather than pointwise strong identi…cation, see Sections 2 and 3 for examples and precide de…nition of identi…cation.

$F_{Y|X}$, were known, then the functional form of the conditional expectation

$$m(x;\theta) \equiv \int \rho(y; x_z; \mu; h(\zeta))\, dF_{Y|X=x}(y)$$

would be known. The minimum distance estimation of μ_o would be the appropriate method since condition (1.3) implies:

$$\theta_o : \inf_{\theta=(\mu;h)\in\Theta\times H} E\left[m(X;\theta)'[\Sigma(X)]^{-1}m(X;\theta)\right] \tag{1.4}$$

where $\Sigma(X)$ is a positive definite matrix for any given X, (and " $'$ " denotes the matrix transpose in this paper). Since the functional forms of $h_o(\zeta)$ and $F_{Y|X}$ (hence $m(\zeta;\zeta)$) are unknown, we estimate $m(X;\theta)$ nonparametrically with $\hat{m}(X;\theta)$ using observations on $\rho(Z;\theta)$ and X. The true values θ_o would then be estimated by minimizing the sample analog of (1.4) with $m(X;\theta)$ replaced by $\hat{m}(X;\theta)$. However, the resulting estimator could be inconsistent or converging arbitrarily slowly when the parameter space H is too large. To address this problem, we follow the sieve literature (Grenander, 1981) by approximating H with a sequence of sieve spaces H_n. The sieve spaces are "computable" (often finite-dimensional) parameter spaces which increase with sample size and become dense in the original functional space H. We then estimate the true values θ_o by minimizing the sample analog of a nonparametric version of (1.4) with h restricted to the sieve spaces H_n:

$$\hat{\theta}_n = (\hat{\mu}_n; \hat{h}_n) : \min_{\theta=(\mu;h)\in\Theta\times H_n} \frac{1}{n}\sum_{i=1}^{n} \hat{m}(X_i;\theta)'[\hat{\Sigma}(X_i)]^{-1}\hat{m}(X_i;\theta); \tag{1.5}$$

where $\hat{\Sigma}(X)$ is a consistent estimator of $\Sigma(X)$.

Under a set of sufficient conditions, we show that: (1) the sieve estimator $\hat{h}_n$ is consistent under certain metric, and may attain the usual nonparametric convergence rates; and (2) the estimator $\hat{\mu}_n$ is $\sqrt{n}$ consistent and asymptotically normally distributed. These results extend the work of Shen (1997) and Chen and Shen (1998) to the semiparametric conditional moments framework. In addition, we show that: (3) the estimator for the asymptotic covariance of $\hat{\mu}_n$ is consistent and easy to compute; and (4) when the weighting matrix $\Sigma(X)$ is set to $\Sigma_o(X) \equiv Var[\rho(Z;\theta_o)|X]$ (assumed to be positive-definite), the estimator $\hat{\mu}_n$ (hereafter the optimally weighted minimum distance estimator) attains the semiparametric efficiency bound of model (1.3). While establishing the latter result, we generalize the efficiency bound result of Chamberlain's (1992) to the much more general model (1.3).

The literature on semiparametric e¢cient estimation is large. The papers most closely related to our approach are those by Severini and Wong (1992), Shen (1997), Hansen (1982) and Newey (1990, 1993). The …rst two sets of authors studied e¢cient estimation in the semiparametric maximum likelihood setting. Hansen proposed e¢cient estimation for the parametric unconditional moments restrictions (1.2), while Newey investigated e¢cient estimation of the parametric conditional moments restrictions (1.1). Our approach is an extension of all those studies to the semiparametric conditional moments restrictions (1.3). Most importantly, our paper provides the semiparametric e¢cient estimation of models contain unknown functions of endogenous variables, such as the semiparametric IV example and the transformation with unknown link example.

The rest of the paper is organized as follows. Section 2 formally introduces the estimator for model (1.3) and three examples as illustration: partially additive model with unknown heteroskedastic variance; semiparametric Engel curves estimation; semiparametric instrumental variables regression. Section 3 provides general results on consistency with convergence rates of $\hat{\theta}_n$. Section 4 derives the $\sqrt{n}$ asymptotic normal distribution of $\hat{\beta}_n$. Section 5 provides a consistent estimator for the asymptotic covariance of $\hat{\beta}_n$. Section 6 investigates the e¢ciency property of $\hat{\beta}_n$. Section 7 presents "low-level" su¢cient conditions to obtain convergence rates of $\hat{\theta}_n$ and e¢cient estimators of $\hat{\beta}_n$ for the three examples. Section 8 illustrates our estimation procedure by a Monte Carlo study and Section 9 concludes. All technical proofs are relegated to the Appendices.

2. Estimator and Examples

Throughout the paper, we assume that the sample observations $\{(Y_i; X_i) : i = 1; 2; ¢¢¢; n\}$ are drawn independently from the distribution of $(Y; X)$ on $\mathcal{Y} £ \mathcal{X}$ with $\mathcal{X} = [X_l; X_u]$ a compact subset of R^s, and that the unknown distribution of $(Y; X)$ satis…es the conditional moment restriction (1.3), where $½ : \mathcal{Z} £ \mathcal{A} ! R^r$ a given known mapping, $Z ´ (Y^{\prime}; X_z^{\prime})^{\prime} 2 \mathcal{Z} ´ \mathcal{Y} £ \mathcal{X}_z$, $X_z µ X$; and $®_o ´ (µ_o; h_o) 2 \mathcal{A} ´ £ £ \mathcal{H}$, with £ a compact (with nonempty interior) subset of R^b and $\mathcal{H} ´ \mathcal{H}^1 £ ¢¢¢ £ \mathcal{H}^q$ a functional space of q-dimensional continuous functions.

To implement the proposed procedure, we need an estimator of $m(x; ®)$. We use the kernel method described in Silverman (1986) to estimate $m(x; ®)$ for arbitrary $(x; ®)$. Let $K : R^s ! R$ denote a known symmetric function and a_n a

bandwidth satisfying $a_n \to 0$ as $n \to +1$. We estimate the density of X, f_X, by

$$\hat{f}_{X_i} = \frac{1}{(n-1)a_n^s} \sum_{j \neq i; j=1}^{n} K\left(\frac{X_i - X_j}{a_n}\right)$$

and the conditional expectation $m(X;\theta)$ by

$$\hat{m}(X_i;\theta) = \frac{[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1}^{n} \rho(Z_j;\mu;h(\cdot)) K\left(\frac{X_i - X_j}{a_n}\right)}{\hat{f}_{X_i}};$$

here the dot " $\cdot$ " in $h(\cdot)$ denotes the fact that h may depend on data and hence varies with observations Z_j.

As noted in previous studies (e.g. Ichimura and Lee (1991)), one drawback of the kernel estimator is that it may be biased at the boundary points. To avoid the bias, trimming is often introduced. Suppose that X_l and X_u are known. For some $0 < \tau < 1$, define $X_n = [X_l + a_n^{\tau}1; X_u - a_n^{\tau}1]$. We use the indicator function $1\{X \in X_n\}$ to trim the boundaries. The proposed estimator $\hat{\theta}_n = (\hat{\mu}_n; \hat{h}_n) \in A_n \times H_n$ is now formally defined as the solution to:

$$\min_{\theta \in A_n} \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \in X_n\} \hat{m}(X_i;\theta)'[\hat{S}(X_i)]^{-1} \hat{m}(X_i;\theta): \tag{2.1}$$

To compute the estimator, the boundary points X_l and X_u are unknown and must be replaced by some consistent estimates. The order statistics, for example, are such consistent estimates. Since the order statistics converge to the boundary points at the rate $\frac{1}{n}$, replacing X_l and X_u with the order statistics will not affect the asymptotic results of the proposed estimator. In addition, we must be specific about the sieve approximation H_n. The most popular sieves in econometrics are linear sieves, which can be written as $H_n = H_n^1 \times \cdots \times H_n^q$ with H_n^j given by

$$H_n^j \equiv \{h_{nj}(\cdot) = \sum_{k=1}^{k_{jn}} \gamma_{jk} B_k^j(\cdot) \text{ for all } \gamma_n^j = (\gamma_{j1}; \cdots; \gamma_{jk_{jn}}) \in \Gamma_n^j\};$$

where $\{B_k^j(\cdot), k = 1; 2; :::\}$ denotes a sequence of known base functions (e.g. Fourier series, orthogonal polynomials, splines, power series, wavelets, etc.). The unknown functions h_o; in this case, are replaced by $h_n = (h_{n1}; \cdots; h_{nq})'$. The proposed estimator is then computed by estimating the sieve coefficients $\gamma_n^j; j = 1; :::; q$ jointly with μ. Our estimation procedure and the large sample results below

permit any types of sieve approximations including neural networks and other nonlinear sieves. The sieve method is easy to implement and convenient here since h_o may enter $\psi(Z;\theta_o)$ nonlinearly. Moreover, the sieve method allows us to impose any known structures such as additivity and shape restrictions on h_o implied by the population restrictions (1.3)[3].

Throughout this paper, we denote the (pathwise) directional derivative of $\psi(Z;\theta_o)$ with respect to θ_o at direction $[\theta - \theta_o]$ as:

$$\psi_{\theta_o}[Z;\theta - \theta_o] \equiv \lim_{\tau \to 0} \frac{\left(\psi(Z;\theta(\tau)) - \psi(Z;\theta_o)\right)}{\tau}$$

where $\theta(\tau)$ is an one-dimensional continuous path in A satisfying $\theta(0) = \theta_o$ and $\theta(1) = \theta$. In many applications such as Robinson's (1988) partial linear regression and Ichimura's (1993) single index model, the directional derivative $\psi_{\theta_o}[Z;\theta - \theta_o]$ does not depend on the endogenous variables Y. Then the kernel estimation (hence the trimming) or any nonparametric estimation of the conditional expectations function $m(x;\theta) \equiv \int \psi(y;x_z;\mu;h(\cdot))dF_{Y|X=x}(y)$ is not needed. One can simply perform the following sieve generalized least squares regression:

$$\min_{\theta=(\mu;h)\in\Theta\times H_n} \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i;\theta)'[\hat{S}(X_i)]^{-1}\psi(Z_i;\theta), \tag{2.2}$$

which is much earlier to compute when s (the dimension of X) is big.

Throughout this paper, we shall consider three examples as illustration of the usefulness of our new results. The first example is a partially additive model with conditional heteroskedastic variance of unknown form. This example slightly extends the popular partially linear regression of Robinson (1988). Although this is a simple example, but as far as we know there is still no published results on semiparametric efficient estimators under unknown heteroskedastic variance[4]. The second example is semiparametric Engel curves estimation, which is a modification of the model studied by Blundell, Browning and Crawford (2000) in their empirical investigation but nevertheless is still consistent with the rational consumer's demand behavior as discussed in Blundell et al (2000). The third example

[3] See e.g., Newey, Powell and Vella (1999) for using the spline sieve to estimate unknown functions with additive structure, Chen and Conley (1999) for using the shape-preserving wavelet cardinal B-spline sieve to estimate a conditional isotropic covariance function.

[4] The estimators proposed by Robinson (1988) and Li (1999) are semiparametric efficient only under conditional homoskedastic variance.

is semiparametric instrumental regression, which can be regarded as an alternative to the empirical model studied by Newey, Powell and Vella (1999, page 584, equation (7.1)).

Example 2.1 (Partially additive regression with unknown conditional heteroskedastic variance):

$$Y_i = X_{0;i}^{0}\mu_o + \sum_{j=1}^{q} h_{oj}(X_{j;i}) + \varepsilon_i;$$
$$E[\varepsilon_i | X_i] = 0; \quad E[\varepsilon_i^2 | X_i] = \sigma_o(X_i);$$

where $\theta_o = (\mu_o; h_{o1}; :::; h_{oq})$; $Z = (Y; X_z)$ and $X_z = X$ = non-overlap union of $X_0; X_1; :::; X_q$: Without loss of generality, we assume that $dim(X_0) = b$, $dim(X_j) = s_j$ and $dim(X) = s \quad b + \sum_{j=1}^{q} s_j$. Obviously $dim(\mu) = b$ and $dim(h) = q$. It is easy to see that $E[\rho(Z; \theta_o) | X] = 0$ where $\rho(Z; \theta) = Y - X_0^{0}\mu - \sum_{j=1}^{q} h_j(X_j)$: Since the directional derivative

$$\rho_{\theta_o}[Z; \theta - \theta_o] = - X_0^{0}[\mu_1 - \mu_{o1}] - \sum_{j=1}^{q} [h_j(X_j) - h_{oj}(X_j)]$$

does not depend on the endogenous variable Y, we may apply the simpler sieve generalized least squares regression (2.2) to estimate $(\mu_o; h_{o1}; :::; h_{oq})$ instead of the method of generalized minimum distance (2.1).

Assume that $h_{oj} \in H^j = B_{p;1}^{m_j}(C_j)$; $(m_j > 0; 2 \quad p \quad 1; 0 < C_j < 1)$, a Besov ball which consists of functions in $L_p(R^{s_j})$ with $\|h\|_{B_{p;1}^{m_j}} \quad C_j$. There are several equivalent de…nitions of a Besov space, we follow the one given in Meyer (1990). Suppose there exists a multivariate scaling function ϕ which has a compact support and continuous derivatives up to order $r_j \geq m_j$. Then there also exist $2^{s_j} - 1$ associated wavelets $\psi^e; e \in \{1; :::; 2^{s_j} - 1\}$ such that for a given integer k_0, $\{\phi_{k_0;i}; \psi_{k;l}^e : k \geq k_0; (i; l) \in Z^{2s_j}; e \in \{1; :::; 2^{s_j} - 1\}\}$ is an orthonormal basis of $L_2(R^{s_j})$, where $\phi_{k;i}(x) = 2^{k=2}\phi(2^k x - i)$ and $\psi_{k;i}^e(x) = 2^{k=2}\psi^e(2^k x - i)$ for $i \in Z^{s_j}$, we have for any $h \in L_2(R^{s_j})$,

$$h = \sum_{i \in Z^{s_j}} a_{k_0;i}\phi_{k_0;i} + \sum_{k \geq k_0} \sum_{i \in Z^{s_j}} \sum_{1 \ e \ 2^{s_j} - 1} c_{k;i}^e \psi_{k;i}^e$$

Now for $m_j > 0; 1 \quad p \quad 1; h \in B_{p;1}^{m_j}(R^{s_j})$ if and only if

$$\|h\|_{B_{p;1}^{m_j}} = \|a_{0;:}\|_p + \sup_{k \geq 0} 2^{k(m_j + s_j(1/2 - 1/p))} \|c_{k;:}\|_p < 1$$

where

$$\|a_{0,\cdot}\|_p^p = \sum_{i \in Z^{s_j}} |a_{0,i}|^p; \quad \|c^j_{k,\cdot}\|_p^p = \sum_{i \in Z^{s_j}} \sum_{1 \le e \le 2^{s_j}-1} |c^e_{k,i}|^p$$

We consider the wavelet sieve space:

$$H^j_n = \left\{ h \in L_2(R^{s_j}) : h = \sum_{k=0}^{K_n} \sum_{i \in Z^{s_j}} \sum_{1 \le e \le 2^{s_j}-1} c^e_{k,i} \tilde{A}^e_{k,i} : \|h\|_{B^{m_j}_{p,1}} \le C_j \right\} \tag{2.3}$$

In (2.2), we let $\hat{\Sigma}$ be the identity matrix and H^j_n, $j = 1, \ldots, q$, be the wavelet sieve space (2.3). Then by Corollary 3.2 in Section 3, the root-mean square convergence rate for the wavelet sieve estimator of h_{oj}, $j = 1, \ldots, q$, will be $O_p(n^{-m_j/(2m_j+s_j)})$, and by Theorem 4.1, the corresponding estimator of μ_o will be $\sqrt{n}$ asymptotic normal. Moreover when $\hat{\Sigma}(X)$ in (2.2) is a consistent estimator of $\Sigma_o(X)$ as provided in Section 6, then by Theorems 6.1 and 6.2 in Section 6, the corresponding estimator of μ_o will be $\sqrt{n}$ efficient with asymptotic variance V_o^{-1}:

$$V_o = \inf_{w_j} E\left\{ [X_0 - \sum_{j=1}^{q} w_j(X_j)]' \Sigma_o(X)^{-1} [X_0 - \sum_{j=1}^{q} w_j(X_j)] \right\} \tag{2.4}$$

where $w = (w_1, \ldots, w_q)$; w_j measurable function of X_j, and $w_j \in B^{m_j}_{p,1}(C_j) - \{h_{oj}\}$ for all $j = 1, \ldots, q$.

Example 2.2 (Partially additive IV regression with unknown conditional heteroskedastic variance, see Chamberlain (1992, model (3.5), page 579)):

$$\begin{aligned} Y_{1i} &= Y_{2i}'\mu_o + \sum_{j=1}^{q} h_{oj}(X_{j,i}) + \varepsilon_i; \\ E[\varepsilon_i | X_i] &= 0; \quad E[\varepsilon_i^2 | X_i] = \Sigma_o(X_i); \end{aligned}$$

where $\theta_o = (\mu_o, h_{o1}, \ldots, h_{oq})$; $Z = (Y, X_z)$; $Y = (Y_1, Y_2')'$; X_z = non-overlap union of $X_1, \ldots, X_q$; and $X = (X_z', X_0')'$. Without loss of generality, we assume that $\dim(X_0) = \dim(Y_2) = b$, and the rest of the dimensions and the assumptions on h_{oj} are the same as those in Example 2.1. It is easy to see that $E[\rho(Z, \theta_o)|X] = 0$ where $\rho(Z, \theta) = Y_1 - Y_2'\mu - \sum_{j=1}^{q} h_j(X_j)$. Since the directional derivative

$$\rho_{\theta_o}[Z, \theta - \theta_o] = -Y_2'[\mu_1 - \mu_{o1}] - \sum_{j=1}^{q} [h_j(X_j) - h_{oj}(X_j)]$$

depend on the endogenous variable Y_2, we apply the generalized minimum distance (2.1) to estimate $(\mu_o; h_{o1}; :::; h_{oq})$. By Lemma A.1 in Appendix, we have $\hat{m}(X;\alpha) - m(X;\alpha) = \{\hat{E}[Y_1|X] - E[Y_1|X]\} - \mu'\{\hat{E}[Y_2|X] - E[Y_2|X]\} = o_p(n^{-1=4})$ uniformly over $X \in \mathcal{X}_n; \alpha \in \mathcal{A}_n$. In (2.1), we let $\hat{\Sigma}$ be the identity matrix and H_n^j, $j = 1; :::; q$; be the wavelet sieve space (2.3). Then by Theorem 3.1 in Section 3, the root-mean square convergence rate for the wavelet sieve estimator of h_{oj}, $j = 1; :::; q$; will be $o_p(n^{-1=4})$, and by Theorem 4.1, the corresponding estimator of μ_o will be $\sqrt{n}$ asymptotic normal. Moreover when $\hat{\Sigma}(X)$ in (2.1) is a consistent estimator of $\Sigma_o(X)$ as provided in Section 6, then by Theorems 6.1 and 6.2 in Section 6, the corresponding estimator of μ_o will be $\sqrt{n}$ efficient with asymptotic variance V_o^{-1}:

$$V_o = \inf_{w_j(\cdot)} E\left[\{E[Y_2|X] - \sum_{j=1}^{q} w_j(X_j)g\}'[\Sigma_o(X)]^{-1}\{E[Y_2|X] - \sum_{j=1}^{q} w_j(X_j)g\}\right] \quad (2.5)$$

where $w = (w_1; :::; w_q)$ with w_j measurable function of X_j, and $w_j \in B_{p;1}^{m_j}(C_j)$ for all $\{h_{oj}\}$ for all $j = 1; :::; q$.

