A Nonparametric Maximum Likelihood Estimation of Conditional Moment Restriction Models

Chunrong Ai
Department of Economics, University of Florida,
Gainesville, FL 32611, USA

March, 2004 (Preliminary draft)

Abstract

This paper studies estimation of a conditional moment restriction model using the nonparametric maximum likelihood approach proposed by Gallant and Nychka (1987). Under some sufficient conditions, we show that the estimator of some finite dimensional parameters is asymptotically normally distributed and attains the semiparametric efficiency bound and that the estimator of the density function is consistent. The asymptotic distribution of smooth functionals of the estimated density is also derived. An easy to compute covariance estimator is presented.

1 Introduction

The moment restriction model is often the model of choice for analyzing economic data. And the Generalized Method of Moment estimation (hereafter GMM) proposed by Hansen (1982) is often the method for estimating the moment restriction model. Under some sufficient conditions, Hansen (1982) showed that the GMM estimator is asymptotically normally distributed and that the optimally weighted GMM estimator is efficient for the unconditional moment restriction model. Newey (1990) extended Hansen's work to the conditional moment restriction model by showing that some weighted version of the GMM estimator attains the semiparametric efficiency bound of Chamberlain (1992). Despite of its popularity and desirable large sample properties, it has been documented that the optimally weighted GMM estimator for the unconditional moment restriction model has poor finite sample performance (see Altonji and Segal (1996)). Although no formal arguments have been made, it is widely expected that Newey's weighted version of GMM estimator for the conditional case could also have poor finite sample performance. Thus, it is imperative to find alternatives that are asymptotically as good as the GMM estimators

but have better finite sample performance. Recently, the empirical likelihood estimation has been suggested as such an alternative; see Owen (1990, 1991), Kitamura and Stutzer (1997), Qin and Lawless (1994), Imbens (1997), Imbens, Spady, and Johnson (1998), and Newey and Smith (2004) for the unconditional moment restriction model, and Kitamura, Tripathi and Ahn (2004) and Donald, Imbens, and Newey (2004) for the conditional moment restriction model. Newey and Smith (2004) showed that, for the unconditional moment restrictions model, the empirical likelihood estimator indeed has better second order properties than the GMM estimator. It remains to be seen whether the same result holds for the conditional moment restriction model.

Another alternative to the GMM estimation is the nonparametric maximum likelihood (hereafter ML) estimation proposed by Gallant and Nychka (1987) and Gallant and Tauchen (1989). Surprisingly, this alternative has received little attention from the literature. One possible explanation for the lack of attention is that the large sample properties of the nonparametric ML estimator have not been completely established. Gallant and Nychka (1987) and Fengton and Gallant (1996) only proved consistency of the estimator. The asymptotic distribution of the estimator of the finite dimensional parameters has not been derived. The main objective of this paper is to establish the large sample properties of the nonparametric ML estimator.

There are some reasons to believe that the nonparametric ML estimation might be better than the empirical likelihood estimation on the higher order terms. To see them, consider the following conditional moment restriction model

$$E\{\rho(Y, X, \theta_o)|X\} = 0 \tag{1}$$

where $Z = (Y, X) \in \mathcal{Y} \times \mathcal{X} = \mathcal{Z}$ denotes data and ρ is a vector of functions known up to a finite dimensional unknown parameters θ_o. Throughout the paper, we will always use capital letters to denote random variables and lowercase letters to denote their realizations. Let $f_o(y|x)$ denote the true conditional density of Y given X, and let $f(y|x)$ denote any density function that satisfies the moment restriction for arbitrary θ:

$$\int \rho(y, x, \theta) f(y|x) dy = 0, \ f(y|x) > 0, \int f(y|x) dy = 1. \tag{2}$$

The nonparametric ML estimation chooses the unknown density and the model parameters jointly to maximize the log likelihood function subject to the above restriction:

$$\max_{f(.) \in \mathcal{F}, \theta \in \Theta} E\{\ln(f(Y|X))\} \text{ subject to} \tag{3}$$
$$\int \rho(y, X, \theta) f(y|X) dy = 0, \ f(y|X) > 0, \int f(y|X) dy = 1,$$

where Θ denotes the parameter space of θ and $\mathcal{F}$ denotes the space of the unknown density that contain the true value. Clearly, the model is identified if and

only if the true value $(\theta_o, f_o(y|x))$ is the unique solution to the above optimization problem. The nonparametric maximum likelihood estimation proposed in Gallant and Nychka (1987) and Gallant and Tauchen (1989) is to use a sieve to approximate the unknown density function and then to estimate the parameters by maximizing the sample version of (3).

Now, consider the following less restrictive problem:

$$\max_{f(.)\in\mathcal{F},\theta\in\Theta} E\{\ln(f(Y|X))\} \text{ subject to} \qquad (4)$$
$$\int \rho(y,X,\theta)f(y|X)dy = 0, \ \int f(y|X)dy = 1.$$

Problem (4) is the same as problem (3) except that the positive restriction is dropped. Let $\lambda(x)$ denote the Lagrange Multiplier associated with the moment restriction and let $\mu(x)$ denote the multiplier associated with the density restriction. Then, the Lagrangian for problem (4) is

$$\begin{aligned} & L(\theta, f, \lambda, \mu) \\ = \ & E\left\{\int \left(\ln(f(y|X))f_o(y|X) - \lambda(X)'\rho(y,X,\theta)f(y|X) - \mu(X)f(y|X)\right)dy\right\} \end{aligned}$$

where the expectation is taken with respect to the true density of X. The true value $(\theta_o, f_o(y|x), \lambda_o(x), \mu_o(x))$ solves

$$(\theta_o, f_o(y|x), \lambda_o(x), \mu_o(x)) = \arg\min_{\lambda(.),\mu(.)} \max_{f(.)\in\mathcal{F},\theta\in\Theta} L(\theta, f, \lambda, \mu).$$

For arbitrary θ, $\lambda(.)$ and $\mu(.)$, let $f(y|x,\theta,\lambda)$ denote the solution to:

$$f(y|x,\theta,\lambda) = \arg\max_{f(.)\in\mathcal{F}} L(\theta, f, \lambda, \mu)$$

Applying calculus of variation, we obtain

$$f(y|x,\theta,\lambda) = \frac{f_o(y|x)}{\mu(x) + \lambda(x)'\rho(y,x,\theta)}.$$

Using the constraint $\int f(y|x)dy = 1$ and $\int \rho(y,x,\theta)f(y|x)dy = 0$, we obtain $\mu(x) = 1$. Hence,

$$f(y|x,\theta,\lambda) = \frac{f_o(y|x)}{1 + \lambda(x)'\rho(y,x,\theta)}. \qquad (5)$$

Substituting the solution back into the log likelihood function we obtain

$$E\left\{\ln(f(Y|X,\theta,\lambda))\right\} = E\left\{\ln\left(\frac{1}{1+\lambda(X)'\rho(Y,X,\theta)}\right)\right\} - E\left\{\ln\left(f_o(Y|X)\right)\right\},$$

which is exactly the criterion function used in the empirical likelihood estimation. Hence, the empirical likelihood estimation can be interpreted as the profile Lagrangian approach, where the unknown density function is concentrated out.

