Sieve-based Empirical Likelihood under Semiparametric Conditional Moment Restrictions

Martin Burda*
Department of Economics, University of Toronto

First draft: November 7, 2006
This version: June 8, 2007

Abstract

In this paper we propose a new Sieve-based Locally Weighted Conditional Empirical Likelihood (SLWCEL) estimator for models of conditional moment restrictions containing finite dimensional unknown parameters θ and infinite dimensional unknown functions h. The SLWCEL is a one-step information-theoretic alternative to the Sieve Minimum Distance estimator analyzed by Ai and Chen (2003). We approximate h with a sieve and estimate both θ and h simultaneously conditional on exogenous regressors. Thus, the estimator permits dependence of h on endogenous regressors and θ. We establish consistency and convergence rates for the estimator and asymptotic normality for its parametric component of θ. The SLWCEL generalizes in two ways the Conditional Empirical Likelihood (CEL) of Kitamura, Tripathi and Ahn (2004). First, we construct the CEL's dual global MD-objective function with a new weighting scheme that adapts to local inhomogeneities in the data. Second, we extend the resulting new estimator into the semiparametric environment defined by the presence of h. We show that the corresponding estimator of θ exhibits better finite-sample properties than found in the previous literature.

Keywords: Semi-/nonparametric conditional moment restrictions, empirical likelihood, sieve estimation, endogeneity.
JEL Classification: C13, C14, C20, C30.

*I am grateful to Mehmet Caner, Xiaohong Chen, George-Levi Gayle, Soiliou Daw Namoro, Taisuke Otsu, Eric Renault and Nese Yildiz for insightful comments and suggestions. I would also like to thank participants of the 16^{th} Annual Meeting of the Midwest Econometrics Group, Cincinnati, OH, October 2006, the 2^{nd} PhD Presentation Meeting at LSE, London, UK, January 2007, and seminar participants at Georgetown, Pittsburgh, Purdue, Oxford (Nuffield), Simon Fraser, Tilburg, Toronto, UNC Chapel Hill and Warwick. Any updates of the paper will be made available at www.pitt.edu/∼mab256. Address for correspondence: Martin Burda, Department of Economics, University of Pittsburgh, 4521 Posvar Hall, 230 S. Bouquet St., Pittsburgh, PA 15260, USA. E-mail: mab256@pitt.edu; Phone: 412-600-7293; Fax: 412-648-1793.

1 Introduction

Moment restrictions frequently provide the basis for estimation and inference in economic problems. A general framework for analyzing economic data (Y, X) is to postulate conditional moment restrictions of the form

$$E\left[g\left(Z, \alpha_0\right)|X\right] = 0 \tag{1}$$

where $Z \equiv (Y', X_z')'$, Y is a vector of endogenous variables, X is a vector of conditioning variables (instruments), X_z is a subset of X, $g(\cdot)$ is a vector of functions known up a parameter α, and $F_{Y|X}$ is assumed unknown. The parameters of interest $\alpha_0 \equiv (\theta_0', h_0')'$ contain a vector of finite dimensional unknown parameters θ_0 and a vector of infinite dimensional unknown functions $h_0(\cdot) \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))'$. The inclusion of h_0 renders the condition (1) semiparametric, encompassing many important economic models. It includes for example the partially linear regression $g\left(Z, \alpha_0\right) = Y - X_1'\theta_0 - h_0(X_2)$ analyzed by Robinson (1988) and the index regression $g\left(Z, \alpha_0\right) = Y - h_0(X'\theta_0)$ studied by Powell et al. (1989) and Ichimura (1993).

Recently, Kitamura, Tripathi and Ahn (2004) analyzed the Conditional Empirical Likelihood (CEL)[1] based on a parametric counterpart of (1) (with θ only) that was shown to exhibit finite-sample properties superior to the Generalized Method of Moments. In this paper we first suggest a new Locally Weighted CEL (LWCEL) that fundamentally changes the form of CEL and further improves on it in terms of finite-sample properties. Then we extend the LWCEL to the semiparametric environment of model (1) proposing new Sieve-based Locally Weighted Conditional Empirical Likelihood (SLWCEL) estimator. The SLWCEL can be viewed as a one-step information-theoretic alternative to the Sieve Minimum Distance (SMD) estimator analyzed by Ai and Chen (2003). In the remainder of the introduction we will elaborate on the heuristic origins of both estimators, and further analysis will follow thereafter.