Example 2.3 (Semiparametric sample selection): for $l = 1; 2$
does not depend on the endogenous variable Y, we may apply the simpler sieve generalized least squares regression (2.2) to estimate $(\mu_{o1}; \mu_{o21}; \mu_{o22}; h_{o1}; h_{o2})$ instead of the method of generalized minimum distance (2.1).

Example 2.4 (Semiparametric IV regression):

$$\begin{aligned} Y_{1i} &= X_{1i}'\mu_o + h_o(Y_{2i}) + \varepsilon_i; \\ E[\varepsilon_i|X_{1i}; X_{2i}] &= 0; \quad E[\varepsilon_i^2|X_{1i}; X_{2i}] = \Sigma_o(X_{1i}; X_{2i}); \end{aligned}$$

where $\alpha_o = (\mu_o; h_o)$; $Z = (Y; X_z)$; $Y = (Y_1; Y_2)'$; $X_z = X_1$ and $X = (X_1'; X_2)'$. Without loss of generality, we assume that $\dim(X_1) = b$, $\dim(X_2) = 1$ and $\dim(X) = s = b + 1$, also $\dim(Y_1) = \dim(Y_2) = 1$, $\dim(\mu) = b$ and $\dim(h) = 1$. It is easy to see that $E[\psi(Z;\alpha_o)|X] = 0$ where $\psi(Z;\alpha) = Y_1 - X_1'\mu - h(Y_2)$: Since the directional derivative

$$\psi_{\alpha_o}[Z;\alpha - \alpha_o] = -X_1'[\mu - \mu_o] - [h(Y_2) - h_o(Y_2)]$$

depends on the endogenous variable Y_2, we apply the method of generalized minimum distance (2.1) to estimate $(\mu_o; h_o)$. By Lemma A.1 in Appendix, we have $\hat{m}(X;\alpha) - m(X;\alpha) = \{\hat{E}[Y_1|X] - E[Y_1|X]\} - \{\hat{E}[h(Y_2)|X] - E[h(Y_2)|X]\} = o_p(n^{-1=4})$ uniformly over $X \in \mathcal{X}_n; \alpha \in \mathcal{A}_n$. In (2.1), we let $\hat{\Sigma}$ be the identity

matrix and H_n be the sieve. Then by Theorem 3.1 in Section 3, the convergence rate in certain metric (to be specified later) for the sieve estimator of h_o will be $o_p(n^{-1/4})$, and by Theorem 4.1, the corresponding estimator of μ_o will be $\sqrt{n}$ asymptotic normal. Moreover when $\hat{\Sigma}(X)$ in (2.1) is a consistent estimator of $\Sigma_o(X)$ as provided in Section 6, then by Theorems 6.1 and 6.2 in Section 6, the corresponding estimator of μ_o will be $\sqrt{n}$ efficient with asymptotic variance V_o^{-1}:

$$V_o = \inf_w E\left[\{X_1^* - E[w(Y_2)|X]\}[\Sigma_o(X)]^{-1}\{X_1^* - E[w(Y_2)|X]\}'\right] \tag{2.6}$$

where w is a measurable function of Y_2, and $w \in B^m_{p,1}(C) - \{h_o\}$.

3. Consistency

In this section, we first present a set of sufficient conditions, then discuss those conditions, and finally state the theorem on consistency and rates of convergence of $\hat{\theta}_n$. We begin by introducing notations which will aid presentation of sufficient conditions. Let $\|\cdot\|_2$ denote a pseudo-norm on A, and $N(A_n, \varepsilon)$ the minimum bracket-number of radius ε balls that cover the sieve space A_n. In this paper, we require the pseudo-norm $\|\alpha - \alpha_o\|_2$ to be locally equivalent to $\sqrt{K(\alpha_o, \alpha)}$, where

$$K(\alpha_o, \alpha) \equiv E\{1\{X \in X_n\}[m(X, \alpha) - m(X, \alpha_o)]'\Sigma(X)^{-1}[m(X, \alpha) - m(X, \alpha_o)]\}/2.$$

That is, there exist two constants $c_1, c_2 > 0$ such that $c_1\sqrt{K(\alpha_o, \alpha)} \leq \|\alpha - \alpha_o\|_2 \leq c_2\sqrt{K(\alpha_o, \alpha)}$ hold for any α in a neighborhood of α_o. For example, when $\rho(Z, \alpha)$ is Lipschitz continuous with respect to α, we can choose the pseudo-metric $\|\alpha - \alpha_o\|_2^2$ to be $E[(\mu - \mu_o)'\omega_\mu(Z)(\mu - \mu_o)] + \sum_{j=1}^q E[\omega_{hj}(Z)\{h_j(\tau) - h_{oj}(\tau)\}^2]$ for some known weighting functions $\omega_\mu(Z)$ and $\omega_{hj}(Z)$. To avoid confusion, all vectors refer to column vectors unless stated otherwise.

An estimator $\hat{\theta}_n$ is said to be consistent for θ_o if $\|\hat{\theta}_n - \theta_o\|_2 \to 0$ in probability, denoted as $\|\hat{\theta}_n - \theta_o\|_2 = o_p(1)$. Let $\{\delta_n\}$ denote a positive sequence that decreases to zero as $n \to \infty$. $\hat{\theta}_n$ is said to be consistent for θ_o at a rate (strictly) faster than δ_n if $\|\hat{\theta}_n - \theta_o\|_2/\delta_n \to 0$ in probability, denoted as $\|\hat{\theta}_n - \theta_o\|_2 = o_p(\delta_n)$. Our goal in this section is to show that $\hat{\theta}_n$ is consistent at a rate faster than $n^{-1/4}$. To attain this goal, we need conditions to ensure that H_n converges to H, $\hat{m}(X, \alpha)$ converges to $m(X, \alpha)$ uniformly over X and α, and $\hat{\Sigma}(X)$ converges to $\Sigma(X)$ uniformly over X at rates faster than $n^{-1/4}$. We also need conditions on the local curvature of $\rho(\tau, \tau)$. Below we present such conditions.

Assumption 1. The data $f(Y_i; X_i) : i = 1; 2; :::; ng$ are i.i.d..

Assumption 2. $E[½(Z; ®)jX] = 0$ holds for any $® 2 A$ i¤ $k® ¡ ®_0 k_2 = 0$.

Assumption 3. (i) $\hat{§}(X) = §(X) + o_p(n^{¡1=4})$ uniformly over $X 2 X_n$; and (ii) there exist some positive constants c_1 and c_2 such that $c_1 \quad \min_¸(§(X)) \quad \max_¸(§(X)) \quad c_2$ for all $X 2 X$, where $¸(§(X))$ denote eigenvalues of $§(X)$.

Assumption 4. (i) The density f_X is strictly positive on the compact support X; and (ii) f_X has up to J-th derivatives, where $J ¸ 2$.

Assumption 5. $K(x)$: (i) is a bounded function with bounded support, and Lipschitz continuous; (ii) is symmetric around origin, and integrates to one; and (iii) is of order $J ¸ 2$, i.e.

$$\int (x^1)^{j_1} ¢¢¢ (x^s)^{j_s} K(x)dx = 0 \text{ for } 1 \quad j_1 + ¢¢¢ + j_s \quad J ¡ 1;$$

$$\int (x^1)^{j_1} ¢¢¢ (x^s)^{j_s} K(x)dx \neq 0 \text{ for } j_1 + ¢¢¢ + j_s = J:$$

where x^i denotes the i-th element of x.

Assumption 6. $E[½(Z; ®)jX]$ has up to J-th (total) derivatives with respect to X; and the J-th derivative (w.r.t. X) is uniformly bounded in $(X; ®) 2 X £ A_n$.

Assumption 7. (i) There exists a measurable function $Á_1 : Z ! [0; 1)$ such that $E[Á_1(Z)^p] < 1$ for some integer $p ¸ 2$ and $j½(Z; ®)j \quad Á_1(Z)$ for all $® 2 A_n$; and (ii) there exist a sequence of positive numbers $fc_n g$ and a measurable function $Á_2 : Z ! [0; 1)$ such that $E[Á_2(Z)jX] < 1$ and

$$j½(Z; ®_1) ¡ ½(Z; ®_2)j \quad c_n Á_2(Z) k®_1 ¡ ®_2 k_2 \quad \text{for all } ®_1; ®_2 2 A_n:$$

Assumption 8. (i) $\dfrac{\sqrt{n a_n^s}}{\ln[N(A_n; c_n^{¡1} n^{¡1=4}) £ a_n^{¡s(s+1)} n^{s=4}] £ \max\{1; n^{(2¡p)=4p} £ a_n^{¡2(s+1)=p}\}} ! +1$,

and (ii) $n a_n^{4J} ! 0$ as $n ! +1$.

Assumption 9. For any $® 2 A$, there exists $¦_n ® 2 A_n$ such that $k® ¡ ¦_n ® k_2 = o(n^{¡1=4})$ as $n ! 1$.

Assumption 10. $_n = o(n^{¡1=4})$, where $_n 2 (0; 1]$ denote the solution to:

$$\inf\left\{ 2 (0; 1] : \frac{1}{\sqrt{2}} \int_{2}^{} \sqrt{\log N\left(A_n; \sqrt{p}\right)} d! \quad O(\sqrt{n}) \right\};$$

In the following, we use $@_o(") = f® 2 A_n : k® ¡ ®_o k_2 \quad "g$ to denote a (radius "-) neighborhood of $®_o$, and $U_n = f\frac{® ¡ ®_o}{k® ¡ ®_o k_2} : ® 2 @_o(o(n^{¡ 1=8}))g$ the unit circle.

Assumption 11. (i) $E f½_{®_o}[Z; À]jXg$ has up to J-th (total) derivatives w.r.t X; and the J-th derivative (w.r.t X) is uniformly bounded over $X 2 X$, and $À 2 U_n$; (ii) there exists a measurable function $Á_3(Z)$ such that $E[Á_3(Z)^p]$ is …nite and $j½_{®_o}[Z; À]j \quad Á_3(Z)$ holds for any $Z 2 Z$ and $À 2 U_n$; and (iii) there exist a sequence of positive numbers fc_ng and a measurable function $Á_4(Z)$ such that $E[Á_4(Z)jX]$ is …nite and

$$j½_{®_o}[Z; À_1] ¡ ½_{®_o}[Z; À_2]j \quad c_n Á_4(Z) kÀ_1 ¡ À_2 k_2$$

holds for any $Z 2 Z$ and $À_1; À_2 2 U_n$.

De…ne the linear approximation error by

$$r_½[Z; ® ¡ ®_o] ´ ½(Z; ®) ¡ ½(Z; ®_o) ¡ ½_{®_o}[Z; ® ¡ ®_o]:$$

Assumption 12. (i) there exists a measurable function $Á_5(Z)$ such that $E[Á_5(Z)^p jX]$ exists and

$$jr_½[Z; ® ¡ ®_o]j \quad Á_5(Z) k® ¡ ®_o k_2^2$$

holds for any $Z 2 Z$ and $® 2 @_o(o(n^{¡ 1=8}))$;

(ii) there exist a sequence of positive numbers fc_ng and a measurable function $Á_6(Z)$ such that $E[Á_6(Z)jX]$ is …nite and

$$j\frac{r_½[Z; ®_1 ¡ ®_o]}{k®_1 ¡ ®_o k_2} ¡ \frac{r_½[Z; ®_2 ¡ ®_o]}{k®_2 ¡ ®_o k_2}j \quad c_n Á_6(Z) k®_1 ¡ ®_2 k_2$$

holds for any $Z 2 Z$ and $®_1; ®_2 2 @_o(o(n^{¡ 1=8}))$.

Theorem 3.1. Suppose Assumptions 1-12 are satis…ed. If $b®_n$ is the solution to (2.1), then $jjb®_n ¡ ®_o jj_2 = o_p(n^{¡ 1=4})$:

Corollary 3.2. Suppose Assumptions 1-3, 7(ii), 9-10 are satis…ed, and that the directional derivative $½_{®_o}[Z; ® ¡ ®_o]$ does not depend on the variables Y. If $b®_n$ is the solution to (2.2), then $jjb®_n ¡ ®_o jj_2 = o_p(n^{¡ 1=4})$:

Discussions: (1) Assumptions 1, 2 and 3 are basic regularity conditions. Assumption 1 rules out dependent data. This condition, though restrictive, is not critical for our results on consistency of $b®$ and $\sqrt{n}$ asymptotic normality of $b¯$.

These results still can be proved for weakly dependent data using the approach developed here and those developed in Chen and Shen (1998). The independence assumption is, however, critical for the semiparametric e¢ciency result. Assumption 2 is a global identi…cation condition, and must be satis…ed for our results to hold. Veri…cation of this condition requires the knowledge of structures of speci…c models. Note that this condition only identi…es θ_o up to the equivalence class with respect to the norm $\|\cdot\|_2$. In Section 4, we shall impose additional condition to identify μ_o uniquely. Assumption 3(i) requires that the estimator of the weighting matrix converges to some positive de…nite matrix uniformly over the regressors at a rate faster than $n^{-1/4}$. This assumption is not restrictive and can be satis…ed by many estimators. For example, the identity weighting matrix satis…es this condition. The kernel estimator proposed in Section 6 also satis…es this condition. Assumption 3(ii) requires the weighting matrix to be bounded from above and below. Conditions of this sort are common in the weighted Least Squares regression literature.

(2) Assumptions 4-8, 11 and 12 are conditions related to the kernel estimation. They (except for 7(ii)) should be dropped whenever the initial estimation of $m(X;\theta)$ is not needed. Assumptions 4 - 8 imply that the kernel estimator $\hat{m}(X;\theta)$ converges to $m(X;\theta)$ at a rate faster than $n^{-1/4}$ uniformly over $X \in \mathcal{X}_n; \theta \in \Theta_n$ (see Corollary A.2 in the Appendix A). These conditions together with Assumption 3(i) imply that the kernel-estimated sample criterion function in (2.1) converges to $\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\} g m(X_i;\theta)'[\Sigma(X_i)]^{-1} m(X_i;\theta)$ uniformly over θ at a rate faster than $n^{-1/4}$. Under additional Assumptions 11 and 12, the convergence rate at a local neighborhood of true θ_o is even faster, (see the proof of Theorem 3.1 in the Appendix B for details).

Assumption 4(i) rules out unbounded regressors and zero density values. This condition can be dropped when trimming is used to control for small density values of f_X, see e.g. Andrews (1995). Assumption 4(ii) is a smoothness condition. It is not critical for the consistency of $\hat{\theta}_n$ but critical for the $\sqrt{n}$ consistency of $\hat{\mu}_n$. Nevertheless, it is familiar in the semiparametric literature. Assumption 5 also is familiar in the semiparametric econometric literature. It is satis…ed by many commonly used kernel functions. Assumption 6 requires $m(\cdot)$ to be su¢ciently smooth with respect to X. Conditions of this sort are familiar in the semiparametric literature, (e.g. Powell et al. (1989), Horowitz (1996)). Again, this condition is not critical for the consistency of $\hat{\theta}_n$ but critical for the $\sqrt{n}$ consistency of $\hat{\mu}_n$. Assumption 7(i) is an envelope condition which resembles the envelope condition commonly imposed in the parametric literature. The envelope condition plays

an important role in the parametric asymptotics literature and it is expected to play a similar role in this paper. Assumption 7(ii) is a Lipschitz condition which requires that the derivative satisfy an envelope condition. Assumption 8 imposes restrictions on the rates at which the kernel bandwidth $a_n \to 0$. Assumption 8(i) controls the variance component of the kernel estimator and Assumption 8(ii) controls the bias component, both converging at a rate faster than $n^{-1/4}$ uniformly over $x \in \mathcal{X}_n, \alpha \in \mathcal{A}_n$. Assumptions 11 and 12 are conditions on the local curvature of $\rho(\tau;\tau)$. Assumption 11 implies that, on the unit circle centered at α_0, the directional derivative function is smooth, satisfies an envelope condition and is Lipschitz continuous. While Assumption 12 requires that, in a neighborhood of α_0, the linear approximation error is sufficiently smooth, bounded by a quadratic function of $\alpha - \alpha_0$, and Lipschitz continuous.

(3) Assumptions 7, 9 and 10 are conditions related to the sieve estimation of α_0. Assumption 9 requires that the sieve approximation error must converge to zero faster than $n^{-1/4}$, while Assumption 10 requires that the size of the sieves $\mathcal{A}_n$ should not be too large. To verify these conditions, we must compute the sieve approximation error $\|\alpha - \Pi_n\alpha\|_2$ and the bracketing number $N(\mathcal{A}_n;\varepsilon)$. While both depend on the smoothness and dimensionality of $\mathcal{A}_n \times \Xi \times \mathcal{H}_n$, $\|\alpha - \Pi_n\alpha\|_2$ also depends on the structure of $\mathcal{A} \times \Xi \times \mathcal{H}$. Let $|\mu - \mu_0|$ denote the Euclidean norm on Ξ and $d(h;h_0)$ a distance metric on $\mathcal{H}$ such that the product metric $|\mu - \mu_0| \vee d(h;h_0)$ is locally equivalent to $\|\alpha - \alpha_0\|_2$ on $\mathcal{A}$. Let $N(\mathcal{H}_n;\varepsilon)$ denote the minimum bracketing number of radius ε balls (in d distance) covering the sieve space $\mathcal{H}_n$. Then obviously $\|\alpha - \Pi_n\alpha\|_2 = d(h;\Pi_n h)$; the approximation error rate between $h \in \mathcal{H}$ and its corresponding sieve $\Pi_n h \in \mathcal{H}_n$. And $N(\mathcal{A}_n;\varepsilon) = N(\mathcal{H}_n;\varepsilon) \times \varepsilon^{-b}$; where ε^{-b} is the minimum number of radius ε balls (in Euclidean distance) covering Ξ. For many sieves, calculations of $d(h;\Pi_n h)$ and $N(\mathcal{H}_n;\varepsilon)$ can be found in DeVore and Lorentz (1993) and Van der Vaart and Wellner (1996). Such calculations can also be found in econometrics applications such as Fenton and Gallant (1996) for Hermite polynomials, Newey (1997) for spline and power series, Chen, Hansen and Scheinkman (1997) for shape-preserving wavelet B-spline with unbounded support, and Chen and White (1999) for multivariate neural networks.

Theorem 3.1 shows that the proposed estimator is consistent at a rate faster than $n^{-1/4}$. This result is stronger than the simple consistency $\|\hat{\alpha}_n - \alpha_0\|_2 = o_p(1)$. The stronger result is needed for obtaining the $\sqrt{n}$ consistency of $\hat{\beta}_n$, to which we now turn to.