There are at least four differences between the nonparametric ML and the empirical likelihood estimation. First and the most obvious difference is that problem (4) does not impose the restriction $f(y|x) > 0$. As a result, the solution in (5), $f(y|x, \theta, \lambda)$, is not guaranteed to be positive everywhere. For example, when $\rho(y, x, \theta)$ is unbounded, ranging from $-\infty$ to $+\infty$, $f(y|x, \theta, \lambda)$ takes negative values for some large y unless $\lambda(x) = 0$. But if we set $\lambda(x) = 0$, the parameter θ disappears from the criterion function and consequently cannot be estimated by the empirical likelihood approach. Second difference is that, for arbitrary θ and $\lambda(.)$, $f(y|x, \theta, \lambda)$ does not necessarily satisfy the density restriction $\int f(y|x, \theta, \lambda)dy = 1$ and the moment restriction $\int \rho(y, x, \theta)f(y|x, \theta, \lambda)dy = 0$ except for when $\lambda(x) = \lambda(x, \theta)$, where

$$\lambda(x, \theta) = \arg\min_{\lambda(.)} E\left\{\ln\left(\frac{1}{1 + \lambda(X)'\rho(Y, X, \theta)}\right)\right\}.$$

Hence, $f(y|x, \theta, \lambda)$ is not necessarily a density function and does not necessarily satisfy the moment restriction. In contrast, the nonparametric estimation always imposes the density and moment restrictions. One would expect that imposition of the density and moment restrictions should help (at least not hurt) the finite sample performance of the nonparametric ML estimator. Third difference is that the empirical likelihood estimation does not estimate the density function, while the nonparametric ML estimation estimates the density function directly. The density estimator allows us to compute other interesting estimands such as quantiles. Fourth and the last difference is that, in some applications, the empirical likelihood function is not differentiable while the nonparametric maximum likelihood function is differentiable. To see this difference, consider a simple example of quantile regression with $x = 1$: $\rho(y, x, \theta) = 1\{y < \theta\} - 1\{y > \theta\}$. Obviously, $\rho(y, x, \theta)$ is not differentiable at $\theta = y$. Hence, the empirical likelihood function $\ln\left(\frac{1}{1+\lambda(x)'\rho(y,x,\theta)}\right)$ is not differentiable at $\theta = y$. On the other hand, the nonparametric ML estimation smooths the moment function by integration:

$$\int \rho(y, x, \theta)dy = \int^{\theta} f(y)dy - \int_{\theta} f(y)dy = 0$$

where $f(y)$ denotes the density function of Y. Obviously the left hand side of the moment restriction is differentiable everywhere. The differentiability is a desirable property that should help both estimation and the finite sample performance.

Despite of those potential advantages, there are at least two potential criticisms of the nonparametric ML estimation. First is that the analytical expressions for the integrations $\int \rho(y, x, \theta)f(y|x)dy$ and $\int f(y|x)dy$ in most applications do not exist. Although in some applications these integrations can be computed with numerical methods, they generally require high dimensional integration that is beyond the capacity of the current computing technology. This criticism will be addressed by replacing the numerical integration with a simulated integration. The simulated integration will undoubted have an effect

on the second order properties of the nonparametric ML estimator. But the effect can be kept small and negligible by using a large number of simulation draws. The second criticism is that, since the unknown density function will be approximated by sieve, the approximation error may have effect on the nonparametric ML estimator of θ. Although the sufficient conditions we present below make sure that the approximation error will not affect the first order properties of the proposed estimator, the approximation error may affect its second order properties. Investigation of how the second order properties are affected, however, is beyond the scope of this paper and will be pursued in a separate paper.

The outline of the paper is as follows: Section 2 formally introduces the nonparametric ML estimation method; Section 3 proves consistency of the estimator; Section 4 derives the asymptotic distribution of the estimator of θ; Section 5 provides a consistent covariance estimator for the θ estimator; and Section 6 concludes. Technical derivations are relegated to an Appendix.

2 Nonparametric MLE

Throughout the paper, we assume that $\{(y_i, x_i), i = 1, 2, ..., n\}$ is a sample of observations on $Z = (Y, X)$, drawn from the joint density $f_o(y|x)f_o(x)$, where $f_o(x)$ is the marginal density of X. The joint density is unknown but satisfies

the moment restriction (1) for some true value θ_o. Our primary interest is the estimation of $(\theta_o, f_o(y|x))$ through empirically implementation of (**??**).

There are two difficulties with implementing (**??**). First is that the density is infinite dimensional and is impossible to estimate from finite data points. Second is that the density and moment restrictions on the infinite dimensional parameter (i.e. density function) are highly nonlinear and difficult to impose. To overcome these difficulties, Gallant and Nychka (1987) proposed a series expansion of the unknown density. To describe their approach, let $g(y|x)$ denote some known conditional density function with unbounded support and let $q(u)$ denote some known and positive transformation function that is monotone over $[0, +\infty)$. The density function $g(y|x)$ is practitioner's initial guess of the true density function. It also plays the role of weighting function, ensuring that integration of power functions over unbounded support exists. Obviously, $g(y|x)$ should be chosen as close to the true density as possible. Since the true density is unknown, this may not be possible. So, at least, one should choose $g(y|x)$ such that $\frac{f_o(y|x)}{g(y|x)}$ is bounded. The function $g(y|x)$ may also depend on some other parameters. Gallant and Nychka (1987), for example, choose $g(y|x)$ to be normal density function with unknown mean and variance. Adding additional parameters to $g(y|x)$ only complicates notation with no additional insight. So, to simplify exposition, we assume that $g(y|x)$ is known. The transformation function q is introduced to ensure that the density function is positive every-

where. In addition, q is chosen so that $q^{-1}\left(\frac{f_o(y|x)}{g(y|x)}\right)$ has a series expansion:

$$q^{-1}\left(\frac{f_o(y|x)}{g(y|x)}\right) = p(y)'\pi_o(x)$$

where $p(y) = (p_1(y), p_2(y), ...)'$ denotes the known series basis functions and $\pi_o(x) = (\pi_{o1}(x), \pi_{o2}(x), ...)'$ denotes the expansion coefficients which are obviously functions of x. The true conditional density is now expressed as

$$f_o(y|x) = q(p(y)'\pi_o(x))g(y|x).$$

Common choices of $g(y|x)$ include any probability density function with support $(-\infty, +\infty)$, while possible choices of q include the power function $q(u) = u^2 + c_n$ with c_n a known and small constant possibly depending on the sample size, the exponential function $q(u) = \exp(u)$, and any other positive function that is invertible over $[0, +\infty)$. Common choices for the series basis functions include power functions, wavelets, and B-splines.