1.1 Conditional Moments Based on θ_0

Without the unknown functions h_0, model (1) becomes the parametric model of conditional moment restrictions

$$E\left[g\left(Z, \theta_0\right)|X\right] = 0 \tag{2}$$

Typically, faced with the model (2) for estimation of θ_0, researchers would pick an arbitrary matrix-valued function $a(X)$ and estimate the unconditional moment model $E\left[a(X)g\left(Z, \theta_0\right)\right] = 0$ implied by (2) with an estimator such as the Generalized Method of Moments (GMM) (see e.g. Kitamura, 2006,

[1] A note on terminology: CEL is called "smoothed" and "sieve" empirical likelihood in KTA and Zhang and Gijbels (2003), respectively. Other types of smoothing have been introduced by Otsu (2003a) on moment restrictions in the quantile regression setting and hence KTA's original method is referred to as "conditional" empirical likelihood to avoid confusion. The CEL terminology was also adopted in Kitamura (2006).

p 26 for a discussion). This procedure is used under the presumption that the chosen instrument $a(X)$ identifies θ, which may not be true even if θ is identified in the conditional model (2) (Domínguez and Lobato, 2004). Moreover, the conversion to unconditional moments results in a loss of efficiency with respect to the information contained in (2). Chamberlain (1987) showed that such loss can be avoided by using the optimal IV estimator $a^*(X) = D'(X)V^{-1}(X)$ where $D(X) = E\left[\nabla_\theta g\left(Z,\theta_0\right)|X\right]$ and $V(X) = E\left[g\left(Z,\theta_0\right)g\left(Z,\theta_0\right)'|X\right]$. In practice, $a^*(X)$ can be estimated with a two-step procedure (Robinson, 1987; Newey, 1993). First an inefficient preliminary estimator $\widetilde{\theta}$ for θ_0 is obtained and the unknown functions $D(X)$ and $V(X)$ are estimated via a nonparametric regression of $\nabla_\theta g(Z,\widetilde{\theta})$ and $g(Z,\widetilde{\theta})g(Z,\widetilde{\theta})'$ on X. Second, the estimate of $a^*(X)$ is constructed with the estimates of $D(X)$ and $V(X)$ from the first step. However, as noted by Domínguez and Lobato (2004), the resulting moment condition $E\left[a^*(X)g\left(Z,\theta_0\right)\right] = 0$ may fail to identify θ while θ is identified under the original model (2). Moreover, satisfactory implementation of the nonparametric regression may require large samples thereby affecting the finite-sample performance of the feasible estimator of $a^*(X)$.

The methods typically employed for estimation of the unconditional model $E\left[a(X)g\left(Z,\theta_0\right)\right] = 0$ have also been subject to criticism. While the optimally-weighted two-step GMM (Hansen, 1982) is first-order asymptotically efficient, its finite sample properties have been reported as relatively poor. For example, a simulation study by Altonji and Segal (1996) documented a substantial small-sample bias of GMM when used to estimate covariance models. Other Monte Carlo experiments have shown that tests based on GMM often have true levels that differ greatly from their nominal levels when asymptotic critical values are used (Hall and Horowitz, 1996). Indeed, it has been widely recognized that the first-order asymptotic distribution of the GMM estimator provides a poor approximation to its finite-sample distribution (Ramalho, 2005).