4. Asymptotic Normality

The standard approach for deriving the asymptotic distribution is to assume that $\hat{\theta}_n = (\hat{\mu}_n, \hat{h}_n)$ is an interior solution and then to apply Taylor expansion. This approach may not work well here for two reasons. First, while it is reasonable to expect $\hat{\mu}_n$ to be an interior solution, $\hat{h}_n$ often is not and hence the first order condition may not hold. Second, even if the first order condition is satisfied, we need to compute the projection of the derivatives with respect to μ onto the derivatives with respect to h. Because h is infinite dimensional, computing such projection is often complicated and may not have a close-form solution. An alternative approach is to apply the Riesz representation theorem. The Riesz representation theorem links $\hat{\mu}_n$ to the derivatives of the sample criterion function without relying on the first order condition and hence does not have the problems mentioned above. This approach was used in Shen (1997) and Chen and Shen (1998) and is adopted in this paper.

To derive the asymptotic distribution of $\hat{\mu}_n$, it suffices to derive the asymptotic distribution of $f(\hat{\theta}_n) = \lambda' \hat{\mu}_n$ for any fixed non-zero $\lambda \in R^b$. Let

$$m_{\theta_o}[X; \theta] \equiv E[\psi_{\theta_o}[Z; \theta] | X]$$

denote the conditional expectation of the directional derivative. Define a Hilbert-norm on A as

$$\|\theta - \theta_o\|^2 \equiv E\left[m_{\theta_o}[X; \theta - \theta_o]' \Sigma(X)^{-1} m_{\theta_o}[X; \theta - \theta_o]\right].$$

Let $\bar{A}$ denote $\Theta \times \overline{H}$ with $\overline{H}$ the linear completion of H under the norm $\|\cdot\|$, and let $\langle \cdot, \cdot \rangle$ denote the inner product induced by the norm $\|\cdot\|$ on $\bar{A}$. Since $f(\theta)$ is a linear functional on $\bar{A}$, it is bounded (i.e. continuous) if and only if

$$\sup_{\|\theta - \theta_o\| > 0; \theta \in \bar{A}} \frac{|f(\theta) - f(\theta_o)|}{\|\theta - \theta_o\|} < 1.$$

The Riesz representation theory states that there exists a representor $v^* \in \bar{A}$ satisfying $\|v^*\| = \sup_{\|\theta - \theta_o\| > 0; \theta \in \bar{A}} \frac{|f(\theta) - f(\theta_o)|}{\|\theta - \theta_o\|}$ and

$$f(\theta) = f(\theta_o) + \langle \theta - \theta_o, v^* \rangle.$$

Hence, $f(\hat{\theta}_n) - f(\theta_o) = \langle \hat{\theta}_n - \theta_o, v^* \rangle$. Our aim in this section is to compute the representor v^* and then link $f(\hat{\theta}_n) - f(\theta_o)$ to the pathwise directional derivatives of the sample criterion function.

First, we compute the representor $v^\star$. Let $W = H - h_o$ and let $\overline{W}$ denote the linear completion of W under the Hilbert norm. For any $h \in \overline{H}$, there exist $w^j(\cdot) \in \overline{W}$ for $j = 1,...,b$ such that $h - h_o = -(w^1(\cdot), \cdots, w^b(\cdot))(\mu - \mu_o) = -w(\cdot)(\mu - \mu_o)$. Let $\partial_h[Z; \cdot h]$ denote the directional derivative with respect to h at direction $\cdot h$. Define $\partial_{h_o}[Z; w(\cdot)] = (\partial_{h_o}[Z; w^1(\cdot)], ..., \partial_{h_o}[Z; w^b(\cdot)])$ and $D_w(X)$ the $r \times b$ matrix valued function with the $j-$th column $D_{w^j}(X); j = 1,...,b$ defined as $D_{w^j}(X) \equiv E\left[\partial_{\mu_j}(Z; \theta_o) - \partial_{h_o}[Z; w^j(\cdot)] | X\right]$. Then

$$D_w(X) \equiv E\left[\partial_{\mu^0}(Z; \theta_o) - \partial_{h_o}[Z; w(\cdot)] | X\right] \tag{4.1}$$

where $\partial_{\mu^0}(Z; \theta)$ is the ordinary derivative with respect to $\mu^0 = (\mu_1, ..., \mu_b)$. By definition we have $m_{\theta_o}[X; \theta - \theta_o] = D_w(X)(\mu - \mu_o)$, and

$$\|\theta - \theta_o\|^2 = (\mu - \mu_o)' E\left[D_w(X)' \Sigma(X)^{-1} D_w(X)\right](\mu - \mu_o).$$

Let $w^\star(\cdot) = (w^{\star 1}(\cdot), ..., w^{\star b}(\cdot))$ be the solution to

$$\inf_{w^j(\cdot) \in \overline{W}, j=1,...,b} E[D_w(X)' \Sigma(X)^{-1} D_w(X)] \tag{4.2}$$

where "inf " is in positive semidefinite matrix sense. Simple calculation yields

$$\begin{aligned} \|v^\star\|^2 &= \sup_{\mu - \mu_o \neq 0} \frac{(\mu - \mu_o)' \Gamma \Gamma' (\mu - \mu_o)}{(\mu - \mu_o)' [\inf_w E[D_w(X)' \Sigma(X)^{-1} D_w(X)]] (\mu - \mu_o)} \\ &= \Gamma' E[D_{w^\star}(X)' \Sigma(X)^{-1} D_{w^\star}(X)]^{-1} \Gamma. \end{aligned} \tag{4.3}$$

Clearly, $\|v^\star\| < 1$ if and only if $E[D_{w^\star}(X)' \Sigma(X)^{-1} D_{w^\star}(X)]$ is positive-definite. The Riesz representor $v^\star = (v^\star_\mu; v^\star_h)$ is now given by

$$\begin{aligned} v^\star_\mu &= \left\{E\left[D_{w^\star}(X)' [\Sigma(X)]^{-1} D_{w^\star}(X)\right]^{-1}\right\} \Gamma, \\ v^\star_h &= -w^\star(\cdot)\left\{E\left[D_{w^\star}(X)' [\Sigma(X)]^{-1} D_{w^\star}(X)\right]^{-1}\right\} \Gamma. \end{aligned}$$

Hence, $m_{\theta_o}[X_i; v^\star] = D_{w^\star}(X)\left(E\left[D_{w^\star}(X)' [\Sigma(X)]^{-1} D_{w^\star}(X)\right]\right)^{-1} \Gamma$.

Next, we link $f(\hat{\theta}_n) - f(\theta_o)$ to the pathwise directional derivatives of the sample criterion function. After some calculation in the Appendix C, we obtain:

$$\sqrt{n}(f(\hat{\theta}_n) - f(\theta_o)) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_{\theta_o}[X_i; v^\star]' [\Sigma(X_i)]^{-1} \partial(Z_i; \theta_o) + o_p(1). \tag{4.4}$$

Applying the Lindeberg-Levy central limit theorem proves that $\sqrt{n}(f(\hat{\theta}_n) - f(\theta_o))$ is asymptotically normally distributed with mean zero and variance $\gamma' V^{-1} \gamma$, where

$$V = \begin{bmatrix} E\{D_w \rho(X)' [\sigma(X)]^{-1} D_w \rho(X)\} \\ (E\{D_w \rho(X)' [\sigma(X)]^{-1} \sigma_o(X) [\sigma(X)]^{-1} D_w \rho(X)\})^{-1} \\ E\{D_w \rho(X)' [\sigma(X)]^{-1} D_w \rho(X)\} \end{bmatrix} \quad (4.5)$$

and $\sigma_o(X) = Var[\psi(Z; \theta_o) | X]$. Hence, $\hat{\mu}_n$ is $\sqrt{n}$ consistent and asymptotically normally distributed with covariance V^{-1}.

We now present a set of conditions in addition to those in Section 3 to ensure validity of equation (4.4).

Assumption 13. There exists a measurable function $\dot{A}_7(Z)$ such that $E[\dot{A}_7(Z) | X]$ is …nite and

$$|r_\psi[Z; \theta_1 - \theta_o] - r_\psi[Z; \theta_2 - \theta_o]|$$
$$\dot{A}_7(Z) \cdot \|\theta_1 - \theta_2\| \cdot \max\{\|\theta_1 - \theta_o\|, \|\theta_2 - \theta_o\|\}$$

holds for any $Z \in \mathcal{Z}$ and $\theta_1, \theta_2 \in \mathcal{B}_o(o(n^{-1/4}))$:

Assumption 14. (i) $E[D_w \rho(X)' \sigma(X)^{-1} D_w \rho(X)]$ is …nite and positive-de…nite; and (ii) $\sigma_o(X)$ is …nite positive de…nite uniformly over $X \in \mathcal{X}$.

Assumption 13 is similar to but not implied by Assumption 12. It requires the moment functions $\psi(Z; \theta)$ to have an extra curvature in the neighborhood of θ_o. Assumption 14 is a local identi…cation condition of μ_o. Assumptions 2, 3 and 14 together uniquely identify μ_o. Moreover, they ensure that μ_o is $\sqrt{n}$ consistently estimated as we show in the next theorem.

Theorem 4.1. Suppose Assumptions 1-14 hold. Then $\sqrt{n}(\hat{\mu}_n - \mu_o)$ is asymptotically normally distributed with mean zero and covariance matrix V^{-1}.

Theorem 4.1 shows that the proposed estimator for the parametric component (μ_o) attains the parametric rates. It provides an asymptotic distribution for the parametric component, which is needed for statistical inference. The asymptotic covariance has three terms. If we can interpret $D_w \rho(X)$ as regressors in a system of equations with $\psi_{\mu'}(Z; \theta_o) - \psi_{h_o}[Z; w\rho]$ as dependent variables, then V^{-1} has an interpretation of the covariance of the weighted minimum distance estimator. The three terms in V correspond to the three terms in the covariance of the (incorrectly) weighted LS estimator in the presence of heteroskedasticity.

As we noted in the introduction, the nonparametric component (h_o) can be concentrated out analytically in some applications such as the partial linear regression (Robinson, 1988) and the single index model (Ichimura, 1993). Although the proposed procedure and Theorems 3.1 and 4.1 are applicable to those applications, it may be worth taking advantage of the concentrated out nonparametric component. Let $h^*(\zeta;\mu)$ denote the concentrated out nonparametric component, which satisfies $h^*(\zeta;\mu_o) = h_o(\zeta)$, and often depends on some unknown functions. Suppose that, for each fixed μ, $h^*(\zeta;\mu)$ is estimated by some nonparametric estimator $\hat{h}^*(\zeta;\mu)$ using kernel or sieve methods. The estimator, $\hat{\mu}_n$, for the parametric component is then obtained as the solution to

$$\min_{\mu\in\Gamma} \sum_{i=1}^{n} 1(X_i \in \mathcal{X}_n)\, m(X_i;\mu;\hat{h}^*(\zeta;\mu))'[\hat{\Sigma}(X_i)]^{-1} m(X_i;\mu;\hat{h}^*(\zeta;\mu)).$$

Suppose that $\hat{h}^*(\zeta;\mu)$ is differentiable with respect to μ and that $\hat{h}^*(\zeta;\mu)$ and its derivatives with respect to μ converge to the true values in probability uniformly at rates faster than $n^{-1/4}$. Then Theorem 4.1 can also be proved using the techniques developed in Andrews (1994), Newey (1994) or Pakes and Olley (1995).

5. Covariance Estimator

To conduct statistical inference on the parametric component, a consistent covariance estimator is needed. In this section, we provide such an estimator. We first present consistent estimators of $\Sigma(X)$; $\Sigma_o(X)$; and $D_{w^*}(X)$ respectively, and then present a consistent estimator of V. Consistent estimator of $\Sigma(X)$ already exists. To estimate $\Sigma_o(X)$, we use the kernel method:

$$\hat{\Sigma}_o(X_i) \equiv \frac{[(n-1)a_n^s]^{-1}\sum_{j\neq i;\, j=1}^{n} \psi(Z_j;\hat{\theta}_n)\psi(Z_j;\hat{\theta}_n)' K\left(\frac{X_i - X_j}{a_n}\right)}{f_{X_i}}.$$

To estimate $D_{w^*}(X)$, we estimate $w^*(\zeta)$. Recall that $D_{w^k}(X)$ and $w^k(\zeta)$ are the k-th columns of $D_w(X)$ and $w(\zeta)$ respectively, and that $w^{*k}(\zeta)$; $k = 1,...,b$ is the solution to:

$$\inf_{w^k(\zeta)\in\overline{W}} E\left[D_{w^k}(X)'[\Sigma(X)]^{-1}D_{w^k}(X)\right].$$

Hence, $w^{*k}(\zeta)$ can be estimated by applying a nonparametric least squares regres-

sion. Speci…cally, let μ_k denote the k-th element of μ and let

$$\hat{D}_{w^k}(X_i) \equiv \frac{[(n-1)a_n^s]^{-1}\sum_{j\neq i;j=1}^{n}\left\{\frac{\partial \rho(Z_j;\hat{h}_n)}{\partial \mu_k} - \rho_{\hat{h}_n}[Z_j;w^k(\tau)]\right\}K\left(\frac{X_i - X_j}{a_n}\right)}{\hat{f}_{X_i}}$$

denote the kernel estimator of $D_{w^k}(X_i)$: Let $\overline{W}_n$ denote the linear completion of $A_n - \iota_n\theta_o$, then $\overline{W}_n$ is a sieve approximation to $\overline{W}$. We now estimate $w^{*k}(\tau)$ by $\hat{w}^k(\tau)$, the solution to

$$\min_{w^k(\tau)\in\overline{W}_n}\sum_{i=1}^{n}1(X_i \in \mathcal{X}_n)\hat{D}_{w^k}(X_i)'[\hat{\Omega}(X_i)]^{-1}\hat{D}_{w^k}(X_i);$$

Let $\hat{V}^*$ denote the matrix whose k-th column is $\hat{w}^k(\tau)$. Then, $D_{w^*}(X)$ is estimated by $\hat{D}_{\hat{V}^*}(X)$ and V is estimated by

$$\hat{V} \equiv \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}1\{X_i \in \mathcal{X}_n\}\hat{D}_{\hat{V}^*}(X_i)'\hat{\Omega}^{-1}(X_i)\hat{D}_{\hat{V}^*}(X_i) & \\ \frac{1}{n}\sum_{i=1}^{n}1\{X_i \in \mathcal{X}_n\}\hat{D}_{\hat{V}^*}(X_i)'\hat{\Omega}^{-1}(X_i)\hat{\Omega}_o(X_i)\hat{\Omega}^{-1}(X_i)\hat{D}_{\hat{V}^*}(X_i) & \end{bmatrix}^{-1} \times \begin{bmatrix} \\ \end{bmatrix}$$
$$\frac{1}{n}\sum_{i=1}^{n}1\{X_i \in \mathcal{X}_n\}\hat{D}_{\hat{V}^*}(X_i)'\hat{\Omega}^{-1}(X_i)\hat{D}_{\hat{V}^*}(X_i)$$

Assumption 15. (i) There exists a measurable function $\acute{A}_8(Z)$ such that $E[\acute{A}_8(Z)|X]$ exists and is …nite, and $|\frac{\partial \rho(Z;\theta)}{\partial \mu_k} - \frac{\partial \rho(Z;\theta_o)}{\partial \mu_k}| \leq \acute{A}_8(Z)\|\theta - \theta_o\|_2$ for any $k = 1;:::;b$; $Z \in \mathcal{Z}$, $\theta \in \Theta_o(o(n^{-1=4}))$; and (ii) there exists a measurable function $\acute{A}_9(Z)$ such that $E[\acute{A}_9(Z)|X]$ exists and is …nite, and $|\rho_h[Z;v(\tau)] - \rho_{h_o}[Z;v(\tau)]| \leq \acute{A}_9(Z)\cdot\|\theta - \theta_o\|_2$ for any $Z \in \mathcal{Z}$, $v \in \overline{W}_n$ with $\|v\|_2 = 1$, $\theta \in \Theta_o(o(n^{-1=4}))$.

Assumption 16. (i) there exists a measurable function $\acute{A}_{10}(Z)$ such that $E[\acute{A}_{10}(Z)^p]$ is …nite and $|\rho(Z;\theta)\rho(Z;\theta)'| \leq \acute{A}_{10}(Z)$ for any $Z \in \mathcal{Z}$ and $\theta \in \Theta_o(o(n^{-1=4}))$; (ii) there exists a measurable function $\acute{A}_{11}(Z)$ such that $E[\acute{A}_{11}(Z)|X]$ is …nite and $|\rho(Z;\theta_1)\rho(Z;\theta_1)' - \rho(Z;\theta_2)\rho(Z;\theta_2)'| \leq \acute{A}_{11}(Z)\cdot\|\theta_1 - \theta_2\|_2$ for any $Z \in \mathcal{Z}$ and $\theta \in \Theta_o(o(n^{-1=4}))$.

Theorem 5.1. Under Assumptions 1-16, we have $\hat{V} = V + o_P(1)$:

It is worth noting that, for linear sieves, computing $\hat{w}^k(\tau)$ and hence the covariance estimator does not require nonlinear optimizations. A simple pooled and weighted LS regression of the derivative of ρ with respect to μ_k on the derivatives of ρ with respect to h is all we need. Thus, our covariance estimator is easy to compute.

6. Efficiency

Having established the asymptotic distribution of the parametric component, we now investigate the efficiency of $\hat{\beta}_n$. Clearly, the efficiency depends on the choice of the weighting matrix $\Sigma(X)$. It is straightforward to show that the optimal weighting matrix in the sense of minimizing the asymptotic variance of the estimator $\hat{\beta}_n$ is $\Sigma(X) = \Sigma_o(X) \equiv Var(\psi(Z;\beta_o)|X)$. Thus, when $\Sigma(X)$ is set to $\Sigma_o(X)$, the corresponding estimator is the best in the class of all weighted minimum distance estimators and hence called the optimally weighted minimum distance estimator. The asymptotic covariance of the optimally weighted minimum distance estimator is V_o^{-1}, where

$$V_o = E\left[D_{w_o}(X)'[\Sigma_o(X)]^{-1}D_{w_o}(X)\right] \tag{6.1}$$

$$= \inf_{w(\cdot)} E\left[D_w(X)'[\Sigma_o(X)]^{-1}D_w(X)\right]. \tag{6.2}$$

The question now is whether the optimally weighted sieve minimum distance estimator is also the best in the class of all regular and $\sqrt{n}$ consistent estimators for model (1.3). To address this question, one would normally have to compute the semiparametric efficiency bound of model (1.3) and then compare the semiparametric efficiency bound with the asymptotic covariance matrix of the optimally weighted sieve minimum distance estimator. Unfortunately, in our general setup (1.3), there is no analytical expression of the semiparametric efficiency bound, see Chamberlain (1992) for an example. Fortunately, we do not need the explicit expression of the bound and are still able to present the following efficiency result:

Theorem 6.1. Suppose Assumptions 1-14 hold with $\Sigma = \Sigma_o$. Then the optimally weighted sieve minimum distance estimator for the parametric component in model (1.3) attains the semiparametric efficiency bound V_o^{-1}.