Decompose $p(y) = (p^1(y)', p^2(y)')'$ and $\pi_o(x) = (\pi_o^1(x), \pi_o^2(x))'$. For arbitrary coefficients $\pi(x) = (\pi_1(x), \pi_2(x), ...)'$, decompose $\pi(x)$ accordingly and write

$$p(y)'\pi(x) = p^1(y)'\pi^1(x) + p^2(y)'\pi^2(x) = p^1(y)'\pi^1(x) + h(y, x).$$

Hence, the true values of $\pi^1(x)$ and $\pi^2(x)$ are $\pi_o^1(x)$ and $\pi_o^2(x)$ respectively and the true value of $h(y, x)$ is $h_o(y, x) = p^2(y)'\pi_o^2(x)$. Write

$$f(y|x) = q\left(p^1(y)'\pi^1(x) + h(y, x)\right) g(y|x).$$

Suppose that $h(.)$ has support $\mathcal{H} = \{p^2(y)'\pi^2(x) : \left\|p^2(y)'\pi^2(x)\right\|_\infty \leq C$ for some constant $C\}$. For arbitrary $h(y, x)$ and θ, let $\pi^1(x, \theta, h)$ solve:

$$\begin{aligned} \int \rho(y, x, \theta) q\left(p^1(y)'\pi^1(x) + h(y, x)\right) g(y|x) dy &= 0, \\ \int q\left(p^1(y)'\pi^1(x) + h(y, x)\right) g(y|x) dy &= 1. \end{aligned}$$

Then $q\left(p^1(y)'\pi^1(x, \theta, h) + h(y, x)\right))$ satisfies the moment and density restrictions for arbitrary θ and h. The constrained optimization problem (**??**) can now be rewritten as the following unconstrained problem

$$\max_{h \in \mathcal{H}, \theta \in \Theta} E\{\ln\left[q\left(p^1(y)'\pi^1(x, \theta, h) + h(y, x)\right)\right]\}. \tag{6}$$

Let $\{a_1(x), a_2(x), ...\}$ denote series basis functions that can approximate any square-integrable function of x arbitrarily well. For some integers K_1 and K_2, denote

$$\begin{aligned} B^K(y, x) &= (b_1(y, x), b_2(y, x), ..., b_K(y, x))' \\ &= (a_1(x), a_2(x), ..., a_{K_1}(x))' \otimes (p_1^2(y), ..., p_{K_2}^2(y))', \end{aligned}$$

where $\otimes$ is the Kronecker product. $B^K(y,x)$ obviously denote basis functions that can approximate any function $h \in \mathcal{H}$ arbitrarily well in the sense that $\sup_{y,x} |h(y,x) - B^K(y,x)'\beta_K| \to 0$ as $K \to +\infty$ for some coefficients β_K. Denote $h_K(y,x) = B^K(y,x)'\beta_K$. The nonparametric ML estimator is defined as

$$(\widehat{\theta}, \widehat{\beta}_K) = \arg\max_{\theta,\beta_K} \sum_{i=1}^{n} \ln \left[q \left(p^1(y_i)'\pi^1(x_i,\theta,h_K) + B^K(y_i,x_i)'\beta_K \right) \right].$$

The nonparametric ML estimator for the density is

$$\widehat{f}(y|x) = q \left(p^1(y)'\pi^1(x,\widehat{\theta},\widehat{h}_K) + \widehat{h}_K(y,x) \right) q(y|x),$$

where $\widehat{h}_K(y,x) = B^K(y,x)'\widehat{\beta}_K$. Our main objective is to derive the asymptotic distribution of $\widehat{\theta}$ and prove consistency of $\widehat{f}(y|x)$.

Notice that the nonparametric ML estimation requires integration with respect to the endogenous variables y. In many applications, y has a low dimension. In these applications, the integration can be computed with numerical method (see Gallant and Nychka (1987) and Gallant and Tauchen (1989) for examples). In other applications, the nonparametric ML estimation requires high dimensional integration that cannot be computed accurately with numerical methods. In those applications, we propose to replace the integration by simulation draws from the density $g(y|x)$. To be specific, let $\{y^{ir}, r = 1, 2, ..., R\}$ denote independent simulation draws from the conditional density $g(y|x_i)$ for each x_i. The moment and density restrictions are replaced by

$$\begin{aligned} \frac{1}{R}\sum_{r=1}^{R} \rho(y^{ir},x_i,\theta) q \left(p(y^{ir})'\widetilde{\pi}^1(x_i,\theta,h_K) + B^K(y^{ir},x_i)'\beta_K \right) &= 0, \\ \frac{1}{R}\sum_{r=1}^{R} q \left(p(y^{ir})'\widetilde{\pi}^1(x_i,\theta,h_K) + B^K(y^{ir},x_i)'\beta_K \right) &= 1, \end{aligned}$$

where $\widetilde{\pi}^1(x_i,\theta,h_K)$ solves the above equations. The simulated nonparametric ML estimator is defined as

$$(\widetilde{\theta}, \widetilde{\beta}_K) = \arg\max_{\theta,\beta_K} \sum_{i=1}^{n} \ln \left[q \left(p^1(y_i)'\widetilde{\pi}^1(x_i,\theta,h_K) + B^K(y_i,x_i)'\beta_K \right) \right].$$

The nonparametric ML estimator for the density is

$$\widetilde{f}(y|x) = q \left(p^1(y)'\pi^1(x,\widetilde{\theta},\widetilde{h}_K) + \widetilde{h}_K(y,x) \right) q(y|x),$$

with $\widetilde{h}_K(y,x) = B^K(y,x)'\widetilde{\beta}_K$

It is worth noting that the above simulation approach requires $R * n$ simulation draws, which can be very large even for a moderate sample size and require large memory space to store them. To reduce the number of simulations and

consequently to conserve memory usage, one can replace $g(y|x)$ with a unconditional density function $g(y)$. With the unconditional density $g(y)$, we only need to generate a fixed simulation draws $\{y^r, r = 1, 2, ..., R\}$ from the density $g(y)$ for all of the observations and solve $\widetilde{\pi}^1(x_i, \theta, h_K)$ from the following equations:

$$\begin{aligned} \frac{1}{R}\sum_{r=1}^{R} \rho(y^r, x_i, \theta) q\left(p(y^r)'\widetilde{\pi}^1(x_i, \theta, h_K) + B^K(y^r, x_i)'\beta_K\right) &= 0, \\ \frac{1}{R}\sum_{r=1}^{R} q\left(p(y^r)'\widetilde{\pi}^1(x_i, \theta, h_K) + B^K(y^r, x_i)'\beta_K\right) &= 1. \end{aligned}$$

Asymptotically, both the conditional and the unconditional simulations will have no effects on the parameter estimators as long as the number of simulations, R, is sufficiently large.