A number of alternative estimators have been suggested to overcome this problem: Empirical Likelihood (EL) (Owen, 1988; Qin and Lawless, 1994; Imbens, 1997), the Euclidean Likelihood (EuL) corresponding to the Continuous Updating Estimator (CUE) (Hansen et al., 1996) the Exponential Tilting Estimator (ET) (Kitamura and Stutzer, 1997; Imbens et al., 1998), and variations on these such as the Exponentially Tilted Empirical Likelihood (ETEL) (Schennach, 2006). The EL, EuL and ET share some common properties and can be derived from a common model basis for estimation. Thus, they and can be viewed as members of broader classes of estimators such as the Generalized Empirical Likelihood (GEL) estimators (Smith, 1997; Newey and Smith, 2004) and the Generalized Minimum Contrast (GMC) estimators (Bickel et al., 1998). Recently, Kitamura (2006) showed that for unconditional moment restriction models, the GEL class is essentially equivalent to the GMC class even if the GEL are derived somewhat differently from the GMC. Both GEL and GMC lead to the same saddle-point optimization problem yielding the same form the individual estimators.

The GEL/GMC estimators circumvent the need for estimating a weight matrix in the two-step GMM procedure by directly minimizing an information-theory-based concept of closeness between

the estimated distribution and the empirical distribution. A growing body of Monte Carlo evidence has revealed favorable finite-sample properties of the GEL/GMC estimators compared to GMM (see e.g. Ramalho, 2005, and references therein).

Recently, Newey and Smith (2004) showed analytically that while GMM and GEL share the same first-order asymptotic properties, their higher-order properties are different. Specifically, while the asymptotic bias of GMM often grows with the number of moment restrictions, the relatively smaller bias of EL does not. Moreover, after EL is bias corrected (using probabilities obtained from EL) it is higher-order efficient relative to other bias-corrected estimators.[2]

It is worth emphasizing that the GMM and GEL estimators mentioned so far are all based on *unconditional* moment restrictions burdened by the potential pitfalls described above. In addressing this problem, Kitamura, Tripathi, and Ahn (2004) (henceforth KTA) recently developed a Conditional Empirical Likelihood (CEL) estimator that makes efficient use of the information contained in (2). Their one-step estimator achieves the semiparametric efficiency bound without explicitly estimating the optimal instruments. Similar analysis has been performed by Antoine, Bonnal, and Renault (2006a) (henceforth ABR) for the case of Conditional Euclidean Likelihood[3] and Smith (2003, 2006) for the Cressie-Read family of estimators.

As the first contribution of this paper, we propose a new form of the CEL estimator for models of conditional moment restrictions (2). Our estimator, the Locally Weighted Conditional Empirical Likelihood (LWCEL), extends the one proposed by KTA. In particular, the LWCEL utilizes information about local inhomogeneities in the data that has not been previously exploited. Consequently, the new Locally Weighted CEL estimator (LWCEL) takes on a new form that differs from the currently available CEL format.

Moreover, using the GMC information-theoretic framework we show that in constructing the estimators for the *conditional* moment restrictions (2) previous literature implicitly use an arbitrary uniform weighting scheme. This leads to minimizing a discrepancy from a probability measure that is different from the one under which the data was distributed. The reason for this phenomenon is that the previously analyzed estimators for (2) are based on local kernel smoothing of the *unconditional* version of (2). In contrast, we consider an information-theoretic dual locally weighted GMC optimization problem built directly on (2) that minimizes a discrepancy from a probability measure according to which the data was distributed.

In a Monte Carlo study we show that the LWCEL estimator exhibits better finite-sample properties than found in the previous literature. However, additional complications arise in the asymptotic analysis due to a newly introduced weighting term. An extension of LWCEL to a more generic es-

[2] Accordingly, the initial focus of this paper lies in EL as opposed to any other member of the GEL family of estimators.

[3] ABR show that the Euclidean empirical likelihood estimator coincides with the continuously updated GMM (CUE-GMM) as first proposed by Hansen et al. (1996).

timation form is currently subject to our research. Assessment of analytical higher-order properties along the lines of Newey and Smith (2004) remains beyond the scope of this paper.