Theorem 6.1 shows that the optimally weighted sieve minimum distance estimator is efficient. It also shows that the V_o^{-1} defined in (6.1) characterizes the semiparametric efficiency bound of model (1.3). While the former result improves upon the existing literature on semiparametric efficient estimation, the latter result extends Chamberlain's (1992) which computes the semiparametric efficiency bound for a special case of (1.3) where $h_o = (h_{o1}(\tau_1(Z)), ..., h_{oq}(\tau_q(Z)))'$, $\tau_j(Z)$, $j = 1, ..., q$ are completely known.

It is known that the generalized sieve minimum distance estimator is e¢cient for a parametric regression model when the variance of the error term is used as the weight. For a more general parametric conditional moment restrictions model (1.1), the GMM estimator is e¢cient when the moment functions are weighted by the covariance matrix of the moment functions (Newey 1990, 1993). For a semiparametric conditional moment restrictions setting (1.3), one would expect the same e¢ciency result to hold when the covariance of the moment functions is used as the weighting matrix. Theorem 6.2 simply con…rms this intuition.

To obtain the e¢cient estimator, we need the weighting matrix $\Sigma_o(X)$, which can be estimated by the kernel estimator $\hat{\Sigma}_o(X)$ if an initial consistent estimator $\hat{\theta}_n$ is available. The following three-step procedure describes how to obtain the e¢cient estimator:

Step 1. Obtain an initial consistent estimator $\hat{\theta}_n$ as the solution to:

$$\min_{\theta \in A_n} \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\} [\rho(X_i;\theta)]' \rho(X_i;\theta).$$

Step 2. Obtain a consistent estimator of the optimal weighting matrix by

$$\hat{\Sigma}_o(X_i) = \frac{[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1}^{n} \psi(Z_j;\hat{\theta}_n)\psi(Z_j;\hat{\theta}_n)' K\left(\frac{X_i - X_j}{a_n}\right)}{\hat{f}_{X_i}}.$$

Step 3. Obtain the optimally weighted estimator $\hat{\theta}_n = (\hat{\mu}_n, \hat{\eta}_n)$ by solving

$$\min_{(\mu,\eta) \in \mathcal{E} \times H_n} \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\} [\rho(X_i;\theta)]' \hat{\Sigma}_o(X_i)^{-1} \rho(X_i;\theta).$$

From Theorem 3.1, the initial estimator in Step 1, $\hat{\theta}_n$, converges to θ_o at a rate faster than $n^{-1/4}$. To apply Theorem 4.1 to $\sqrt{n}(\hat{\mu}_n - \mu_o)$, we need to show the estimated weighting matrix in Step 2 converges in probability to $\Sigma_o(X_i)$ uniformly over $X_i \in \mathcal{X}_n$ at a rate faster than $n^{-1/4}$. This uniform convergence can be proved with the following additional condition.

Assumption 17. $E[\psi(Z;\theta)\psi(Z;\theta)' \mid X]$ has up to J-th derivatives with respect to X and the J-th derivative is uniformly bounded in $X \in \mathcal{X}$ and $\theta \in \mathcal{N}_o(o(n^{-1/4}))$.

Theorem 6.2. Under Assumptions 1, 2, 4 - 12, 14(ii), 16 and 17, we obtain: (i) $\hat{\Sigma}_o(X_i) = \Sigma_o(X_i) + o_p(n^{-1/4})$ uniformly over $X_i \in \mathcal{X}_n$: (ii) If in addition, V_o is positive-de…nite and Assumption 13 hold, then: $\sqrt{n}(\hat{\mu}_n - \mu_o) \Rightarrow N(0, V_o^{-1})$.

It is worth noting that Step 2 can be repeated using $\hat{\theta}_n$ to re-estimate the weighting matrix and Step 3 is then repeated to obtain a new estimator. The new estimator for μ_o is asymptotically as efficient as $\hat{\mu}_n$. Thus, there is no efficiency gain by repeating Step 2 and Step 3. This remains true even if Step 2 and Step 3 are repeated many times. It is possible, however, the new estimator have better finite sample properties.

7. Applications

In this section we provide relative "low-level" sufficient conditions to obtain efficient estimators for the three examples in Section 2.

Example 2.1 Continued: We estimate $\theta_o = (\mu_o; h_{o1}; :::; h_{oq})$ using the simpler sieve generalized least squares regression (2.2). Consider the distance:

$$K(\theta_o; \theta) = \frac{1}{2} E \begin{pmatrix} [X_0'(\mu - \mu_o) + \sum_{j=1}^{q} [h_j(X_j) - h_{oj}(X_j)] g]' \Sigma(X)^{-1} \\ [X_0'(\mu - \mu_o) + \sum_{j=1}^{q} [h_j(X_j) - h_{oj}(X_j)] g] \end{pmatrix}$$
$$= \frac{1}{2} \|\theta - \theta_o\|_2^2 = \frac{1}{2} \|\theta - \theta_o\|^2$$

Suppose V_o given by (2.4) is finite positive definite. Then the three-step procedure in Section 6 leads to $\sqrt{n}$-efficient estimator of μ_o with asymptotic variance V_o^{-1}.

Example 2.2 Continued: we apply the generalized minimum distance (2.1) to estimate $(\mu_o; h_{o1}; :::; h_{oq})$. Consider the distance:

$$K(\theta_o; \theta) = \frac{1}{2} E \begin{pmatrix} [(\mu - \mu_o)' E[Y_2 | X] + \sum_{j=1}^{q} [h_j(X_j) - h_{oj}(X_j)] g]' \Sigma(X)^{-1} \\ [(\mu - \mu_o)' E[Y_2 | X] + \sum_{j=1}^{q} [h_j(X_j) - h_{oj}(X_j)] g] \end{pmatrix}$$
$$= \frac{1}{2} \|\theta - \theta_o\|_2^2 = \frac{1}{2} \|\theta - \theta_o\|^2$$

Suppose V_o given by (2.5) is finite positive definite. Then the three-step procedure in Section 6 leads to $\sqrt{n}$-efficient estimator of μ_o with asymptotic variance V_o^{-1}.

Example 2.3 (Semiparametric sample selection): for $l = 1; 2$
does not depend on the endogenous variable Y, we may apply the simpler sieve generalized least squares regression (2.2) to estimate $(\mu_{o1}; \mu_{o21}; \mu_{o22}; h_{o1}; h_{o2})$ instead of the method of generalized minimum distance (2.1).

Example 2.4 Continued: we apply the method of generalized minimum distance (2.1) to estimate $(\mu_o; h_o)$. Consider the distance:

$$K(\theta_o; \theta) = \frac{1}{2} \|\theta - \theta_o\|_2^2 = \frac{1}{2} \|\theta - \theta_o\|^2$$

$= E\{f X_1^{\prime}(\mu - \mu_o) + E[h(Y_2) - h_o(Y_2)|X]g^{\prime}\Sigma^{-1}\{X_1^{\prime}(\mu - \mu_o) + E[h(Y_2) - h_o(Y_2)|X]g\}$

For identification, we assume that there is no linear combination of X_1 enters additively in $E[h_o(Y_2)|X_1, X_2]$. We also assume that Assumptions 1-5 are satisfied.

Suppose V_o given by (2.6) is finite positive definite. Then the three-step procedure in Section 6 leads to $\sqrt{n}$-efficient estimator of μ_o with asymptotic variance V_o^{-1}.

8. Conclusion

In this paper, we present a consistent estimator with rate for the general semiparametric conditional moments restrictions model (1.3). we derive the $\sqrt{n}$ asymptotic normality of the estimator for the parametric component and provide a consistent estimator of its asymptotic covariance matrix. We show that the optimally weighted sieve minimum distance estimator of the parametric component can attain the semiparametric efficiency bound of model (1.3). We finally provide a three-step procedure to obtain the efficient estimator.

Our results can be extended in several directions. First, model (1.3) assumes all moment restrictions hold conditional on the same set of regressors. This assumption may rule out many interesting applications. For instance, consider a partially linear regression model with endogenous regressors: $Y = X_1\mu_o + h_{o1}(X_2) + u_1$ and $X_2 = h_{o2}(X_3) + u_2$, where u_1 is correlated with u_2, $E[u_1|u_2, X_1, X_3] = E[u_1|u_2]$ and $E[u_2|X_3] = 0$. Let $\psi_1(Z; \mu_o, h_o) = Y - X_1\mu_o - h_{o1}(X_2) - h_{o3}(X_2 - h_{o2}(X_3))$ and $\psi_2(Z; \mu_o, h_o) = X_2 - h_{o2}(X_3)$. Then we have $E[\psi_1(Z; \mu_o, h_o)|X_1, X_2, X_3] = 0$ and $E[\psi_2(Z; \mu_o, h_o)|X_3] = 0$.

Our next example is a modified version of the sample selection model considered by Das, Newey and Vella (1999):

$$Y_2 = Y_1 \cdot [h_{o2}(X_2) + X_3^{\prime}\mu_o + U], Y_1 = 1\{g(X_1) \geq e\}$$

where g and h_{o2} are unknown functions. Let $X = (X_1, X_2, X_3)$ and suppose $E[Y_1|X] = h_{o1}(X_1)$ and $E[U|Y_1 = 1, X] = h_{o3}(h_{o1}(X_1))$ where both h_{o1} and h_{o3} are unknown functions. It is easy to show that equation (1.3) is satisfied with

$$\psi(Z; \theta_o) = (Y_1 - h_{o1}(X_1), Y_2 - h_{o1}(X_1)[h_{o2}(X_2) + X_3^{\prime}\mu_o + h_{o3}(h_{o1}(X_1))])^{\prime},$$

where $h_o(\cdot) = (h_{o1}(\cdot), h_{o2}(\cdot), h_{o3}(\cdot))^{\prime}$. In this example, the argument of h_{o1} is X_1, the argument of h_{o2} is X_2, and the argument of h_{o3} is the unknown function h_{o1}. This example shows that different moment restrictions may hold conditional on

different set of regressors. A related problem is that some conditioning variables have index structure such as those in Ai (1997) and Andrews (1995). Our procedure and results need to be extended to models of this sort. Second, although our results are sufficient for statistical inference on the parametric component, they are not sufficient for statistical inference on the nonparametric component. In some applications, we may be interested in testing the functional forms of $h_o(\cdot)$ (e.g. completely unspecified $h_o(\cdot)$ against the index form or the index form against a parametric specification of $h_o(\cdot)$). We may also be interested in testing the joint restrictions of both the parametric component and the nonparametric component. We are currently undertaking a project to develop more efficient tests for all sorts of restrictions of model (1.3). Third, as we noted before, although our results on convergence rates and $\sqrt{n}$ asymptotic normality can be easily extended to weakly dependent time series data, the problem of semiparametric efficiency bound with time series data is non-trivial. Since most dynamic models of asset pricing fit into the framework (1.3) with stationary Markov data, see e.g. Hansen and Richard (1987), we plan to address the efficiency issue in a separate paper. Lastly but not the least importantly, we need to address problems such as finite sample performance, higher-order analysis and choice of smoothing parameters in subsequence research.

References

[1] Ai, C. (1997): "A Semiparametric Maximum Likelihood Estimator", Econometrica, 65, 933-964.

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

[3] Andrews, D. (1995): "Nonparametric kernel estimation for semiparametric models", Econometric Theory, 11, 560-596.

[4] Blundell, R., M. Browning and I. Crawford (2000): "Nonparametric Engel Curves and Revealed Preference", UCL Working Paper.

[5] Chamberlain, G. (1987): "Asymptotic E¢ciency in Estimation with Conditional Moment Restrictions," Journal of Econometrics, 34, 305-334.

[6] Chamberlain, G. (1992): "E¢ciency Bounds for Semiparametric Regression," Econometrica, 60, 567-596.

[7] Chen, X. and T. Conley (1999): "A New Semiparametric Spatial Model for Panel Time Series." Northwestern Working Paper.

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

[9] Chen, X. and H. White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators", IEEE Tran. Information Theory, 45, 682-691.

[10] Chen, X., L.P. Hansen, and J. Scheinkman (1997): "Shape-preserving Estimation of Di¤usions", manuscript, University of Chicago, Dept. of Economics.

[11] Das, M., W.K. Newey and F. Vella (1999): "Non-parametric Estimation of the Sample Selection Model", manuscript, MIT Dept. of Economics.

[12] DeVore, R.A. and G. Lorentz (1993): Constructive Approximation. New York: Springer-Verlag.

[13] Fan, J., W. Härdle and E. Mammen (1998): "Direct Estimation of Low Dimensional Components in Additive Models," Annals of Statistics 26, 943-971.

[14] Fenton, V. and A.R. Gallant (1996): "Convergence Rate of SNP Density Estimators," Econometrica 64, 719-727.

[15] Grenander, U. (1981): Abstract Inference, New York: Wiley Series.

[16] Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, 50, 1029-1054.

[17] Hansen, L.P. and S.F. Richard (1987): "The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models", Econometrica, 55, 587-613.

[18] Horowitz, J. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," Econometrica 64:1, 103-137.

[19] Horowitz, J. (1998): Semiparametric Methods in Econometrics, Springer-Verlag.

[20] Ichimura, H. and L.F. Lee (1991): "Semiparametric Least Squares Estimation of Multiple Models: Single Equation Estimation," in Nonparametric and Semiparametric Models in Econometrics and Statistics, ed. by W. Barnett, J. Powell, and G. Tauchen. Cambridge: Cambridge University Press.

[21] Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," Journal of Econometrics 58, 71-120.

[22] Linton, O. (2000): "Edgeworth Approximations for Semiparametric Instrumental Variable Estimators and Test Statistics", LSE Working Paper.

[23] Newey, W.K. (1990): "E¢cient Instrumental Variables Estimation of Nonlinear Models," Econometrica, 58, 809-837.

[24] Newey, W.K. (1993): "E¢cient Estimation of Models with Conditional Moment Restrictions, in Handbook of Statistics, Vol. 11, G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., Amsterdam: North-Holland.

[25] Newey, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators", Econometrica 62, 1349-1382.

[26] Newey, W.K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators", Journal of Econometrics 79, 147-168.

[27] Newey, W.K. and J.L Powell (1989): "Nonparametric Instrumental Variables Estimation", MIT Working Paper.

[28] Newey, W.K., J.L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models", Econometrica 67, 565-603.

[29] Pakes, A. and S. Olley (1995): "A Limit Theorem for A Smooth Class of Semiparametric Estimators", Journal of Econometrics 65, 295-332.

[30] Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coe¢cients", Econometrica 57, 1403-1430.

[31] Powell, J. (1994): Estimation of Semiparametric Models, in Handbook of Econometrics, ed. by R. Engle and D. McFadden.

[32] Robinson, P. (1987): "Asymptotically E¢cient Estimation in the Presence of Heteroskedasticity of unknown Form," Econometrica 55, 875-892.

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

[34] Robinson, P. (1991): "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," Econometrica 59, 755-786.

[35] Severini, T. and H.W. Wong (1992): "Pro…le Likelihood and Conditionally Parametric Models," The Annals of Statistics, 20(4), 1768-1802.

[36] Shen, X. (1997): "On Methods of Sieves and Penalization," The Annals of Statistics 25(6), 2555-2591.

[37] Silverman, B. (1986): Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

[38] Van der Vaart, A. and J. Wellner (1996): Weak Convergence and Empirical Processes: with Applications to Statistics. New York: Springer

Mathematical Appendices

This appendix consists of several parts. In Appendix A, we present some uniform convergence results for the kernel estimators. In Appendix B, we prove the consistency of the proposed estimator $\hat{\theta}_n$ and compute the convergence rates. In Appendix C, we prove several Lemmas which are then used to establish the asymptotic distribution of the estimator $\hat{\beta}_n$. In Appendix D, we prove the consistency of the variance estimator $\hat{V}$ and verify the efficiency property of the optimally weighted LS estimator.

Appendix A: Some Uniform Convergence Results

Let $G(Z;\theta)$ denote a generic measurable function of the data $Z \in \mathcal{Z}$ and the parameters $\theta \in \Theta$. Recall that $f_X(x)$ is the joint density function of the s-dimensional random variables X, $K(x)$ the kernel function, and a_n the bandwidth. Define

$$g(x;\theta) = f_X(x)E[G(Z;\theta)|X = x]$$

Denote the kernel estimator of $g(X_i;\theta)$ by

$$\hat{g}(X_i;\theta) = \frac{1}{(n-1)a_n^s} \sum_{j \neq i, j=1}^{n} G(Z_j;\theta(\gamma)) K\left(\frac{X_i - X_j}{a_n}\right).$$

Also, recall that $\mathcal{X}_n$ is the trimmed support of X; Θ_n the sieve approximation of Θ, $\|\gamma\|_2$ the norm on Γ, and $N(\Theta_n;\varepsilon)$ the minimum number of mutually exclusive partitions of $\Theta_n = \bigcup_{k=1}^{N(\Theta_n;\varepsilon)} \Theta_{kn}$ such that for any $\theta_1, \theta_2 \in \Theta_{kn}$, we have $\|\theta_1 - \theta_2\|_2 \leq \varepsilon$.

Lemma A.1: Suppose that the data $\{Z_i; i = 1, 2, ..., n\}$ satisfy Assumption 1, the density of X satisfies Assumption 4, and the kernel function satisfies Assumption 5. In addition, suppose the followings hold:

A.1.1. $E[G(Z;\theta)|X]$ has up to J-th total derivatives with respect to X, and the J-th derivative is uniformly bounded over $X \in \mathcal{X}$ and $\theta \in \Theta_n$;

A.1.2. there exists a measurable function $C_1(Z) : \mathcal{Z} \to [0, \infty)$ such that $E[C_1(Z)^p] < \infty$ for some $p > 2$, and $|G(Z;\theta)| \leq C_1(Z)$ for all $\theta \in \Theta_n$;

A.1.3. there exists a measurable function $C_{2n}(Z) : \mathcal{Z} \to [0, \infty)$ such that $E[C_{2n}(Z)|X] = c_{2n} < \infty$ and $|G(Z;\theta_1) - G(Z;\theta_2)| \leq C_{2n}(Z)\|\theta_1 - \theta_2\|_2$ for any $\theta_1, \theta_2 \in \Theta_n$;

A.1.4. $\{\varepsilon_n\}$ is a sequence of small positive values satisfying

$$\frac{n a_n^s \varepsilon_n^2}{\ln[N(\Theta_n; \frac{\varepsilon_n}{c_{2n}}) \cdot a_n^{-s(s+1)} \varepsilon_n^{-s}] \cdot \max\{1; \varepsilon_n^{1-2/p} a_n^{-2(s+1)/p}\}} \to +\infty \text{ as } n \to \infty.$$

Then uniformly over $(X;\alpha) \in X_n \times A_n$:

$$\hat{g}(X;\alpha) = g(X;\alpha) + o_p(\pm_n) + O_p\left(a_n^J\right):$$

Proof. Let $E[\hat{g}(X_i;\alpha)|X_i]$ denote the conditional expectation given X_i: Write

$$\hat{g}(X_i;\alpha) - g(X_i;\alpha) = \hat{g}(X_i;\alpha) - E[\hat{g}(X_i;\alpha)|X_i] + E[\hat{g}(X_i;\alpha)|X_i] - g(X_i;\alpha):$$

Under Assumptions 1, 5(ii), using the variable transformation $t = \frac{x - X_i}{a_n}$, we have:

$$\begin{aligned} E[\hat{g}(X_i;\alpha)|X_i] &= \int \frac{1}{a^s} E[G(Z;\alpha(\zeta))|X = x] K\left(\frac{X_i - x}{a_n}\right) f_X(x)dx \\ &= \int E[G(Z;\alpha(\zeta))|X = X_i + a_n t] K(t) f_X(X_i + a_n t)dt: \end{aligned}$$

Under Assumptions 4, 5 and Condition A.1.1, Taylor expansion of $E[G(Z;\alpha(\zeta))|X = X_i + a_n t] f_X(X_i + a_n t)$ around X_i to the J-th order yields: $E[\hat{g}(X_i;\alpha)|X_i] - g(X_i;\alpha) = O_p\left(a_n^J\right)$ uniformly over $(X_i;\alpha) \in X_n \times A_n$.