3 Consistency

In this and next sections, we derive the asymptotic distribution of the estimators introduced in the section above. For the rest of the paper, we will use $E\{\cdot\}$ to denote the expectation taken with respect to the true density and use $E_f\{\cdot\}$ to denote the expectation taken with respect to the density f. Denote $\mathcal{A} = \Theta \times \mathcal{H}$ and denote $\alpha = (\theta, h)$ with $\alpha_o = (\theta_o, h_o)$. Let $\|\cdot\|_s$ denote a pseudo metric. For example,

$$\begin{aligned} \|\alpha - \alpha_o\|_s^2 &= (\theta - \theta_o)'(\theta - \theta_o) + \int (h(y,x) - h_o(y,x))^2 d\mu(y,x) \\ \text{or } \|\alpha - \alpha_o\|_s &= \max_{1 \le j \le d_\theta} |\theta_j - \theta_{jo}| + \sup_{y,x} |h(y,x) - h_o(y,x)| \mu(y,x) \end{aligned}$$

where $\mu(y, x)$ is a probability measure or waiting function and d_θ denotes the dimension of θ. First, we present sufficient conditions for consistency under the pseudo metric. The first condition is on how the sample is generated.

Assumption 3.1. *$\{(y_i, x_i), i = 1, 2, ..., n\}$ is an independent sample drawn from the joint density $f_o(y|x)f_o(x)$. The joint density $f_o(y|x)f_o(x)$ is unknown but satisfies (1).*

This condition is clearly restrictive since it rules out dependent data. However, the main result can be easily extended to weakly dependent data using the technique developed in Chen and Shen (1998). The next set of conditions identify the true value of the model parameters θ_o and $h_o(y, x)$.

Assumption 3.2. *The true value θ_o is the only value that satisfies (1). The true density $f_o(y|x)$ is the only solution to $\sup_{f(y|x)\in\mathcal{F}} E\{\ln(f(Y|X))\}$.*

Assumption 3.3. *(i) The series basis functions $p(y)$ are chosen such that $p(y)'\pi(x) = p(y)'\pi_o(x)$ for all y, x if and only if $\pi(x) = \pi_o(x)$; (ii) either*

$g(\cdot)$ is a monotone function over $(-\infty,+\infty)$ or $g(\cdot)$ is a monotone function over $[0,+\infty)$ *and $\mathcal{H}$ does not contain both $h(y,x) \neq 0$ and $-h(y,x)$.*

Assumption 3.2 identifies the true values θ_o and $f_o(y|x)$. This condition together with Assumption 3.3 identifies θ_o and $h_o(y,x)$. Assumption 3.3(i) requires that the basis functions are not perfectly correlated. Moreover, it requires that $E\{p(y)|x\}$ are not perfectly corrrelated. This condition is satisfied, for example, if the basis functions are orthonormal conditional on X. If the basis functions are not orthonormal, it is common practice to require that the minimum eigenvalue of $E\{p_{KK}(Y)p_{KK}(y)'|X\}$ is bounded away from zero for all K and X, where $p_{KK}(y) = (p_1(y), p_2(y), ..., p_K(y))'$ (see Newey (1997)). Assumption 3.3(ii) may require some restrictions on $h(\cdot)$. For instance, for the square transformation function, $q(u) = u^2 + c_n$, it is easy to show that

$$q\left(p^1(y)'\pi^1(x,\theta,h) + h(y,x)\right) = q\left(-p^1(y)'\pi^1(x,\theta,h) - h(y,x)\right).$$

Clearly, without some restrictions on $h(\cdot)$, $h_o(.)$.is not identified since both $h_o(.)$ and $-h_o(.)$ give the same density function. This problem can be corrected by a simple restriction on any $h \in \mathcal{H}$ such as $h(1,1) > 0$. Denote $\mathcal{A}_k = \Theta \times \mathcal{H}_k$ with $\mathcal{H}_k = \{h_K(y,x) = B^K(y,x)'\beta_K : \|h_K(y,x)\|_\infty \leq C\}$. Denote $\widetilde{\rho}(z,\theta) = (\rho(z,\theta)', 1)'$.

Assumption 3.4. *The closure of $\mathcal{A}$ with respect to $\|\alpha\|_s$ is compact in the relative topology generated by $\|\alpha\|_s$.*

Assumption 3.5. *(i) For any $x, \theta \in \Theta$ and $h \in \mathcal{H}$, $\pi^1(x,\theta,h)$ is well defined. (ii) $\rho(y,x,\theta)$ is twice continuously differentiable with respect to θ.*

Assumption 3.6. *The random variables X have a bounded support and a density function that is bounded and bounded away from zero.*

Assumption 3.7. *(i) For some constant c, $\ln(g(u))| \leq c * |u|$ and $\left|\frac{g'(u)}{g(u)}\right| \leq c$; (ii) $E\{|p^1(y)|^\gamma\} < +\infty$ for some $\gamma > 4$.*

Assumption 3.8. *(i) $\cup_{k=1}^{\infty}\mathcal{A}_k$ is dense in the closure of $\mathcal{A}$ with respect to $\|\alpha\|_s$. (ii) $K \to +\infty$ and $\frac{K}{n} \to 0$.*

Assumption 3.9. *Both $\rho(z,\theta)$ and its derivative with respect to θ are dominated by some $C(Z)$ satisfying $E\{C(Z)^2\} < \theta$.*

The compact condition of Assumption 3.4 is commonly imposed in the literature (e.g. Gallant and Nychka (1987)). This condition is convenient for establishing consistency. Assumption 3.5(i) basically requires that $p^1(y)$ is highly correlated with $\rho(z,\theta)$ for any $\theta \in \Theta$ so that the solution $\pi^1(x,\alpha)$ always exists and is unique. This condition must hold for the proposed approach to work. Assumption 3.5(ii) is familiar in the nonlinear econometric literature and can be verified by inspection. Assumption 3.5 implies that $\pi^1(x,\theta,h)$ is twice continuously differentiable with respect to θ and has up to second (directional)

derivatives with respect to h. Assumption 3.6 is made for convenience. It can always be satisfied by discarding large regressors' values. This condition together with Assumption 3.4 implies that $\left|\pi^1(x,\alpha)-\pi^1(x,\alpha')\right| \leq c * \|\alpha-\alpha'\|_s$ for some constant c. The dominance condition of Assumption 3.9 is also familiar in the nonlinear econometric literature. It is satisfied by the exponential transformation function. The dominance condition is needed to show that the simulated integration converges to the true integration uniformly and hence $\widetilde{\pi}^1(x,\alpha)$ converges in probability to $\pi^1(x,\alpha)$ uniformly.