Next, consider the residual $\hat{g}(X_i;\alpha) - E[\hat{g}(X_i;\alpha)|X_i]$. Let c denote a generic constant which may have different values in different expressions. Let $\{M_n\}$ be a sequence of positive values that increase with the sample size at certain rates to be specified later. Define $d_{jn} = 1\{C_1(Z_j) \leq M_n\}$. Then we can write

$$\hat{g}(X_i;\alpha) - E[\hat{g}(X_i;\alpha)|X_i] = (\hat{g}_1(X_i;\alpha) - E[\hat{g}_1(X_i;\alpha)|X_i]) + (\hat{g}_2(X_i;\alpha) - E[\hat{g}_2(X_i;\alpha)|X_i])$$

where

$$\hat{g}_1(X_i;\alpha) \equiv \frac{1}{(n-1)a^s} \sum_{j \neq i; j=1} d_{jn} G(Z_j;\alpha) K\left(\frac{X_i - X_j}{a_n}\right);$$

$$\hat{g}_2(X_i;\alpha) \equiv \frac{1}{(n-1)a^s} \sum_{j \neq i; j=1} [1 - d_{jn}] G(Z_j;\alpha) K\left(\frac{X_i - X_j}{a_n}\right):$$

For any small value $\varepsilon > 0$; and a positive sequence $\{\pm_n\}$ satisfying Condition A.1.4, we have:

$$\begin{aligned} & P\left(\sup_{(X_i;\alpha) \in X_n \times A_n} |\hat{g}(X_i;\alpha) - E[\hat{g}(X_i;\alpha)|X_i]| > \varepsilon \pm_n\right) \\ & P\left(\sup_{(X_i;\alpha) \in X_n \times A_n} |\hat{g}_1(X_i;\alpha) - E[\hat{g}_1(X_i;\alpha)|X_i]| > \varepsilon \pm_n/2\right) \\ & + P\left(\sup_{(X_i;\alpha) \in X_n \times A_n} |\hat{g}_2(X_i;\alpha) - E[\hat{g}_2(X_i;\alpha)|X_i]| > \varepsilon \pm_n/2\right) \\ \equiv\ & P_1 + P_2: \end{aligned}$$

Applying the Chebyshev inequality and under Assumption 5(i) and Condition A.1.2, we obtain

$$P_2 \quad \frac{2E\left[\sup_{(X_i;\theta)\in X_n\times A_n} |g_2(X_i;\theta) - E[g_2(X_i;\theta)|X_i]|\right]}{\left(\pm_n\right)^2}$$

$$\frac{4E\left[[1 - d_{jn}]\sup_{(X_i;\theta)\in X_n\times A_n} G(Z_j;\theta)K\left(\frac{X_i - X_j}{a_n}\right)\frac{1}{a_n^s}\right]}{\left(\pm_n\right)^2}$$

$$\frac{cE\left[[1 - d_{jn}]C_1(Z_j)\right]}{\pm_n a_n^s}$$

$$c\sqrt{\frac{E[1 - d_{jn}]\times E[C_1(Z_j)^2]}{\pm_n a_n^s}} \quad \frac{c}{M_n^{p=2}\pm_n a_n^s} = O(a_n);$$

where the last inequality follows from Condition A.1.2 and

$$E[1 - d_{jn}] = P\left(C_1(Z_j) > M_n\right) \quad \frac{E[C_1(Z_j)^p]}{M_n^p};$$

and the last equality is obtained by letting $M_n = (\pm_n a_n^{s+1})^{-2=p}$. Then $P_2 \to 0$ as $n \to +1$:

Let $W_n = X_n \times A_n$. For any pair $(X_i^1;\theta^1) \in W_n$ and $(X_i^2;\theta^2) \in W_n$, Assumption 5(i), Conditions A.1.2 and A.1.3 imply:

$$\begin{aligned}
&\frac{1}{a^s}d_{jn}G(Z_j;\theta^1)K\left(\frac{X_i^1 - X_j}{a_n}\right) - \frac{1}{a^s}d_{jn}G(Z_j;\theta^2)K\left(\frac{X_i^2 - X_j}{a_n}\right)\\
&\frac{1}{a^s}d_{jn}\left[G(Z_j;\theta^1) - G(Z_j;\theta^2)\right]K\left(\frac{X_i^1 - X_j}{a_n}\right)\\
&+\frac{1}{a^s}d_{jn}G(Z_j;\theta^2)\left[K\left(\frac{X_i^1 - X_j}{a_n}\right) - K\left(\frac{X_i^2 - X_j}{a_n}\right)\right]\\
&\frac{c}{a^s}C_{2n}(Z_j)\left|K\left(\frac{X_i^1 - X_j}{a_n}\right)\right|\left\|\theta^1 - \theta^2\right\|_2 + \frac{c}{a^{s+1}}C_1(Z_j)\left|X_i^1 - X_i^2\right|;
\end{aligned}$$

where $|X_i^1 - X_i^2|$ denotes the Euclidean norm of $X_i^1 - X_i^2$. Since $E[C_1(Z)]$ is …nite, by the Law of Large Number for i.i.d. processes, we have:

$$\frac{1}{(n-1)}\sum_{j\neq i;j=1}^{n} C_1(Z_j) = O_p(1):$$

Since $E[C_{2n}(Z)|X] = c_{2n} < 1$, we have uniformly over $X_i^1 \in X$:

$$\frac{1}{c_{2n}}\frac{1}{(n-1)a_n^s}\sum_{j\neq i;j=1}^{n} C_2(Z_j)\cdot K\left(\frac{X_i^1 - X_j}{a_n}\right) = O_p(1).$$

Thus, for some arbitrarily small value $\varepsilon > 0$, there exists a sufficiently large n such that, with $\bar{X}_i^2 - X_i^1 < a_n^{s+1}\delta_n/(12c)$ and $\|\alpha^2 - \alpha^1\| < \delta_n/(12c_{2n})$, we have with probability at least $1-\varepsilon$:

$$|\psi(X_i^2,\alpha^2) - \psi(X_i^1,\alpha^1)| < \delta_n/6.$$

Partition W_n into b_n mutually exclusive subsets $W_{nm}, m = 1,2,...,b_n$, where $(X_i^1,\alpha^1) \in W_{nm}$ and $(X_i^2,\alpha^2) \in W_{nm}$ imply $\bar{X}_i^2 - X_i^1 < a_n^{s+1}\delta_n/(12c)$ and $\|\alpha^2 - \alpha^1\| < \delta_n/(12c_{2n})$. Since X_n is a compact subset in R^s, we have

$$b_n = O\left((a_n^{s+1}\delta_n)^{-s}N\left(A_n,\frac{\delta_n}{c_{2n}}\right)\right).$$

Let (X_i^m,α^m) denote an arbitrarily fixed point in W_{nm}. We obtain:

$$\begin{aligned} P_1 &< \varepsilon + \sum_{m=1}^{b_n} P\left(|\psi_1(X_i^m,\alpha^m) - E[\psi_1(X_i^m,\alpha^m)|X_i^m]| > \delta_n/6\right) \\ &= \varepsilon + \sum_{m=1}^{b_n} P\left(|\psi_1^*(X_i^m,\alpha^m) - E[\psi_1^*(X_i^m,\alpha^m)|X_i^m]| > a_n^s\delta_n/6\right) \end{aligned}$$

where $\psi_1^*(X_i^m,\alpha^m) = a_n^s\psi_1(X_i^m,\alpha^m) = \frac{1}{(n-1)}\sum_{j\neq i;j=1}^{n} d_{jn}G(Z_j,\alpha^m)K\left(\frac{X_i^m - X_j}{a_n}\right)$; which is bounded by $O(M_n)$. Applying the Bernstein inequality for i.i.d. processes, see e.g. Ichimura and Lee (1991, p.26), we obtain:

$$\begin{aligned} &P\left(|\psi_1^*(X_i^m,\alpha^m) - E[\psi_1^*(X_i^m,\alpha^m)|X_i^m]| > a_n^s\delta_n/6\right) \\ &2\exp\left(-na_n^{2s}\delta_n^2/(c\sigma_m^2 + 8M_n a_n^s\delta_n)\right) \end{aligned}$$

where σ_m^2 is the conditional variance of $d_{jn}G(Z_j,\alpha^m)K\left(\frac{X_i^m - X_j}{a_n}\right)$ given X_i^m. Under Assumptions 1, 4, 5 and Condition A.1.2, it is easy to see that σ_m^2 is $O(a_n^s\int K^2(t)f_X(X_i^m + a_n t)dt) = O(a_n^s)$. Hence,

$$P_1 < \varepsilon + 2b_n\exp\left(-na_n^s\delta_n^2/(c + 8M_n\delta_n)\right)$$

which is arbitrarily small if $\frac{na_n^s\pm_n^2}{\max\{1;M_n\pm_n\}} - \ln(b_n) \to +1$: By the construction of M_n and b_n, we have $b_n \to +1$ and

$$\frac{\frac{na_n^s\pm_n^2}{\ln(b_n)\max\{1;M_n\pm_n\}}}{} = O\left(\frac{na_n^s\pm_n^2}{\ln[N(A_n;\frac{\pm_n}{c_{2n}})\cdot a_n^{-s(s+1)}\pm_n^{-s}]\cdot\max\{1;\pm_n^{1-2/p}a_n^{-2(s+1)/p}\}}\right) \to 1$$

where the limit to in…nity is by Condition A.1.4. This completes the proof. ■

The uniform convergence rate of the kernel estimator $g(X_i;\hat{\theta})$ has two components. The …rst term, $O(a_n^J)$, is the convergence rate of the bias term. This rate depends on the order of the kernel and on the smoothness condition of $g(X;\theta)$. Clearly, higher the order of the kernel function and the degree of smoothness, faster the bias term shrinks to zero. This component is obtained under Assumptions 1, 4, 5 and Condition A.1.1, while Conditions A.1.2, A.1.3 and A.1.4 are not needed. The second component, $O_p(\pm_n)$, is the uniform convergence rate of the residual term. This rate is limited by: (i) p, the number of the moments of $G(Z;\theta)$, (ii) s, the dimension of X, and (iii) $N(A_n;")$; the size of the sieve space. Higher the number of moments, lower the dimension of X; and the smaller the size of the sieve space, faster the residual term converges to zero. This component is obtained under Assumptions 1, 4, 5 and Conditions A.1.2, A.1.3 and A.1.4 without using Condition A.1.1. It must be noted that Lemma A.1 holds only on the sieve parameter space A_n, not on the original parameter space A:

Lemma A.1 still holds when we weaken Conditions A.1.1-A.1.2. For example, if Condition A.1.1 is weakened to assume that the $J-th$ total derivative of $g(X;\theta)$ with respect to X is bounded by B_n over the sieve space where B_n could grow slowly with the sample size, then Lemma A.1 holds with the convergence rate of the bias term replaced by $O(B_n a_n^J)$. Likewise, we can relax Condition A.1.2 by replacing the envelope function $C_1(Z)$ by $C_{1n}(Z)$ where $c_{1n}^p = E[C_{1n}(Z)^p]$ may increase with the sample size. Then Lemma A.1 still holds with the convergence rate of the residual term $o_p(\pm_n)$ if the term

$$\ln[N(A_n;\frac{\pm_n}{c_{2n}})\cdot a_n^{-s(s+1)}\pm_n^{-s}]\cdot\max\{1;\pm_n^{1-2/p}a_n^{-2(s+1)/p}\}$$

in Condition A.1.4 is replaced by

$$\ln[N(A_n;\frac{\pm_n}{c_{2n}})\cdot\frac{c_{1n}^s}{a_n^{s(s+1)}\pm_n^s}]\cdot\max\{1;c_{1n}^{2/p}\pm_n^{1-2/p}a_n^{-2(s+1)/p}\}:$$

Corollary A.1: Under Assumptions 1, 4, 5 and 8, we have: $\hat{f}_{X_i} = f_{X_i} + o_p(n^{-1=4})$ uniformly over $X_i \in \mathcal{X}_n$:

Proof. We apply Lemma A.1 to $G(Z;\theta) = 1$. Conditions A.1.1 - A.1.3 are trivially satisfied with $p = +1$. Because $G(Z;\theta)$ does not depend on the parameters θ, we have $N(\mathcal{A}_n;\pm_n) = 1$. Because $p = +1$, we have $\max\{1;\pm_n^{1-2=p} a_n^{-2(s+1)=p}\} = \max\{1;\pm_n\} = 1$. Hence, with $\pm_n = n^{-1=4}$, Condition A.1.4 is satisfied if

$$\frac{\sqrt{n}a_n^s}{\ln[a_n^{-s(s+1)} n^{s=4}]} \to +1 \text{ as } n \to +1$$

which is implied by Assumption 8. Assumption 8 also implies $n^{1=4} a_n^J \to 0$. ■

Corollary A.2: Under Assumptions 1, 4, 5, 6, 7 and 8, we obtain uniformly over $(X;\theta) \in \mathcal{X}_n \times \mathcal{A}_n$; $\hat{m}(X;\theta) = m(X;\theta) + o_p(n^{-1=4})$:

Proof. We apply Lemma A.1 to $G(Z;\theta) = \psi(Z;\theta)$. Condition A.1.1 is implied by Assumption 6, Conditions A.1.2 and A.1.3 are implied by Assumption 7. Assumption 8 implies Condition A.1.4 with $\pm_n = n^{-1=4}$ and that $n^{1=4} a_n^J \to 0$. The result now follows from Lemma A.1 and Corollary A.1. ■

Define

$$\hat{m}_\theta[X_i;\phi\theta] \equiv \frac{[(n-1)a_n^s]^{-1} \sum_{j\neq i;j=1}^{n} \psi_\theta[Z_j;\phi\theta] K\left[\frac{X_i - X_j}{a_n}\right]}{\hat{f}_{X_i}}.$$

Corollary A.3: Under Assumptions 1, 4, 5, 8 and 11, we obtain uniformly over $X \in \mathcal{X}_n$ and $\grave{A} \in \mathcal{U}_n$: $\hat{m}_{\theta_o}[X;\grave{A}] = m_{\theta_o}[X;\grave{A}] + o_p(n^{-1=4})$:

Proof. We apply Lemma A.1 to $G(Z;\grave{A}) = \psi_{\theta_o}[Z;\grave{A}]$ and $\pm_n = n^{-1=4}$. Assumption 11 implies Conditions A.1.1 - A.1.3 hold in $\mathcal{U}_n$. Restricting the parameter space to $\mathcal{U}_n$ and Noting that $N(\mathcal{U}_n;\pm_n) < N(\mathcal{A}_n;\pm_n)$, Condition A.1.4 is implied by Assumption 8. The result follows from Lemma A.1 and Corollary A.1. ■

For some sequence of positive values $\pm_n \to 0$ as $n \to 1$, Corollary A.3 implies that uniformly over $X \in \mathcal{X}_n$ and $k\theta - \theta_o k_2 \leq \pm_n$,

$$\hat{m}_{\theta_o}[X;\theta - \theta_o] = m_{\theta_o}[X;\theta - \theta_o] + o_p(n^{-1=4}\pm_n) .$$

Recall that $r_\psi[Z;\theta - \theta_o]$ is the remainder term in the linear approximation of $\psi(Z;\theta)$. Define

$$r_m[X;\theta - \theta_o] = E[r_\psi[Z;\theta - \theta_o]|X]$$

and the kernel estimator as

$$\hat{r}_m[X_i;\theta - \theta_o] = \frac{[(n-1)a_n^s]^{-1}\sum_{j\neq i;j=1}^{n} r_\nu[Z_j;\theta - \theta_o]K\left(\frac{X_i - X_j}{a_n}\right)}{\hat{f}_{X_i}}.$$

Corollary A.4: Under Assumptions 1, 4, 5, 6, 8, 11(i) and 12, we obtain: (i) uniformly over $X \in \mathcal{X}_n$ and $\theta \in \Theta_o$:

$$\frac{\hat{r}_m[X;\theta - \theta_o]}{\|\theta - \theta_o\|_2} = \frac{r_m[X;\theta - \theta_o]}{\|\theta - \theta_o\|_2} + o_p(n^{-1/4})$$

(ii) uniformly over $X \in \mathcal{X}_n$, $\theta \in \Theta_n$ with $\|\theta - \theta_o\|_2 \geq \delta_n$: $\hat{r}_m[X;\theta - \theta_o] = O_p(\delta_n^2) + o_p(n^{-1/4}\delta_n)$.

Proof. (i) We apply Lemma A.1 with $G(Z_j;\theta) = \frac{r_\nu[Z_j;\theta - \theta_o]}{\|\theta - \theta_o\|_2}$ and $\delta_n = n^{-1/4}$, Condition A.1.1 is implied by definition of $r_\nu[Z;\theta - \theta_o]$ and Assumptions 6 and 11(i); Conditions A.1.2 and A.1.3 are implied by Assumption 12, Condition A.1.4 is implied by Assumption 8. The result follows from Lemma A.1 and Corollary A.1.

(ii) Immediately follows from Result (i) and Assumption 12(i). ■

Corollary A.4(i) implies that uniformly over $X \in \mathcal{X}_n$ and $\|\theta - \theta_o\|_2 \geq \delta_n$,

$$\hat{r}_m[X;\theta - \theta_o] = r_m[X;\theta - \theta_o] + o_p(n^{-1/4}\delta_n)$$

Appendix B: Convergence Rates for Sieve Estimators

We now prove the consistency and compute the convergence rates of the proposed estimator. Define

$$\hat{L}_n(\theta) \equiv -\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\}\hat{m}(X_i;\theta)'[\hat{S}(X_i)]^{-1}\hat{m}(X_i;\theta)/2$$

$$L_n(\theta) \equiv -\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\}m(X_i;\theta)'[S(X_i)]^{-1}m(X_i;\theta)/2$$

$$\hat{L}_{n\theta_o}[\theta - \theta_o] = -\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\}\hat{m}_{\theta_o}[X_i;\theta - \theta_o]'[\hat{S}(X_i)]^{-1}\hat{m}(X_i;\theta_o),$$

$$L_{n\theta_o}[\theta - \theta_o] = -\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\}m_{\theta_o}[X_i;\theta - \theta_o]'[S(X_i)]^{-1}m(X_i;\theta_o)$$
$$\leq 0 \quad \text{since } m(X_i;\theta_o) \equiv 0 \text{ by definition.}$$

Also de…ne:

$$R[\theta - \theta_o] = L_n(\theta) - L_n(\theta_o) - L_{n\theta_o}[\theta - \theta_o] = L_n(\theta);$$

$$\dot{R}[\theta - \theta_o] = \dot{L}_n(\theta) - \dot{L}_n(\theta_o) - \dot{L}_{n\theta_o}[\theta - \theta_o]:$$

By de…nition of $r_m[X;\theta - \theta_o]$, and $m(X;\theta_o) = 0$; we have:

$$m(X;\theta) = m_{\theta_o}[X;\theta - \theta_o] + r_m[X;\theta - \theta_o]:$$

Write

$$\begin{aligned}
& -2R[\theta - \theta_o] \\
= \; & \frac{1}{n}\sum_{i=1} 1\{X_i \in X_n\} m_{\theta_o}[X_i;\theta - \theta_o]'[\xi(X_i)]^{-1} m_{\theta_o}[X_i;\theta - \theta_o] + \\
& \frac{2}{n}\sum_{i=1} 1\{X_i \in X_n\} m_{\theta_o}[X_i;\theta - \theta_o]'[\xi(X_i)]^{-1} r_m[X_i;\theta - \theta_o] + \\
& \frac{1}{n}\sum_{i=1} 1\{X_i \in X_n\} r_m[X_i;\theta - \theta_o]'[\xi(X_i)]^{-1} r_m[X_i;\theta - \theta_o] \\
\equiv \; & R1 + 2 \cdot R3 + R4;
\end{aligned}$$

$$\begin{aligned}
& -2\dot{R}[\theta - \theta_o] \\
= \; & \frac{1}{n}\sum_{i=1} 1\{X_i \in X_n\} \dot{m}_{\theta_o}[X_i;\theta - \theta_o]'[\xi(X_i)]^{-1} \dot{m}_{\theta_o}[X_i;\theta - \theta_o] + \\
& \frac{2}{n}\sum_{i=1} 1\{X_i \in X_n\} \dot{m}(X_i;\theta_o)'[\xi(X_i)]^{-1} \dot{b}_m[X_i;\theta - \theta_o] + \\
& \frac{2}{n}\sum_{i=1} 1\{X_i \in X_n\} \dot{m}_{\theta_o}[X_i;\theta - \theta_o]'[\xi(X_i)]^{-1} \dot{b}_m[X_i;\theta - \theta_o] + \\
& \frac{1}{n}\sum_{i=1} 1\{X_i \in X_n\} \dot{b}_m[X_i;\theta - \theta_o]'[\xi(X_i)]^{-1} \dot{b}_m[X_i;\theta - \theta_o] \\
\equiv \; & \dot{R}1 + 2 \cdot \dot{R}2 + 2 \cdot \dot{R}3 + \dot{R}4:
\end{aligned}$$

The following corollaries are needed for deriving the asymptotic results of the proposed estimator.

Corollary A.5: (i) Under Assumptions 1, 3-8, we have uniformly over $\theta \in A_n$:

$$\hat{L}_n(\theta) - L_n(\theta) = o_p(n^{-1/4}).$$

(ii) Under Assumptions 1-8, 11, we have uniformly over $\theta \in A_n$, $\|\theta - \theta_0\|_2 \leq \varepsilon_n$:

$$\hat{L}_{n\theta_0}[\theta - \theta_0] = o_p(n^{-1/4}\varepsilon_n).$$

Proof. (i) A direct consequence of Assumptions 1, 3, Corollary A.2 and the following decomposition:

$$\begin{aligned}
&\hat{m}(X_i;\theta)'[\hat{\Sigma}(X_i)]^{-1}\hat{m}(X_i;\theta) - m(X_i;\theta)'[\Sigma(X_i)]^{-1}m(X_i;\theta) \\
= &\{\hat{m}(X_i;\theta) - m(X_i;\theta)\}'[\hat{\Sigma}(X_i)]^{-1}\hat{m}(X_i;\theta) + \\
&+ m(X_i;\theta)'\left[[\hat{\Sigma}(X_i)]^{-1} - [\Sigma(X_i)]^{-1}\right]\hat{m}(X_i;\theta) + \\
&+ m(X_i;\theta)'[\Sigma(X_i)]^{-1}\{\hat{m}(X_i;\theta) - m(X_i;\theta)\}
\end{aligned}$$

(ii) By definition, we have:

$$\begin{aligned}
&\hat{L}_{n\theta_0}[\theta - \theta_0] \\
= &-\|\theta - \theta_0\|_2 \cdot \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in X_n)\hat{m}_{\theta_0}[X_i; \frac{\theta - \theta_0}{\|\theta - \theta_0\|_2}]'[\hat{\Sigma}(X_i)]^{-1}\hat{m}(X_i;\theta_0).
\end{aligned}$$

Now by the decomposition for any $\lambda \in U_n$;

$$\begin{aligned}
&\hat{m}_{\theta_0}[X_i;\lambda]'[\hat{\Sigma}(X_i)]^{-1}\hat{m}(X_i;\theta_0) \\
= &\hat{m}_{\theta_0}[X_i;\lambda]'\left[[\hat{\Sigma}(X_i)]^{-1} - [\Sigma(X_i)]^{-1}\right]\hat{m}(X_i;\theta_0) \\
&+ (\hat{m}_{\theta_0}[X_i;\lambda] - m_{\theta_0}[X_i;\lambda])'[\Sigma(X_i)]^{-1}\hat{m}(X_i;\theta_0) \\
&+ m_{\theta_0}[X_i;\lambda]'[\Sigma(X_i)]^{-1}\hat{m}(X_i;\theta_0)
\end{aligned}$$

and Assumptions 1, 2, 3, Corollaries A.2 and A.3, we have uniformly over $\lambda \in U_n$;

$$\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in X_n)\hat{m}_{\theta_0}[X_i;\lambda]'[\hat{\Sigma}(X_i)]^{-1}\hat{m}(X_i;\theta_0) = o_p(n^{-1/4}).$$

Hence uniformly over $\theta \in A_n$; $\|\theta - \theta_0\|_2 \leq \varepsilon_n$; $\hat{L}_{n\theta_0}[\theta - \theta_0] = \varepsilon_n \cdot o_p(n^{-1/4})$: ■

Corollary A.6: Under Assumptions 1-8, 11 and 12, we have uniformly over $\theta \in A_n, \|\theta - \theta_o\|_2 \leq \delta_n$ for any $\delta_n = o(n^{-1/8})$:

$$\text{(i)} \quad \hat{R}[\theta - \theta_o] = R[\theta - \theta_o] + o_p(n^{-1/2});$$

$$\text{(ii)} \quad \hat{L}_n(\theta) - \hat{L}_n(\theta_o) - \{L_n(\theta) - L_n(\theta_o)\} = o_p(n^{-1/4}\delta_n).$$

Proof. (i) Recall $-2\hat{R}[\theta - \theta_o] \equiv \hat{R}1 + 2\hat{R}2 + 2\hat{R}3 + \hat{R}4$, and $-2R[\theta - \theta_o] \equiv R1 + 2R3 + R4$. Write

$$\begin{aligned}
&\hat{R}1 - R1 \\
&= \frac{1}{n}\sum_{i=1} 1(X_i \in X_n)\{\hat{m}_{\theta_o}[X_i;\theta - \theta_o] - m_{\theta_o}[X_i;\theta - \theta_o]\}'[\hat{\Sigma}(X_i)]^{-1}\hat{m}_{\theta_o}[X_i;\theta - \theta_o] \\
&+ \frac{1}{n}\sum_{i=1} 1(X_i \in X_n) m_{\theta_o}[X_i;\theta - \theta_o]'\{[\hat{\Sigma}(X_i)]^{-1} - [\Sigma(X_i)]^{-1}\}\hat{m}_{\theta_o}[X_i;\theta - \theta_o] \\
&+ \frac{1}{n}\sum_{i=1} 1(X_i \in X_n) m_{\theta_o}[X_i;\theta - \theta_o]'[\Sigma(X_i)]^{-1}\{\hat{m}_{\theta_o}[X_i;\theta - \theta_o] - m_{\theta_o}[X_i;\theta - \theta_o]\}.
\end{aligned}$$

Corollary A.3 and Assumption 3 imply $\hat{R}1 - R1 = o_p(\delta_n^2 n^{-1/4}) = o_p(n^{-1/2})$ uniformly over $\|\theta - \theta_o\|_2 \leq \delta_n$ for any $\delta_n = o(n^{-1/8})$.

Similarly it is easy to check that Corollaries A.2 and A.4 imply $\hat{R}2 = o_p(n^{-1/2})$ uniformly over $\|\theta - \theta_o\|_2 \leq \delta_n$: Corollaries A.3 and A.4(ii) imply $\hat{R}3 - R3 = o_p(n^{-1/2})$ uniformly over $\|\theta - \theta_o\|_2 \leq \delta_n$: Corollary A.4(ii) implies $\hat{R}4 - R4 = o_p(n^{-1/2})$ uniformly over $\|\theta - \theta_o\|_2 \leq \delta_n$:

(ii) A direct consequence of Corollaries A.5(ii) and A.6(i). ■

We now are ready to prove Theorem 3.1.

Proof. (Theorem 3.1): Let $\tau = 1/4$; $\varepsilon_{0n} = o(n^{-\tau})$ and $\delta_{0n} = 2\sqrt{\varepsilon_{0n}} = o(n^{-\tau/2})$: For any fixed $k > 1$,

$$\begin{aligned}
&P(k\hat{\theta}_n - \theta_o\|_2 \geq k\delta_{0n}) \\
&P\left(\sup_{\{\|\theta - \theta_o\|_2 \geq k\delta_{0n}, \theta \in A_n\}} \hat{L}_n(\theta) \geq \hat{L}_n(\theta_{0n})\right) \\
&P\left(\sup_{\{\theta \in A_n\}} |\hat{L}_n(\theta) - L_n(\theta)| > \varepsilon_{0n}\right)
\end{aligned}$$

$$
\begin{aligned}
&+P\left(\sup_{\{\theta\in A_n\}}\left(\bar{\hat{L}}_n(\theta)-L_n(\theta)\right)-\bar{\hat{L}}_n(\theta_o)>\varepsilon_{0n}\right.\\
&\qquad\left.\sup_{\{\|\theta-\theta_o\|_2\le\delta_{0n},\theta\in A_n\}}\left(\hat{L}_n(\theta)-\hat{L}_n(\hat{\theta}_n)\right)\ge\varepsilon_{0n}\right)\\
&\le P\left(\sup_{\{\theta\in A_n\}}\left|\bar{\hat{L}}_n(\theta)-L_n(\theta)\right|>\varepsilon_{0n}\right)\\
&+P\left(\sup_{\{\|\theta-\theta_o\|_2\le\delta_{0n},\theta\in A_n\}}\left(L_n(\theta)-L_n(\hat{\theta}_n)\right)\ge 2\varepsilon_{0n}\right)\\
&\equiv P_1+P_2.
\end{aligned}
$$

Corollary A.5(i) implies $P_1 \to 0$ as $n \to \infty$. To show $P_2 \to 0$, we notice that all conditions A.1-A.4 in Chen and Shen (1998) are trivially satisfied given our Assumptions 1, 2, 3(ii), 7(ii), 9, 10 and our definition of $\|\cdot\|_2$ and Θ_n. Hence $P_2 \to 0$ by the theorem 1 in Chen and Shen (1998). This proves: $\|\hat{\theta}_n-\theta_o\|_2 = O_p(\delta_{0n}) = o_p(n^{-\varphi/2})$.

Next, we refine the convergence rate by exploiting the local curvature of $\hat{L}_n(\theta)$ around θ_o. Let $\varepsilon_{1n} = n^{-1/4}\delta_{0n} = o(n^{-(\varphi+\varphi/2)})$ and $\delta_{1n} = 2\sqrt{\varepsilon_{1n}} = o(n^{-(\varphi/2+\varphi/4)})$. For any fixed $k > 1$, we have:

$$
\begin{aligned}
&P\left(\|\hat{\theta}_n-\theta_o\|_2\le k\delta_{1n}\right)\\
&\le P\left(\sup_{\delta_{0n}<\|\theta-\theta_o\|_2\le k\delta_{1n}}\left|\hat{L}_n(\theta)-\hat{L}_n(\theta_o)\right|\ge\left|\hat{L}_n(\hat{\theta}_n)-\hat{L}_n(\theta_o)\right|\right)\\
&\le P\left(\sup_{\delta_{0n}<\|\theta-\theta_o\|_2\le k\delta_{1n}}\left|\bar{\hat{L}}_n(\theta)-\hat{L}_n(\theta_o)-\{L_n(\theta)-L_n(\theta_o)\}\right|>\varepsilon_{1n}\right)\\
&+P\left(\sup_{\delta_{0n}<\|\theta-\theta_o\|_2\le k\delta_{1n}}\left|\bar{\hat{L}}_n(\theta)-\hat{L}_n(\theta_o)-\{L_n(\theta)-L_n(\theta_o)\}\right|>\left|\hat{L}_n(\hat{\theta}_n)-\hat{L}_n(\theta_o)\right|\right)\\
&\le P\left(\sup_{\delta_{0n}<\|\theta-\theta_o\|_2\le k\delta_{1n}}\left|\bar{\hat{L}}_n(\theta)-\hat{L}_n(\theta_o)-\{L_n(\theta)-L_n(\theta_o)\}\right|>\varepsilon_{1n}\right)\\
&+P\left(\sup_{\delta_{0n}<\|\theta-\theta_o\|_2\le k\delta_{1n}}\left|L_n(\theta)-L_n(\theta_o)\right|-\left|\hat{L}_n(\hat{\theta}_n)-\hat{L}_n(\theta_o)\right|\ge 2\varepsilon_{1n}\right)\\
&\equiv P_3+P_4
\end{aligned}
$$

Corollary A.6(ii) implies $\hat{L}_n(\theta)-\hat{L}_n(\theta_o)-\{L_n(\theta)-L_n(\theta_o)\} = o_p(n^{-1/4}\delta_{0n}) = o_p(\varepsilon_{1n})$, hence $P_3 \to 0$ as $n \to \infty$. Now under Assumptions 1, 2, 3(ii), 7(ii), 9

and 10, the Theorem 1 of Chen and Shen again implies $P_4 \to 0$ as $n \to \infty$. This proves $\|\hat{\theta}_n - \theta_o\|_2 = O_p(\delta_{1n}) = o_p(n^{-(1/2+1/4)})$.

Repeat the above proof j times, we obtain $\gamma_{jn} = o(n^{-(1/2+1/4+\cdots+1/2^j)})$ and $\|\hat{\theta}_n - \theta_o\|_2 = O_p(\delta_{jn})$ with $\delta_{jn} = o(n^{-(1/2+1/2^2+\cdots+1/2^{j+1})})$. Repeat the above proof infinite many times, we obtain $\|\hat{\theta}_n - \theta_o\|_2 = O_p(\delta_{1n})$ with $\delta_{1n} = o(n^{-(1/2+1/2^2+\cdots)}) = o(n^{-1/4})$. This completes the proof. ■

Appendix C: Asymptotic Normality

In the section, we derive the asymptotic distribution of the proposed estimator. We begin by first introducing notation, then proving several Lemmas, and finally proving Theorem 4.1.

Define a pseudo-norm on A_n as:

$$\begin{aligned}\|\theta - \theta_o\|_n^2 &= E\left\{1(X_i \in X_n)m_{\theta_o}[X_i;\theta - \theta_o]'[\Sigma(X_i)]^{-1}m_{\theta_o}[X_i;\theta - \theta_o]\right\} \\ &= (\mu - \mu_o)'E\left\{1(X_i \in X_n)D_w(X)'\Sigma(X)^{-1}D_w(X)\right\}(\mu - \mu_o).\end{aligned}$$

Note that the difference between $\|\cdot\|_n$ and the Fisher-like norm $\|\cdot\|$ defined in section 4 is the trimming function. It is easy to show $\|\theta - \theta_o\|_2 \to 0$ implies $\|\theta - \theta_o\|_n \to 0$. Let $\langle\cdot,\cdot\rangle_n$ denote the inner product induced by $\|\cdot\|_n^2$. For the bounded linear functional $f(\theta) - f(\theta_o) = \lambda'(\mu - \mu_o)$, there is again a Riesz representor under the inner product $\langle\cdot,\cdot\rangle_n$. Let $w_n^*(\cdot) = (w_n^{*1}(\cdot);\cdots;w_n^{*b}(\cdot))$ be the solution to

$$\inf_{w^i(\cdot)\in\overline{W};i=1,\ldots,b} E[1(X \in X_n)D_w(X)'\Sigma(X)^{-1}D_w(X)].$$

Since $1(X \in X_n) \to 1$, it is easy to show that $D_{w_n^*}(X) \to D_{w^*}(X)$ for each $X \in X_n$. Let $v_n^* = (v_{n\mu}^*; v_{nh}^*)$, where

$$\begin{aligned}v_{n\mu}^* &= \left\{E\left[D_{w_n^*}(X)'[\Sigma(X)]^{-1}D_{w_n^*}(X)\right]^{-1}\right\}\lambda, \\ v_{nh}^* &= -w_n^*(\cdot)\left\{E\left[D_{w_n^*}(X)'[\Sigma(X)]^{-1}D_{w_n^*}(X)\right]^{-1}\right\}\lambda.\end{aligned}$$

Then, by the Riesz representation theorem, we have $f(\hat{\theta}_n) - f(\theta_o) = \langle\hat{\theta}_n - \theta_o, v_n^*\rangle_n$. In the following we shall link $\langle\hat{\theta}_n - \theta_o, v_n^*\rangle_n$ to the derivative of the sample criterion function, and then replace v_n^* by v^* and removes the trimming function.

Remark A.1: Under Assumptions 1, 11(ii) and 14(i), we have: (i) uniformly over $X_i \in X_n$: $m_{\theta_o}[X_i; v_n^*] = m_{\theta_o}[X_i; v^*] + o_p(1)$; (ii) $\|v_n^*\|_n \to \|v^*\|$.