Applying the uniform convergence result (Lemma A.1 of Ai and Chen (2003)) and the consistency result (Theorem 0 of Gallant and Nychka (1987)), we obtain:

Theorem 3.1. *Under Assumptions 3.1 - 3.8, we have* $\|\widehat{\alpha}-\alpha_o\|_s \to 0$ *in probability. Under additional Assumption 3.9 and* $R \to +\infty$*, we obtain* $\|\widetilde{\alpha}-\alpha_o\|_s \to 0$ *in probability.*

The above consistency result under the strong metric $\|\cdot\|_s$ is useful but not enough for deriving the root-n consistency of the estimator $\widehat{\theta}$ (and $\widetilde{\theta}$). To show root-n consistency of the estimated finite dimensional parameters, the strong metric $\|\cdot\|_s$ is not needed, as pointed out by Ai and Chen (2003) and Chen and Shen (1998). What is needed here is the following weaker metric:

$$\|\alpha-\alpha_o\|^2 = E\left\{\left(\frac{\partial l(Y,X,\alpha_o)}{\partial\theta'}(\theta-\theta_o)+\frac{dl(Y,X,\alpha_o)}{dh}[h-h_o]\right)^2\right\},$$

where $l(Y,X,\alpha)=\ln\left[q\left(p^1(Y)'\pi^1(X,\alpha)+h(Y,X)\right)\right]$ and $\frac{dl(Y,X,\alpha_o)}{dh}[h-h_o]$ denotes the directional derivative with respect to h. The strong metric $\|\cdot\|_s$ is needed, however, to bring the parameter α to the neighborhood of the true value α_o. Thus, $\|\cdot\|_s$ should be chosen so that the weaker metric $\|\alpha-\alpha_o\|^2$ is equivalent to $E\{l(Y,X,\alpha_o)-l(Y,X,\alpha)\}$ in the neighborhood of α_o defined as $\{\alpha\in\mathcal{A}:\|\alpha-\alpha_o\|_s\leq\epsilon\}$ for some small $\epsilon>0$. And the weaker metric $\|\alpha-\alpha_o\|^2$ can be interpreted as the local quadratic approximation to the average Kullback-Leibler information.

We now present additional conditions and compute the convergence rates under the weaker metric $\|\cdot\|$. Let $[N(\varepsilon,\mathcal{A}_n,||\cdot||_s)$ denote the number of covering balls with radius ε that cover the approximating spaces $\mathcal{A}_n$.

Assumption 3.10. *The strong metric* $\|\cdot\|_s$ *is chosen so that the weaker metric* $\|\alpha-\alpha_o\|^2$ *is equivalent to* $E\{l(Y,X,\alpha_o)-l(Y,X,\alpha)\}$ *over* $\{\alpha\in\mathcal{A}:\|\alpha-\alpha_o\|_s\leq\epsilon\}$ *for some small* $\epsilon>0$.

Assumption 3.11. *For any* $\alpha\in\mathcal{A}$*, there exists* $\alpha_k\in\mathcal{A}_K$ *satisfying*

$$\|\alpha-\alpha_K\|_s = O(K^{-\zeta}) = o(n^{-1/4}).$$

Assumption 3.12. $\ln[N(\varepsilon,\mathcal{A}_n,||\cdot||_s)]\leq const.\times k\times\ln(\frac{k}{\varepsilon})$.

Assumption 3.13. $\frac{K\ln(n)}{\sqrt{n}}\to 0$ *as* $n\to+\infty$.

Assumption 3.14. $R = O(n)$.

Assumption 3.11 imposes restriction on the approximation error of the parameter space $\mathcal{A}$ by the sieve spaces $\mathcal{A}_K$ as well as on the number of the approximating functions, K. The approximation error must shrink at a polynomial rate as the sample size goes to infinity. This condition is satisfied if the parameter space $\mathcal{H}$ is a Sobolev, or Besov, or Holder space and the approximating functions are power series, or splines, or wavelet series. The restriction on the number of the approximating functions, K, imposes a lower bound on the rate at which K goes to infinity. For example, $K = \frac{n^{1/(4\zeta)}}{\ln(n)}$ satisfies the rate restriction. Assumption 3.12 restricts the size of the sieve spaces $\mathcal{A}_K$. It requires that the sieve spaces do not grow too fast. This condition is satisfied by commonly used sieves. For instance, it is satisfied by power series and splines since $\ln[N(\varepsilon, \mathcal{A}_n, ||\cdot||_s)] = const * k * \ln(\frac{1}{\varepsilon})$. Assumption 3.13 imposes another restriction on K. It requires that K does not grow too fast as the sample size goes to infinity. This condition and Assumption 3.12 together put a lower and an upper bound on the rate of $K \to +\infty$.

The following theorem is proved in the appendix.

Theorem 3.2. *Under Assumptions 3.1 - 3.13, we obatin* $\|\widehat{\alpha} - \alpha_o\| = o_p(n^{-1/4})$. *Under additional Assumption 3.14, we have* $\|\widetilde{\alpha} - \alpha_o\| = o_p(n^{-1/4})$.

The convergence rate under the weaker metric derived in the above theorem is the minimum rate needed for proving the $\sqrt{n}$ consistency. Faster rate can be obtained if some of the sufficient conditions, particularly Assumptions 3.11 and 3.13, are strengthened. In some applications, especially where the objective function is highly nonlinear, faster rate is absolutely necessary for obtaining $\sqrt{n}$ consistency of the estimated finite dimensional parameter.

4 Asymptotic Distribution

We now present sufficient conditions to derive the asymptotic distribution of $\widehat{\theta}$ and $\widetilde{\theta}$. Notice that it is sufficient to derive the asymptotic distribution of the linear functional $f(\alpha) \equiv \lambda'\theta$ for any fixed and nonzero $\lambda \in \mathcal{R}^{d_\theta}$. Following the approach first developed in Shen (1997) and then applied by Chen and Shen (1998) and Ai and Chen (2003), we first express the linear functional $f(\alpha)$ as a Riesz representation and then derive the asymptotic distribution of the Riesz representation. Specifically, let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A} - \{\alpha_o\}$ under the metric $||\cdot||$. Then $(\overline{\mathbf{V}}, ||\cdot||)$ is a Hilbert space with the inner product:

$$\langle \alpha, \overline{\alpha} \rangle = E \left\{ \begin{array}{c} \left(\frac{\partial l(Y,X,\alpha_o)}{\partial \theta'} \theta + \frac{dl(Y,X,\alpha_o)}{dh}[h] \right) * \\ \left(\frac{\partial l(Y,X,\alpha_o)}{\partial \theta'} \overline{\theta} + \frac{dl(Y,X,\alpha_o)}{dh}[\overline{h}] \right) \end{array} \right\}.$$

By the results in Van der Vaart (1991) and Shen (1997), the linear functional $f(\alpha) = \lambda'\theta$ must be *bounded* (i.e. $\sup_{0\neq\alpha-\alpha_o\in\overline{\mathbf{V}}} \frac{|f(\alpha)-f(\alpha_o)|}{||\alpha-\alpha_o||} < \infty$) in order for it to be estimated at a $\sqrt{n}-$ rate. Also, the Riesz representation exists if and only if $f(\alpha)$ is bounded. Thus, our immediate task is to show that the linear functional is bounded.

Write $\overline{\mathbf{V}} = \mathcal{R}^{d_\theta} \times \overline{\mathcal{W}}$ with $\overline{\mathcal{W}} \equiv \overline{\mathcal{H}} - \{h_o\}$. For each component θ_j (of θ), $j = 1, ..., d_\theta$, let $w_j^* \in \overline{\mathcal{W}}$ denote the solution to

$$\min_{w_j\in\overline{\mathcal{W}}} E\left\{\left(\frac{\partial l(Y,X,\alpha_o)}{\partial\theta_j} - \frac{dl(Y,X,\alpha_o)}{dh}[w_j]\right)^2\right\}. \tag{7}$$

Define $w^* = (w_1^*, ..., w_{d_\theta}^*)$, $\frac{dl(Y,X,\alpha_o)}{dh}[w^*] = (\frac{dl(Y,X,\alpha_o)}{dh}[w_1^*], ..., \frac{dl(Y,X,\alpha_o)}{dh}[w_{d_\theta}^*])$, and

$$D_{w^*}(Y,X) \equiv \frac{\partial l(Y,X,\alpha_o)}{\partial\theta'} - \frac{dl(Y,X,\alpha_o)}{dh}[w^*].$$

It is easy to show that

$$\sup_{0\neq\alpha-\alpha_o\in\overline{\mathbf{V}}} \frac{|f(\alpha)-f(\alpha_o)|^2}{||\alpha-\alpha_o||^2} = \lambda'\left(E\{D_{w^*}(Y,X)'D_{w^*}(Y,X)\}\right)^{-1}\lambda.$$

Suppose that $E\{D_{w^*}(Y,X)'D_{w^*}(Y,X)\}$ is finite positive-definite. Then $f(\alpha) = \lambda'\theta$ is bounded and has the following Riesz representation:

$$f(\alpha) - f(\alpha_o) \equiv \lambda'(\theta-\theta_o) = \langle v^*, \alpha-\alpha_o\rangle \quad \text{for all } \alpha\in\mathcal{A}$$

where $v^* \equiv (v_\theta^*, v_h^*) \in \overline{\mathbf{V}}$ with $v_\theta^* = (E\{D_{w^*}(Y,X)'D_{w^*}(Y,X)\})^{-1}\lambda$, $v_h^* = -w^* \times v_\theta^*$. Hence, the asymptotic distribution of $f(\widehat{\alpha})-f(\alpha_o)$ (and $f(\widetilde{\alpha})-f(\alpha_o)$ in the case of simulated estimation) is the same as the asymptotic distribution of $\langle v^*, \widehat{\alpha}-\alpha_o\rangle$ (and $\langle v^*, \widetilde{\alpha}-\alpha_o\rangle$).

Before derive the asymptotic distribution of the Riesz representation, we notice that the moment restriction (1) implies:

$$E\left\{\rho(Y,X,\theta_o)\frac{\partial l(Y,X,\alpha_o)}{\partial\theta'}|X\right\} + E\left\{\frac{\partial\rho(Y,X,\theta_o)}{\partial\theta'}|X\right\} = 0.$$

Write

$$\begin{aligned}
&\frac{\partial l(Y,X,\alpha_o)}{\partial\theta_j} \\
= \;& -\rho(Y,X,\theta_o)'E\left\{\rho(Y,X,\theta_o)\rho(Y,X,\theta_o)'|X\right\}^{-1}E\left\{\frac{\partial\rho(Y,X,\theta_o)}{\partial\theta_j}|X\right\} \\
& +\upsilon_j(Y,X).
\end{aligned}$$

Then, it is easy to show that $E\left\{\rho(Y,X,\theta_o)\upsilon_j(Y,X)|X\right\} = 0$. Again, the moment restriction (1) implies

$$E\left\{\rho(Y,X,\theta_o)\frac{dl(Y,X,\alpha_o)}{dh}[w]|X\right\} = 0$$

for any $w \in \overline{\mathcal{W}}$. We show in the appendix that the tangent space which is the closure of

$$\Gamma = \{\frac{dl(Y,X,\alpha_o)}{dh}[w] : E\left\{\rho(Y,X,\theta_o)\frac{dl(Y,X,\alpha_o)}{dh}[w]|X\right\} = 0 \text{ and } w \in \overline{\mathcal{W}}$$

contains $v_j(Y,X)$. Hence,

$$\begin{aligned} D_{w^*}(Y,X) &\equiv \frac{\partial l(Y,X,\alpha_o)}{\partial \theta'} - \frac{dl(Y,X,\alpha_o)}{dh}[w^*] \\ &= -\rho(Y,X,\theta_o)' * E\left\{\rho(Y,X,\theta_o)\rho(Y,X,\theta_o)'|X\right\}^{-1} * \\ &\quad E\left\{\frac{\partial \rho(Y,X,\theta_o)}{\partial \theta'}|X\right\} \end{aligned}$$

and

$$\begin{aligned} & E\{D_{w^*}(Y,X)'D_{w^*}(Y,X)\} \\ = \; & E\left\{\begin{array}{c} E\left\{\frac{\partial \rho(Y,X,\theta_o)}{\partial \theta'}|X\right\}' * \\ E\left\{\rho(Y,X,\theta_o)\rho(Y,X,\theta_o)'|X\right\}^{-1} E\left\{\frac{\partial \rho(Y,X,\theta_o)}{\partial \theta'}|X\right\} \end{array}\right\} \\ = \; & E\left\{E\left\{\frac{\partial \rho(Y,X,\theta_o)}{\partial \theta'}|X\right\}' \Sigma_o^{-1}(X) E\left\{\frac{\partial \rho(Y,X,\theta_o)}{\partial \theta'}|X\right\}\right\} \\ = \; & V_o, \end{aligned}$$

which is exactly the semiparametric efficiency information of θ_o for model (1). Thus, our estimator is asymptotically efficient if its asymptotic covariance is the inverse of V_o.