Corollary A.7: Under Assumptions 1-8, 11 and 14(i), we obtain:

$$\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\}\, m_{\theta_o}[X_i; v_n^*]'[\hat{\gamma}(X_i)]^{-1} \psi(X_i; \theta_o)$$
$$= \frac{1}{n}\sum_{i=1}^{n} m_{\theta_o}[X_i; v^*]'[\gamma(X_i)]^{-1} \zeta(Z_i; \theta_o) + o_p(1/\sqrt{n}).$$

Proof. Write

$$\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\}\, m_{\theta_o}[X_i; v_n^*]'[\hat{\gamma}(X_i)]^{-1} \psi(X_i; \theta_o) = A + B + C$$

where

$$A \equiv \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\}\, m_{\theta_o}[X_i; v_n^*]' \left[[\hat{\gamma}(X_i)]^{-1} - [\gamma(X_i)]^{-1}\right] \psi(X_i; \theta_o);$$

$$B \equiv \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\} \left(\psi_{\theta_o}[X_i; v_n^*] - m_{\theta_o}[X_i; v_n^*]\right)' [\gamma(X_i)]^{-1} \psi(X_i; \theta_o);$$

$$C \equiv \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\}\, m_{\theta_o}[X_i; v_n^*]'[\gamma(X_i)]^{-1} \psi(X_i; \theta_o).$$

Applying Corollaries A.2 and A.3, and Assumptions 1 and 3, and Remark A.1, we have: $A = o_p(n^{-1/2})$ and $B = o_p(n^{-1/2})$. Denote $g(x; \theta_o) = m(x; \theta_o) f_{X_i}$, we have $g(x; \theta_o) = m(x; \theta_o) f_{X_i} = 0$ by Assumption 2. Then

$$C = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\} m_{\theta_o}[X_i; v^*]'[\gamma(X_i)]^{-1} \frac{g(X_i; \theta_o)}{f_{X_i}}$$
$$= \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\} m_{\theta_o}[X_i; v_n^*]'[\gamma(X_i)]^{-1} \left(\frac{1}{\hat{f}_{X_i}} - \frac{1}{f_{X_i}}\right) g(X_i; \theta_o)$$
$$+ \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\} m_{\theta_o}[X_i; v_n^*]'[\gamma(X_i)]^{-1} \frac{g(X_i; \theta_o)}{f_{X_i}}.$$

Corollaries A.1, A.3, and Assumptions 1, 2 and 4 imply that

$$g(X_i; \theta_o) = o_p(n^{-1/4}), \quad \frac{1}{\hat{f}_{X_i}} - \frac{1}{f_{X_i}} = o_p(n^{-1/4}) \quad \text{uniformly over } X_i \notin X_n.$$

Hence

$$C = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \notin X_n\} m_{\theta_o}[X_i; v_n^*]'[\gamma(X_i)]^{-1} \frac{g(X_i; \theta_o)}{f_{X_i}} + o_p(1/\sqrt{n}).$$

Note that

$$\hat{g}(X_i;\theta_o) = [(n-1)a_n^s]^{-1}\sum_{j\neq i;j=1}^{n}\psi(Z_j;\theta_o(\zeta))K\left(\frac{X_i - X_j}{a_n}\right):$$

By exchanging summation (since $K(\zeta)$ is symmetric), we obtain

$$C = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{(n-1)a_n^s}\sum_{i\neq j;i=1}^{n}\frac{1(X_i \in X_n)m_{\theta_o}[X_i;v_n^{\alpha}]'[\hat{S}(X_i)]^{-1}}{f_{X_i}}K\left(\frac{X_i - X_j}{a_n}\right);\psi(Z_j;\theta_o)\right\}$$
$$+o_p(n^{-1=2})$$

Notice that the term in the $\{\zeta\}$ bracket,

$$\frac{1}{(n-1)a_n^s}\sum_{i\neq j;i=1}^{n}\frac{1(X_i \in X_n)m_{\theta_o}[X_i;v_n^{\alpha}]'[\hat{S}(X_i)]^{-1}}{f_{X_i}}K\left(\frac{X_i - X_j}{a_n}\right);$$

is a kernel estimator of $1(X_j \in X_n)m_{\theta_o}[X_j;v_n^{\alpha}]'[\hat{S}(X_j)]^{-1}$: This kernel estimator is consistent (by Assumptions 1, 3, 4, 5, 11 and Remark A.1). Because $E[\psi(Z_j;\theta_o(\zeta))|X_1;X_2;:::;X_n] = 0$, it follows from applying the Chebyshev inequality that

$$C = \frac{1}{n}\sum_{j=1}^{n}1(X_j \in X_n)m_{\theta_o}[X_j;v_n^{\alpha}]'[\hat{S}(X_j)]^{-1}\psi(Z_j;\theta_o) + o_p(\frac{1}{\sqrt{n}}):$$

Also, by Remark A.1, $1(X_j \in X_n) \to 1$ and $E[\psi(Z_j;\theta_o(\zeta))|X_j] = 0$ and applying the Chebyshev inequality, we obtain:

$$C = \frac{1}{n}\sum_{j=1}^{n}m_{\theta_o}[X_j;v^{\alpha}]'[\hat{S}(X_j)]^{-1}\psi(Z_j;\theta_o) + o_p(\frac{1}{\sqrt{n}}):$$

This completes the proof. ■

In the rest of the proofs, we denote $\varepsilon_n = o(n^{-1=2})$; $\Theta_o(\sqrt{\varepsilon_n}) = \{\theta \in A_n : \|\theta - \theta_o\|_2 \sqrt{\varepsilon_n}\}$, and $u^{\alpha} = \hat{S}v_n^{\alpha}$: De…ne $\theta^{\alpha} = (1-\varepsilon_n)\theta + \varepsilon_n(u^{\alpha} + \theta_o)$:

Remark A.2: (i) For any $\theta \in \Theta_o(\sqrt{\varepsilon_n})$, we have: $\|\theta^{\alpha} - \theta_o\|_2 = O(\sqrt{\varepsilon_n})$; $\|\theta^{\alpha} - \theta\|_2 = O(\varepsilon_n)$; $\|\hat{\Gamma}_n\theta^{\alpha} - \theta\|_2 = O(\varepsilon_n)$, $\|\hat{\Gamma}_n\theta^{\alpha} - \theta_o\|_2 = O(\sqrt{\varepsilon_n})$:
(ii) Under Assumption 9, for any $\theta \in \Theta_o(\sqrt{\varepsilon_n})$ we have: $\|\hat{\Gamma}_n\theta^{\alpha} - \theta^{\alpha}\|_2 = \varepsilon_n O(\sqrt{\varepsilon_n})$

Lemma A.2: Under Assumptions 1 - 9, 11 - 14(i), we obtain uniformly over $\theta \in \Theta_0(\sqrt{\varepsilon_n})$,

$$\hat{R}[\theta - \theta_0] - \hat{R}[\hat{\theta}_n - \theta_0] = R[\theta - \theta_0] - R[\hat{\theta}_n - \theta_0] + o_p(\varepsilon_n^2).$$

Proof. By definition, we can write:

$$\begin{aligned} &\hat{R}[\theta - \theta_0] - \hat{R}[\hat{\theta}_n - \theta_0] - (R[\theta - \theta_0] - R[\hat{\theta}_n - \theta_0]) \\ &= -\frac{1}{2n}\sum_{i=1}^{n} 1\{X_i \in \mathcal{X}_n\}(A_i + B_i + C_i + D_i), \end{aligned}$$

where $A_i \equiv A1_i - A2_i$, with

$$\begin{aligned} A1_i &= \dot{m}_{\theta_0}[X_i;\theta - \hat{\theta}_n]'[\hat{\Sigma}(X_i)]^{-1}\dot{m}_{\theta_0}[X_i;\theta - \hat{\theta}_n] \\ &\quad + 2\dot{m}_{\theta_0}[X_i;\hat{\theta}_n - \theta_0]'[\hat{\Sigma}(X_i)]^{-1}\dot{m}_{\theta_0}[X_i;\theta - \hat{\theta}_n]; \\ A2_i &= m_{\theta_0}[X_i;\theta - \hat{\theta}_n]'[\Sigma(X_i)]^{-1}m_{\theta_0}[X_i;\theta - \hat{\theta}_n]; \\ &\quad + 2m_{\theta_0}[X_i;\hat{\theta}_n - \theta_0]'[\Sigma(X_i)]^{-1}m_{\theta_0}[X_i;\theta - \hat{\theta}_n]; \end{aligned}$$

$$B_i = 2\dot{m}(X_i;\theta_0)'[\hat{\Sigma}(X_i)]^{-1}(\hat{b}_m[X_i;\theta - \theta_0] - \hat{b}_m[X_i;\hat{\theta}_n - \theta_0]);$$

$C_i \equiv C1_i - C2_i$, with

$$\begin{aligned} C1_i &= 2\dot{m}_{\theta_0}[X_i;\theta - \hat{\theta}_n]'[\hat{\Sigma}(X_i)]^{-1}(\hat{b}_m[X_i;\theta - \theta_0] - \hat{b}_m[X_i;\hat{\theta}_n - \theta_0]) \\ &\quad + 2\dot{m}_{\theta_0}[X_i;\theta - \hat{\theta}_n]'[\hat{\Sigma}(X_i)]^{-1}\hat{b}_m[X_i;\hat{\theta}_n - \theta_0] \\ &\quad + 2\dot{m}_{\theta_0}[X_i;\hat{\theta}_n - \theta_0]'[\hat{\Sigma}(X_i)]^{-1}(\hat{b}_m[X_i;\theta - \theta_0] - \hat{b}_m[X_i;\hat{\theta}_n - \theta_0]); \\ C2_i &= 2m_{\theta_0}[X_i;\theta - \hat{\theta}_n]'[\Sigma(X_i)]^{-1}(r_m[X_i;\theta - \theta_0] - r_m[X_i;\hat{\theta}_n - \theta_0]) \\ &\quad + 2m_{\theta_0}[X_i;\theta - \hat{\theta}_n]'[\Sigma(X_i)]^{-1}r_m[X_i;\hat{\theta}_n - \theta_0] \\ &\quad + 2m_{\theta_0}[X_i;\hat{\theta}_n - \theta_0]'[\Sigma(X_i)]^{-1}(r_m[X_i;\theta - \theta_0] - r_m[X_i;\hat{\theta}_n - \theta_0]); \end{aligned}$$

and $D_i \equiv D1_i - D2_i$, with

$$\begin{aligned} D1_i &= (\hat{b}_m[X_i;\theta - \theta_0] - \hat{b}_m[X_i;\hat{\theta}_n - \theta_0])'\hat{\Sigma}_i^{-1}(\hat{b}_m[X_i;\theta - \theta_0] - \hat{b}_m[X_i;\hat{\theta}_n - \theta_0]) \\ &\quad + 2(\hat{b}_m[X_i;\theta - \theta_0] - \hat{b}_m[X_i;\hat{\theta}_n - \theta_0])'\hat{\Sigma}_i^{-1}\hat{b}_m[X_i;\hat{\theta}_n - \theta_0]; \\ D2_i &= (r_m[X_i;\theta - \theta_0] - r_m[X_i;\hat{\theta}_n - \theta_0])'\Sigma_i^{-1}(r_m[X_i;\theta - \theta_0] - r_m[X_i;\hat{\theta}_n - \theta_0]) \\ &\quad + 2(r_m[X_i;\theta - \theta_0] - r_m[X_i;\hat{\theta}_n - \theta_0])'\Sigma_i^{-1}r_m[X_i;\hat{\theta}_n - \theta_0]. \end{aligned}$$

Assumptions 1, 3(ii) and 11(ii) and Remark A.2 imply:

$$\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\} m_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)]'[\dot{s}(X_i)]^{i-1} m_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)] = O_p(\varepsilon_n^2).$$

Corollary A.3 and Assumptions 1, 3(ii) and 11(ii) and Remark A.2 imply:

$$\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\} \dot{m}_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)]'[\dot{s}(X_i)]^{i-1} \dot{m}_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)] = O_p(\varepsilon_n^2).$$

Notice that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\} \dot{m}_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o]'[\dot{s}(X_i)]^{i-1} \dot{m}_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)] \\
&- \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\} m_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o]'[\dot{s}(X_i)]^{i-1} m_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)] \\
= &\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}(A_{1i} + A_{2i} + A_{3i});
\end{aligned}$$

where

$$A_{1i} = (\dot{m}_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o] - m_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o])'[\dot{s}(X_i)]^{i-1}\dot{m}_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)];$$

$$A_{2i} = m_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o]'[\dot{s}(X_i)]^{i-1}(\dot{m}_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)] - m_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)]);$$

$$A_{3i} = m_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o]'[\ddot{s}(X_i)]^{i-1} - [\dot{s}(X_i)]^{i-1} m_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)].$$

Applying Corollary A.3 and Assumptions 1, 3(ii) and 11(ii) and Remark A.2, we obtain by Holder inequality,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\} A_{1i}\right| \\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}(\dot{m}_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o] - m_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o])'\right)^{1/2} \\
&\quad \cdot (\dot{m}_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o] - m_{\theta_o}[X_i;\hat{\theta}_n(\tau) - \theta_o]) \\
&\quad \cdot \left(\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}\dot{m}_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)]'\dot{m}_{\theta_o}[X_i;\theta - \hat{\theta}_n(\tau)]\right)^{1/2} \\
&= \|\hat{\theta}_n(\tau) - \theta_o\| \cdot O_p(\sqrt{\varepsilon_n}) \cdot \|\theta - \hat{\theta}_n(\tau)\| \cdot O_p(1) = O_p(\varepsilon_n^2).
\end{aligned}$$

Using the same argument, we obtain $\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}A_{2i} = O_p(\varepsilon_n^2)$ and $\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}A_{3i} = O_p(\varepsilon_n^2)$. Hence, $\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}A_i = O_p(\varepsilon_n^2)$:

Assumptions 1, 2, 3, 13 and Corollaries A.2, A.4 and Remark A.2 imply:

$$\left|\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}B_i\right| = O_p(\varepsilon_n^2):$$

Assumptions 1, 3, 11, 12, 13 and Corollaries A.3, A.4, and Remark A.2 imply:

$$\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}C_i = O_p(\varepsilon_n^2).$$

Assumptions 1, 3, 12 - 14 and Corollary A.4 and Remark A.2 imply

$$\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in X_n\}D_i = O_p(\varepsilon_n^2\sqrt{\pi_n}).$$

The lemma now follows. ■

For any function $g(Z)$, de…ne $\nu_n(g) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i) - E[g(Z_i)]$:

Corollary A.8: Under Assumptions 1, 3, 8, 11 and 14(i), we have: uniformly over $X_i \in X_n$ and $\lambda \in U_n$,

$$\nu_n(1(X_i \in X_n)m_{\theta_o}[X_i;\lambda]'[\Sigma(X_i)]^{-1}m_{\theta_o}[X_i;u^*]) = o_p(n^{-1/4}):$$

Proof. We can still apply Lemma A.1 with $G(Z;\lambda) = m_{\theta_o}[X_i;\lambda]'[\Sigma(X_i)]^{-1}m_{\theta_o}[X_i;u^*]$ but without the bias term. Notice that Conditions A.1.2-A.1.4 are trivially satis…ed given Assumptions 1, 3, 8, 11 and 14(i), and with $\pm_n = n^{-1/4}$, hence the result. ■

Lemma A.3: Under Assumptions 1 - 9, 11 - 14(i), we have uniformly over $\theta \in \Theta_o(\sqrt{\pi_n})$,

$$\begin{aligned} &R[\theta - \theta_o] - R[\Gamma_n\theta^* - \theta_o] - E\{R[\theta - \theta_o] - R[\Gamma_n\theta^* - \theta_o]\} \\ &- L_n(\theta) - L_n(\Gamma_n\theta^*) - E\{L_n(\theta) - L_n(\Gamma_n\theta^*)\} = O_p(\varepsilon_n^2):\end{aligned}$$

Proof. Using the same notation as that in the proof of Lemma A.2, we have:

$$R[\theta - \theta_o] - R[\Gamma_n\theta^* - \theta_o] = -\frac{1}{2n}\sum_{i=1}^{n} 1(X_i \in X_n)(A2_i + C2_i + D2_i)$$

From the results (for $A2_i; C2_i; D2_i$) in the proof of Lemma A.2, we have:

$$
\begin{aligned}
& - (R[\theta - \theta_o] - R[\hat{\theta}_n\theta^* - \theta_o] - E\{R[\theta - \theta_o] - R[\hat{\theta}_n\theta^* - \theta_o]\}) \\
&= \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in X_n) m_{\theta_o}[X_i; \hat{\theta}_n\theta^* - \theta_o]'[\Sigma(X_i)]^{-1} m_{\theta_o}[X_i; \theta - \hat{\theta}_n\theta^*] + O_p(\varepsilon_n^2) \\
&\quad - E\left[1(X_i \in X_n) m_{\theta_o}[X_i; \hat{\theta}_n\theta^* - \theta_o]'[\Sigma(X_i)]^{-1} m_{\theta_o}[X_i; \theta - \hat{\theta}_n\theta^*]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in X_n) m_{\theta_o}[X_i; \theta^* - \theta_o]'[\Sigma(X_i)]^{-1} m_{\theta_o}[X_i; \theta - \theta^*] + O_p(\varepsilon_n^2) \\
&\quad - E\left[1(X_i \in X_n) m_{\theta_o}[X_i; \theta^* - \theta_o]'[\Sigma(X_i)]^{-1} m_{\theta_o}[X_i; \theta - \theta^*]\right] \\
&= \varepsilon_n \cdot \iota_n(1(X_i \in X_n) m_{\theta_o}[X_i; \theta - \theta_o]'[\Sigma(X_i)]^{-1} m_{\theta_o}[X_i; -u^*]) + O_p(\varepsilon_n^2) \\
&= \varepsilon_n \sqrt{\pi_n} \cdot \iota_n(1(X_i \in X_n) m_{\theta_o}[X_i; \frac{\theta - \theta_o}{k\theta - \theta_o k_2}]'[\Sigma(X_i)]^{-1} m_{\theta_o}[X_i; -u^*]) + O_p(\varepsilon_n^2)
\end{aligned}
$$

where the last three equalities follow from Remark A.2. The result now follows by applying Corollary A.8. ■

Lemma A.4: Under Assumptions 1, 2, 3, 9, 11(ii), 12, 13 and 14, we have uniformly over $\theta \in \Theta_o(\sqrt{\pi_n})$,

$$E[L_n(\theta) - L_n(\hat{\theta}_n\theta^*)] = \frac{1}{2}\left[\|\hat{\theta}_n\theta^* - \theta_o\|_n^2 - \|\theta - \theta_o\|_n^2\right] + O(\varepsilon_n^2).$$

Proof. Since $m(X; \theta_o) = 0$, we can write

$$
\begin{aligned}
-E[L_n(\theta)] &= \frac{1}{2}E\left[1(X_i \in X_n) m_{\theta_o}[X_i; \theta - \theta_o]'[\Sigma(X_i)]^{-1} m_{\theta_o}[X_i; \theta - \theta_o]\right] + \\
&\quad E\left[1(X_i \in X_n) m_{\theta_o}[X_i; \theta - \theta_o]'[\Sigma(X_i)]^{-1} r_m[X_i; \theta - \theta_o]\right] + \\
&\quad \frac{1}{2}E\left[1(X_i \in X_n) r_m[X_i; \theta - \theta_o]'[\Sigma(X_i)]^{-1} r_m[X_i; \theta - \theta_o]\right].
\end{aligned}
$$

Then

$$E[L_n(\theta) - L_n(\hat{\theta}_n\theta^*)] - \frac{1}{2}\left[\|\hat{\theta}_n\theta^* - \theta_o\|_n^2 - \|\theta - \theta_o\|_n^2\right] = A_n + \frac{1}{2}B_n$$

where after some manipulation, we have

$$
\begin{aligned}
A_n &= E\left[1(X_i \in X_n) m_{\theta_o}[X_i; \hat{\theta}_n\theta^* - \theta]'[\Sigma(X_i)]^{-1}\{r_m[X_i; \hat{\theta}_n\theta^* - \theta_o] - r_m[X_i; \theta - \theta_o]\}\right] \\
&\quad + E\left[1(X_i \in X_n) m_{\theta_o}[X_i; \theta - \theta_o]'[\Sigma(X_i)]^{-1}\{r_m[X_i; \hat{\theta}_n\theta^* - \theta_o] - r_m[X_i; \theta - \theta_o]\}\right] \\
&\quad + E\left[1(X_i \in X_n) m_{\theta_o}[X_i; \hat{\theta}_n\theta^* - \theta]'[\Sigma(X_i)]^{-1} r_m[X_i; \theta - \theta_o]\right]
\end{aligned}
$$

and

$$B_n = E\Big(1(X_i \le X_n)\{r_m[X_i;\hat{\gamma}_n\hat{\theta}^* - \theta_0] - r_m[X_i;\theta - \theta_0]\}^0 \times \{\xi(X_i)\}^{i-1}\{r_m[X_i;\hat{\gamma}_n\hat{\theta}^* - \theta_0] - r_m[X_i;\theta - \theta_0]\}\Big)$$
$$+2E\Big[1(X_i \le X_n) r_m[X_i;\theta - \theta_0]^0 \{\xi(X_i)\}^{i-1}\{r_m[X_i;\hat{\gamma}_n\hat{\theta}^* - \theta_0] - r_m[X_i;\theta - \theta_0]\}\Big]$$

Now Assumptions 11(ii), 12 - 14(i) together with Remark A.2 imply:

$$A_n = O(\varepsilon_n^2); \quad B_n = O(\varepsilon_n^2 \sqrt{\pi_n}).$$

The result now follows. ■

Lemma A.5: Under Assumptions 1-9, 11, we have, uniformly over $\theta \in \Theta_0(\sqrt{\pi_n})$

$$\hat{L}_{n\theta_0}[\theta^* - \hat{\gamma}_n\theta^*] = O_p(\varepsilon_n^2) \text{ and } \hat{L}_{n\theta_0}[\theta - \theta_0] = O_p(\varepsilon_n).$$

Proof.

$$\begin{aligned}
& -\hat{L}_{n\theta_0}[\theta^* - \hat{\gamma}_n\theta^*] \\
= & \frac{1}{n}\sum_{i=1}^{n} m_{\theta_0}[X_i;\theta^* - \hat{\gamma}_n\theta^*]^0 \{\xi(X_i)\}^{i-1} m(X_i;\theta_0) 1(X_i \le X_n) \\
= & \|\theta^* - \hat{\gamma}_n\theta^*\|_2 \times \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le X_n) m_{\theta_0}[X_i; \frac{\theta^* - \hat{\gamma}_n\theta^*}{\|\theta^* - \hat{\gamma}_n\theta^*\|_2}]^0 \{\xi(X_i)\}^{i-1} m(X_i;\theta_0) \\
= & O_p(\varepsilon_n \sqrt{\pi_n}) \times O_p(\sqrt{\pi_n}) = O_p(\varepsilon_n^2)
\end{aligned}$$

where the last equality follows from applying Remark A.2 and using the same decomposition as that in the proof of Corollary A.5(ii). Also Corollary A.5(ii) directly implies $\hat{L}_{n\theta_0}[\theta - \theta_0] = O_p(\varepsilon_n)$ uniformly over $\theta \in \Theta_0(\sqrt{\pi_n})$. ■

Proof. (Theorem 4.1) Define $\theta^* = (1 - \varepsilon_n)\theta + \varepsilon_n(u^* + \theta_0)$. Simple calculations give:

$$\hat{L}_n(\theta) = \hat{L}_n(\hat{\gamma}_n\theta^*) + \hat{L}_{n\theta_0}[\theta - \hat{\gamma}_n\theta^*] + \hat{R}[\theta - \theta_0] - \hat{R}[\hat{\gamma}_n\theta^* - \theta_0].$$

By Lemma A.2, we have:

$$\begin{aligned}
\hat{L}_n(\theta) &= \hat{L}_n(\hat{\gamma}_n\theta^*) + \hat{L}_{n\theta_0}[\theta - \hat{\gamma}_n\theta^*] + R[\theta - \theta_0] - R[\hat{\gamma}_n\theta^* - \theta_0] + o_p(\frac{1}{n}) \\
&= \hat{L}_n(\hat{\gamma}_n\theta^*) + \hat{L}_{n\theta_0}[\theta - \hat{\gamma}_n\theta^*] + E\,[L_n(\theta) - L_n(\hat{\gamma}_n\theta^*)] + \\
&\quad + R[\theta - \theta_0] - R[\hat{\gamma}_n\theta^* - \theta_0] - E\{R[\theta - \theta_0] - R[\hat{\gamma}_n\theta^* - \theta_0]\} + o_p(\frac{1}{n})
\end{aligned}$$

where the last equality follows from rearranging terms. Lemmas A.3 and A.4 imply

$$\begin{aligned} \dot{Q}_n(\theta) &= \dot{Q}_n(\hat{\theta}_n) - \frac{1}{2}\Big[\|\theta - \theta_0\|_n^2 - \|\hat{\theta}_n - \theta_0\|_n^2 \Big] \\ &\quad + \dot{Q}_{n\theta_0}[\theta - \hat{\theta}_n] + o_p(\frac{1}{n}). \end{aligned}$$

Write

$$\begin{aligned} \|\hat{\theta}_n - \theta_0\|_n^2 &= \|\hat{\theta}_n - \theta^*\|_n^2 + (1 - \varepsilon_n)^2 \|\theta - \theta_0\|_n^2 + \varepsilon_n^2 \|u^*\|_n^2 \\ &\quad + 2\langle \hat{\theta}_n - \theta^*, (1 - \varepsilon_n)(\theta - \theta_0) \rangle_n \\ &\quad + 2\langle \hat{\theta}_n - \theta^*, \varepsilon_n u^* \rangle_n + 2\langle (1 - \varepsilon_n)(\theta - \theta_0), \varepsilon_n u^* \rangle_n . \end{aligned}$$

By Theorem 3.1 and that $\|\theta - \theta_0\|_n$ is dominated by $\|\theta - \theta_0\|_2$, we have

$$|\langle \hat{\theta}_n - \theta^*, \theta - \theta_0 \rangle_n| \quad \|\hat{\theta}_n - \theta^*\|_n \, \pounds \, \|\theta - \theta_0\|_n = O_p(\varepsilon_n^2)$$

Hence

$$\|\hat{\theta}_n - \theta_0\|_n^2 = O_p(\varepsilon_n^2) + (1 - \varepsilon_n)^2 \|\theta - \theta_0\|_n^2 + 2\varepsilon_n(1 - \varepsilon_n)\langle (\theta - \theta_0), u^* \rangle_n$$

and

$$\begin{aligned} & \dot{Q}_n(\theta) - \dot{Q}_n(\hat{\theta}_n) - \dot{Q}_{n\theta_0}[\theta - \hat{\theta}_n] \\ = & -\frac{1}{2}\Big[1 - (1 - \varepsilon_n)^2 \Big] \|\theta - \theta_0\|_n^2 + \varepsilon_n(1 - \varepsilon_n)\langle \theta - \theta_0, u^* \rangle_n + O_p(\varepsilon_n^2). \end{aligned}$$

By de…nition of $\hat{\theta}$, $\dot{Q}_n(\theta) - \dot{Q}_n(\hat{\theta}_n) \, , \, 0$. And by Theorem 3.1, $\|\theta - \theta_0\|_n^2$ $O_p(\varepsilon_n)$: It follows that

$$0 \quad \dot{Q}_{n\theta_0}[\theta - \hat{\theta}_n] + \varepsilon_n(1 - \varepsilon_n)\langle \theta - \theta_0, u^* \rangle_n + O_p(\varepsilon_n^2).$$

Note that $\theta - \hat{\theta}_n = -\varepsilon_n(u^* + \theta_0 - \theta) + \theta^* - \hat{\theta}_n$. Write

$$\dot{Q}_{n\theta_0}[\theta - \hat{\theta}_n] = -\varepsilon_n \dot{Q}_{n\theta_0}[u^*] + \varepsilon_n \dot{Q}_{n\theta_0}[\theta - \theta_0] + \dot{Q}_{n\theta_0}[\theta^* - \hat{\theta}_n].$$

Applying Lemma A.5, we obtain

$$\dot{Q}_{n\theta_0}[\theta - \hat{\theta}_n] = -\varepsilon_n \dot{Q}_{n\theta_0}[u^*] + o_p(\frac{1}{n})$$

and

$$0 \quad \geq \varepsilon_n \hat{Q}_{n\theta_o}[u^*] + \varepsilon_n(1 - \varepsilon_n)\langle \hat{\theta} - \theta_o; u^* \rangle_n + O_p(\varepsilon_n^2):$$

Therefore,

$$0 \geq \hat{Q}_{n\theta_o}[u^*] - (1 - \varepsilon_n)\langle \hat{\theta} - \theta_o; u^* \rangle_n + O_p(\varepsilon_n):$$

Since this holds for $u^* = \pm v_n^*$, we obtain:

$$\left| \hat{Q}_{n\theta_o}[v_n^*] - (1 - \varepsilon_n)\langle \hat{\theta} - \theta_o; v_n^* \rangle_n \right| = O_p(\varepsilon_n):$$

This proves that

$$\hat{Q}_{n\theta_o}[v_n^*] = \langle \hat{\theta} - \theta_o; v_n^* \rangle_n + o_p(\frac{1}{\sqrt{n}}):$$

Applying Corollary A.7, we have:

$$\hat{Q}_{n\theta_o}[v_n^*] = -\frac{1}{n}\sum_{i=1}^{n} m_{\theta_o}[X_i; v^*]'[\Sigma(X_i)]^{-1}\psi(Z_i; \theta_o) + o_p(\frac{1}{\sqrt{n}})$$

Hence for any fixed non-zero $\gamma \in R^b$,

$$\gamma'(\hat{\mu}_n - \mu_o) = -\frac{1}{n}\sum_{i=1}^{n} m_{\theta_o}[X_i; v^*]'\Sigma(X_i)^{-1}\psi(Z_i; \theta_o) + o_p(\frac{1}{\sqrt{n}}):$$

Substituting for v^*, we obtain

$$\begin{aligned} &\sqrt{n}\gamma'(\hat{\mu}_n - \mu_o) \\ &= -\left[E\left[D_{w^*}(X)'(\Sigma(X))^{-1}D_{w^*}(X)\right]^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{w^*}(X_i)'(\Sigma(X_i))^{-1}\psi(Z_i; \theta_o) + o_p(1): \end{aligned}$$

The theorem now follows from applying a standard CLT for i.i.d. data. ■

Appendix D: Asymptotic Variance and Efficiency

In this appendix, we first prove the consistency of $\hat{V}$; then show the efficiency of the optimally weighted minimum distance estimator, and finally establish the asymptotic properties of the three-step estimator.

Proof. (Theorem 5.1): To prove the consistency of the estimated covariance matrix, it suffices to show that $\hat{\Sigma}_o(X_i)$ and $\hat{D}_{w^k}(X_i)$ are consistent estimators of $\Sigma_o(X_i)$ and $D_{w^k}(X_i)$ uniformly over $X_i \in \mathcal{X}_n$. Define

$$\hat{\Sigma}_\theta(X_i) \equiv \frac{[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1}^{n} \psi(Z_j;\theta)\psi(Z_j;\theta)' K\left(\frac{X_i - X_j}{a_n}\right)}{f_{X_i}}.$$

We apply Lemma A.1 with $G(Z_j;\theta) = \psi(Z_j;\theta)\psi(Z_j;\theta)'$ and $\pm_n = 1$, Conditions A.1.1-A.1.4 are satisfied by Assumptions 6, 8 and 16. Hence by Lemma A.1 and Corollary A.1 we have uniformly over $X_i \in \mathcal{X}_n$ and $\theta \in \Theta_n$:

$$\hat{\Sigma}_\theta(X_i) = \Sigma_\theta(X_i) + o_p(1).$$

This together with Assumptions 1, 5, 7, Corollary A.1 and $\|\hat{\theta}_n - \theta_o\|_2 = o_p(n^{-1/4})$ implies that uniformly over $X_i \in \mathcal{X}_n$:

$$\hat{\Sigma}_o(X_i) = \Sigma_o(X_i) + o_p(1).$$

To prove the consistency of $\hat{D}_{w^k}(X_i)$, note that Assumption 15(i) implies uniformly over $X_i \in \mathcal{X}_n$:

$$\begin{aligned}
&\Big[[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1} \left(\frac{\partial \psi(Z_j;\hat{\theta}_n)}{\partial \mu_k} - \frac{\partial \psi(Z_j;\theta_o)}{\partial \mu_k}\right) K\left(\frac{X_i - X_j}{a_n}\right)\Big] \\
&\|\hat{\theta}_n - \theta_o\|_2 + [(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1} \dot{\psi}_\theta(Z_j) K\left(\frac{X_i - X_j}{a_n}\right) = o_p(n^{-1/4})
\end{aligned}$$

where the last equality follows from $\|\hat{\theta}_n - \theta_o\|_2 = o_p(n^{-1/4})$. Applying Lemma A.1 with $G(Z_j;\theta) = \frac{\partial \psi(Z_j;\theta)}{\partial \mu_k}$ and $\pm_n = 1$, we have uniformly over $X_i \in \mathcal{X}_n$:

$$[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1} \frac{\partial \psi(Z_j;\theta_o)}{\partial \mu_k} K\left(\frac{X_i - X_j}{a_n}\right) = E\left[\frac{\partial \psi(Z;\theta_o)}{\partial \mu_k} | X = X_i\right] + o_p(1).$$

Hence

$$[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1} \frac{\partial \psi(Z_j;\hat{\theta}_n)}{\partial \mu_k} K\left(\frac{X_i - X_j}{a_n}\right) = E\left[\frac{\partial \psi(Z;\theta_o)}{\partial \mu_k} | X = X_i\right] + o_p(1).$$

Since the objective function

$$\sum_{i=1} 1(X_i \in \mathcal{X}_n) \hat{D}_{w^k}(X_i)' [\hat{\Sigma}(X_i)]^{-1} \hat{D}_{w^k}(X_i)$$

is globally convex in $w^k(:)$, the solution $\hat{w}^k(:)$ is bounded by $\|\hat{w}^k(:)\|_2 \leq C$ for some constant C. Thus, we only need to concern with the subset $\{\lambda \in \overline{W}_n : \|\lambda\|_2 \leq C\}$. Assumption 15(ii) implies uniformly over $X_i \in \mathcal{X}_n$, $\lambda \in \overline{W}_n$ and $\|\lambda\|_2 \leq C$:

$$[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1}^{n} [\psi_{h_n}[Z_j; \lambda(\cdot)] - \psi_{h_o}[Z_j; \lambda(\cdot)]] K\left(\frac{X_i - X_j}{a_n}\right) = o_p(n^{-1/4}).$$

Moreover, Corollary A.3 implies uniformly over $X_i \in \mathcal{X}_n$, $\lambda \in \overline{W}_n$ and $\|\lambda\|_2 \leq C$:

$$[(n-1)a_n^s]^{-1} \sum_{j \neq i; j=1}^{n} \psi_{h_o}[Z_j; \lambda(\cdot)] K\left(\frac{X_i - X_j}{a_n}\right) = E[\psi_{h_o}[Z; \lambda(\cdot)] | X = X_i] + o_p(1).$$

these results imply $\hat{D}_{w^k}(X_i) = D_{w^k}(X_i) + o_p(1)$ uniformly over $X_i \in \mathcal{X}_n$. Employing the same proof as of Theorem 3.1, we show that $\|\hat{w}^k(:) - w^{*k}(:)\|_2 = o_p(1)$. Hence, $\hat{D}_{\hat{w}^k}(X_i) = D_{w^{*k}}(X_i) + o_p(1)$ uniformly over $X_i \in \mathcal{X}_n$. The theorem now follows immediately. ■

Proof. (Theorem 6.1): Since the semiparametric efficiency bound is the lower bound of the variances of all $\sqrt{n}$ regular (consistent) semiparametric estimators and V_o^{-1} is the asymptotic variance of a semiparametric estimator, it follows that V_o^{-1} must be no smaller than the efficiency bound in positive semi-definite matrix sense. To show that it also is the semiparametric efficiency bound, we only need to show that there is a sequence of parametric submodels whose asymptotic covariance matrices converge to V_o^{-1}: Recall that $w^*(\cdot) \in \overline{W}$ and $-\psi_{w^*}(\cdot)(\mu - \mu_o) \in \overline{H} - h_o$. Because $\overline{W}$ is a completion of W under the Hilbert norm, there is a sequence $w_s(\cdot) \in W$ for $s = 1, 2, ...$ such that $w_s(\cdot) \to w^*(\cdot)$ as $s \to +\infty$. Correspondingly, we have $-\psi_{w_s}(\cdot)(\mu - \mu_o) \in H - h_o$. Hence, $g_s(\cdot; \mu) = h_o - \psi_{w_s}(\cdot)(\mu - \mu_o) \in H$ is a parametric submodel passing through h_o. For this particular submodel, Chamberlain (1987) showed that the semiparametric efficiency bound is the inverse of $V_s = E[D_{w_s}(X)'[\Sigma_o(X)]^{-1} D_{w_s}(X)]$, where $D_w(X)$ is given in (4.1). Letting $s \to +\infty$ and noting that $w_s \to w^*(\cdot)$ and that $D_w(X)$ is a continuous functional of $w(\cdot)$, it follows that $V_s \to V_o$. This proves that V_o^{-1} is the semiparametric efficiency bound for model (1.3). ■

Proof. (Theorem 6.2): (i) We apply Lemma A.1 with $G(Z_j; \theta) = \psi(Z_j; \theta)\psi(Z_j; \theta)'$ and $\varepsilon_n = n^{-1/4}$, Conditions A.1.1-A.1.4 are satisfied by Assumptions 8, 16 and 17. Hence by Lemma A.1 and Corollary A.1 we have uniformly over $X_i \in \mathcal{X}_n$ and

$\alpha \in A_n$:

$$\hat{\Sigma}_\alpha(X_i) = \Sigma_\alpha(X_i) + o_p(n^{-1/4}).$$

Notice that the estimator $\hat{\alpha}_n$ in Step 1 has the convergence rate $\|\hat{\alpha}_n - \alpha_o\|_2 = o_p(n^{-1/4})$ under Assumptions 1, 2, 4 -12. These together with Assumptions 1, 14(ii) and 16(ii) imply that uniformly over $X_i \in X_n$:

$$\hat{\Sigma}_{\alpha_o}(X_i) = \Sigma_{\alpha_o}(X_i) + o_p(n^{-1/4}).$$

(ii) A direct consequence of result (i) and Theorem 4.1. ■