The following conditions are sufficient for establishing the $\sqrt{n}$ - consistency of the estimators $\widehat{\theta}_n$: and $\widetilde{\theta}_n$.

Assumption 4.1. *(i)* $\Sigma_o(X) = E\left\{\rho(Y,X,\theta_o)\rho(Y,X,\theta_o)'|X\right\}$ *is nonsingular for all* X*; (ii)* V_o *is nonsingular; (iii)* $\theta_o \in int(\Theta)$.

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

Assumption 4.3. *(i) For some constant* c, $\left|\frac{g"(u)}{g(u)}\right| \leq c$*; (ii)*

$$E\left\{\frac{d^2 l(Y,X,\alpha)}{dh^2}[\alpha - \alpha_o, v_n^*] - \frac{d^2 l(Y,X,\alpha_o)}{dh^2}[\alpha - \alpha_o, v_n^*]\right\} = o(n^{-1/2})$$

for all $\alpha \in \mathcal{A}_n$ *and* $\|\alpha - \alpha_o\| \leq o(n^{-1/4})$.

Assumption 4.4. $R = O(n \ln n)$.

Assumption 4.1(i)(ii) is a local identification condition for θ_o. This condition must be satisfied for the estimated finite dimensional parameters to be $\sqrt{n}$ consistent. Unfortunately, this condition is difficult to verify in practice since

it requires knowing the true value of the model. Assumption 4.2 is a "bias controlling" condition. This condition is needed due to the presence of unknown h_o. Here for simplicity we assume that the same sieve space $\mathcal{H}_n$ approximates the space $\overline{\mathcal{W}} \equiv \overline{\mathcal{H}} - \{h_o\}$ well. Theorem 4.1 can be proved even if $v_h^* = -w^* v_\theta^*$ is approximated by any other sieve spaces. Assumption 4.4 is needed to ensure that the simulation has no effect on the asymptotic distribution of $\widetilde{\theta}$.The following result is proved in the appendix.

Theorem 4.1: *Under Assumptions 3.1 - 3.13 and 4.1 - 4.2,* $\sqrt{n}(\widehat{\theta}_n - \theta_o) \Longrightarrow N(0, V_o^{-1})$; *under additional assumptions 3.14 and 4.4, we obtain:* $\sqrt{n}(\widetilde{\theta}_n - \theta_o) \Longrightarrow N(0, V_o^{-1})$.

The result of Theorem 4.1 simply states that the nonparametric ML estimator of θ_o has the same first order properties as do the estimators proposed in Newey (1990), Kitamura, Tripathi and Ahn (2004), and Donald, Imbens, and Newey (2004). It is not clear whether these same sufficient conditions also ensure that the nonparametric ML estimator of θ_o has better second order properties. Our guess is that these sufficient conditions, particularly Assumptions 3.11, 3,13, and 4.4, need to be strengthened in order for the nonparametric ML estimator to have better second order properties. For example, Assumption 4.4 is sufficient for the simulation to have no effect on the first order properties. To ensure that the simulation has no effect on the second order properties, we probably need larger number of simulation draws such as $R = O(n^2)$. The Assumption 3.11 and 3.13 give a range of values of K that are all sufficient for the first order efficiency. This range probably needs to be tightened up to obtain the second order superiority.

5 Covariance Matrix

The asymptotic distribution derived above can be used for statistical inference only if a practical and easy to compute covariance estimator is available. One way to estimate the asymptotic covariance is to use the estimated conditional density. For example, we can estimate the conditional covariance matrix $\Sigma_o(X)$ by

$$\widehat{\Sigma}_o(x) \equiv \int \rho(y, x, \widehat{\theta}) \rho(y, x, \widehat{\theta})' \widehat{f}(y|x) dy$$

and V_o by

$$\widehat{V}_o = \frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial \rho(y, x_i, \widehat{\theta})'}{\partial \theta} \widehat{f}(y|x_i) dy * \widehat{\Sigma}_o^{-1}(x_i) \int \frac{\partial \rho(y, x_i, \widehat{\theta})}{\partial \theta'} \widehat{f}(y|x_i) dy.$$

In the case of simulation, we can estimate $\Sigma_o(X)$ by

$$\widetilde{\Sigma}_o(x_i) \equiv \frac{1}{R} \sum_{r=1}^{R} \rho(y^{ir}, x_i, \widetilde{\theta}) \rho(y^{ir}, x_i, \widetilde{\theta})' q(p^1(y^{ir})' \widetilde{\pi}^1(x_i, \widetilde{\alpha}) + \widetilde{h}_K(y^{ir}, x_i))$$

and V_o by

$$\begin{aligned}\widetilde{V}_o &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{A}(x_i) * \widetilde{\Sigma}_o^{-1}(x_i) * \widetilde{A}(x_i)', \text{ where} \\ \widetilde{A}(x_i) &= \frac{1}{R}\sum_{r=1}^{R}\frac{\partial\rho(y^{ir}, x_i, \widetilde{\theta})'}{\partial\theta} q(p^1(y^{ir})'\widetilde{\pi}^1(x_i, \widetilde{\alpha}) + \widetilde{h}_K(y^{ir}, x_i)).\end{aligned}$$

An alternative approach is to estimate $w^* = (w_1^*, ..., w_{d_\theta}^*)$, defined in (7) above, by simple OLS regression. Specifically, for each component θ_j, $j = 1, ..., d_\theta$, we approximate w_j^* by $B^K(y, x)'\delta_K$ and estimate δ_K by

$$\text{regressing } \frac{\partial l(y_i, x_i, \widehat{\alpha})}{\partial\theta_j} \text{ on } \frac{dl(y_i, x_i, \widehat{\alpha})}{dh}[b_1(y_i, x_i)], ..., \frac{dl(y_i, x_i, \widehat{\alpha})}{dh}[b_K(y_i, x_i)].$$

Notice that the above regression is the same as

$$\text{regressing } \frac{\partial l(y_i, x_i, \widehat{\alpha})}{\partial\theta_j} \text{ on } \frac{\partial l(y_i, x_i, \widehat{\alpha})}{\partial\beta_{1K}}, ..., \frac{\partial l(y_i, x_i, \widehat{\alpha})}{\partial\beta_{KK}}.$$

The regression residuals from the above regression are the estimates of $D_{w^*}(y, x)$. Let $\widehat{D}_{ji}$ denote the regression residuals and denote $\widehat{D}_i = (\widehat{D}_{1i}, ..., \widehat{D}_{d_\theta i})'$. Then we estimate V_o by

$$\widehat{V}_o = \frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{D}_i'.$$

Similarly, we can construct the covariance estimator for the simulated case.

$$\text{Regress } \frac{\partial l(y_i, x_i, \widetilde{\alpha})}{\partial\theta_j} \text{ on } \frac{\partial l(y_i, x_i, \widetilde{\alpha})}{\partial\beta_{1K}}, ..., \frac{\partial l(y_i, x_i, \widetilde{\alpha})}{\partial\beta_{KK}}.$$

Let $\widetilde{D}_{ji}$ denote the regression residuals and denote $\widetilde{D}_i = (\widetilde{D}_{1i}, ..., \widetilde{D}_{d_\theta i})'$. Then we estimate V_o by

$$\widetilde{V}_o = \frac{1}{n}\sum_{i=1}^{n}\widetilde{D}_i\widetilde{D}_i'.$$

The following theorem is proved in the appendix.

Theorem 5.1: *Under Assumptions 3.1 - 3.14 and 4.1 - 4.4, we have:* $\widehat{V}_o = V_o + o_p(1)$ *and* $\widetilde{V}_o = V_o + o_p(1)$.

It follows from Theorem 4.1 and 5.1 that the usual t-statistics, computed as the ratios of the parameter estimates $\widehat{\theta}$ ($\widetilde{\theta}$) divided by their respective estimated standard errors have standard normal distribution and hence standard t- test for significance is still valid. To test joint restrictions on θ_o, the usual Wald test and Hausman test are still valid. The likelihood ratio test should also work here. To test the restriction on the density function, in principal the likelihood ratio test still applies but the asymptotic distribution of the test statistic needs to be worked out.

It is worth pointing out that the covariance matrices of $\widehat{\theta}$ and $\widetilde{\theta}$ computed in the second approach are the same as the covariance matrices of the ML estimators of $\widehat{\theta}$ and $\widetilde{\theta}$ if $B^K(y,x)'\beta_K$ is treated as the parametric specification of $h(y,x)$ and K is fixed. Thus, the covariance matrices can be computed from any standard maximum likelihood estimation statistical package.

6 Conclusion

In this paper, we study the nonparametric maximum likelihood estimation of the conditional moment restriction model. We present some sufficient conditions and show that the estimated finite dimensional parameter is $\sqrt{n}$ consistent and asymptotically normally distributed and the estimated density function is consistent. We provide an easy to compute and consistent covariance matrix and show that the covariance matrix is the same as the covariance matrix of the maximum likelihood estimator if the sieve approximation is treated as the correct parametric specification, and hence can be computed from any standard statistical package that computes the maximum likelihood estimation. We also argue that it is possible that the nonparametric maximum likelihood estimator may have potential advantages over the empirical maximum likelihood estimator. It is unclear however whether those advantages truly exist in finite samples. Moreover, we are not certain that the nonparametric maximum likelihood estimator has better second order properties than the empirical maximum likelihood estimator since we have to approximate the true density function. The higher order properties of the proposed estimator will be explored in a future study. The issue of testing restrictions on the density function also is important and will be dealt with in a separate paper.

In our moment restriction model, we only permit finite dimensional parameter θ. In some applications, the conditional moment restriction may contain unknown functions. Thus, it is necessary to extend the result here to a model similar to the one studied by Ai and Chen (2003). Ai and Chen proposed a minimum distance estimator which is very similar to the GMM formulation and hence may suffer from the same finite sample problems. The extension of the nonparametric maximum likelihood estimation to that model would be useful addition to the literature. This extension will be pursued in a separate paper.

7 Reference

Ai, C. (1997): "A Semiparametric Maximum Likelihood Estimator," *Econometrica,* 65, pp. 933-964.

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

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

Altonji, J.G. and Segal, L.M. (1996), "Small Sample Bias in GMM Estimation of Covariance Structures", *Journal of Economics and Business Statistics* 14, pp. 353 - 366.

Chamberlain, Gary (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics,* 34, pp. 305-334.

———————(1992): "Efficiency Bounds for Semiparametric Regression", *Econometrica* 60, pp. 567 - 596.

Donald, Imbens, and Newey (2004), "Empirical likelihood estimation and consistent tests with conditional moment restrictions", *Journal of Econometrics* 117, pp. 55-93.

Fenton, V. and A.R. Gallant (1996), "Convergence Rates of SNP Density Estimators", *Econometrica* 64, pp. 719-727

Gallant, A.R. and D.W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation", *Econometrica* 55, pp. 363-390

Gallant, A.R. and G. Tauchen (1989): "Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications", *Econometrica* 57, pp. 1091-1120

Hansen, Lars P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica* 50, pp. 1029 - 1054.

Imbens, G.W. (1997), "One-Step Estimators for Over-identified Generalized Method of Moments Models", *Review of Economic Studies* 64, pp. 359 - 383.

Imbens, G.W., R.H. Spady, and P. Johnson (1998), "Informtion Theoretic Approaches to Inference in Moment Conditions Models", *Econometrica* 66, pp. 333 - 357.

Kitamura, Yuichi and M. Stutzer (1997), "An Information-Theoretic Alternative to Generalized Method of Moments Estimation", *Econometrica* 65, pp. 861 - 874.

Kitamura, Y., G. Tripathi, and H. Ahn (2004), "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models", *Econometrica* 72, pp. 1667-1714.

Newey, Whitney K. (1990) "Efficient Instrumental Variables Estimation of Nonlinear Models", *Econometrica* 58, pp. 809 - 837.

Newey, W.K. and R. Smith (2004), "Higher Order Properties of Gmm and Generalized Empirical Likelihood Estimators", *Econometrica* 72, pp. 219-255.

Owen, A. (1990), "Empirical Likelihood Ratio Confidence Regions", *The Annals of Statistics* 18, pp. 90 - 120.

———(1991), "Empirical Likelihood for Linear Models", *The Annals of Statistics* 19, pp. 1725 - 1747.

Qin, J. and J. Lawless (1994), "Empirical Likelihood and General Estimating Equations", *The Annals of Statistics* 22, pp. 300-325.

Severini, T. and H.W. Wong (1992): "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics* 20, 1768-1802.

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

Van der Vaart, A. (1991): "On Differentiable Functionals," *The Annals of Statistics* 19, pp. 178-204.