1.2 Conditional Moments Based on (θ_0, h_0)

A semiparametric extension of (2) to model (1) is unquestionably desirable because economic theories seldom produce exact functional forms, and misspecifications in functional forms may lead to inconsistent parameter estimates. By specifying the model partially (i.e. including h_0 as part of the unknown parameters), the inconsistency problem can be alleviated. In general, semiparametric literature related to the model (1) has been growing rapidly (see e.g. Powell, 1994; Pagan and Ullah, 1999, for reviews). Most of the available results are derived using a plug-in procedure: first h_0 is estimated nonparametrically by $\widehat{h}$ and then θ_0 is estimated using a parametric method (e.g. GMM or GEL) with h_0 replaced by $\widehat{h}$. However, such plug-in estimators are not capable of handling models where the unknown functions h_0 depend on the endogenous variables Y, because in such models θ_0 affects h_0 as well. Thus, in models where h_0 depends on an endogenous regressor, h_0 and θ_0 need to be estimated simultaneously. There are very few results concerning simultaneous estimators. Earlier applications include a semiparametric censored regression estimator (Duncan, 1986) and a semi-nonparametric maximum likelihood estimator (Gallant and Nychka, 1987).

However, a general estimation method for the model (1) that permits dependence of h_0 on Y and θ_0 was not well analyzed until a recent work by Ai and Chen (2003). These authors proposed a Sieve Minimum Distance (SMD) estimator of α_0 under (1), based on identification and consistency conditions derived by Newey and Powell (2003). Subsequent applications of the SMD estimator include Chen and Ludvigson (2006) in a habit-based asset pricing model (with unknown functional form of the habit) testing various hypotheses on stock return data, Blundell, Chen and Kristensen (2006) in a dynamic optimization model describing the allocation of total non-durable consumption expenditure, and Ai et al. (2006) investigating co-movement of commodity prices.

The first analysis that ventured into the realm of GEL-type estimators subject to conditional moment restrictions containing unknown functions is due to Otsu (2003b).[4] His shrinkage-type estimator is based on a penalized empirical log-likelihood ratio (PELR) which utilizes a penalty function $J(h)$ confining the minimization problem to a parameter space specified by the researcher. Usually, $J(h)$ is used to control some physical plausibility of h such as roughness of h. Otsu's (2003b) penalized likelihood method differs from sieve analysis and hence his treatment of asymptotics differs from ours.[5]

[4] Up to date, the author has not been able to obtain a full copy of this paper. Only a google-cached html version containing parts of the paper's text is publicly available.

[5] In the seminal paper by Shen (1997), penalized likelihood and the method of sieves are treated as two separate concepts. To achieve asymptotic normality, Otsu extends Theorem 2 of Shen (1997), whereas we extend Theorem 1 of Shen (1997) which is a separate result derived under different conditions from the former.

Otsu (2003b) suggests (in Remark 2.2) that it is also possible to use a deterministic sieve approximations, instead of the penalty function approach, resulting in a deterministic sieve empirical likelihood estimator (DSELE) that would also be, under suitable conditions, [first-order] asymptotically equivalent to the SMD of Ai and Chen (2003). Similar conjecture has been raised in Nishiyama et al. (2005) who noted the lack of theoretical justification for such procedure. Chen (2005, footnote 39) made the same type of conjecture in relation to the conditional parametric Euclidean empirical likelihood estimator of Antoine et al. (2006b). However, despite calls for a theoretical justification of such procedures, no previous paper has performed the necessary theoretical analysis. Yet, in analogy to the parametric literature described above, developing a one-step simultaneous GEL-type sieve alternative to the two-step simultaneous SMD in the semiparametric case can lead to a similar type of improvement in terms of bias and higher-order efficiency and is therefore of great theoretical and practical interest.

As the second contribution of this paper, we extend the LWCEL estimator to the semiparametric environment defined by (1). We approximate h with a sieve and estimate θ_0 and h_0 simultaneously with LWCEL. We establish consistency of the resulting one-step estimator and asymptotic normality for its parametric component of θ. Our LWCEL under (1) can be viewed as a direct alternative to the SMD estimators. A Monte Carlo study comparing small sample properties of LWCEL with SMD is planned to be included in future updates of this paper. Analytical comparison of higher-order properties remains beyond the scope of this paper.

All of the simultaneous estimators mentioned above are based on the method of sieves (Grenander, 1981; Chen, 2005) where h_0 is estimated over a compact subspace that is dense in the full parameter space as sample size increases. This feature of sieves conveniently simplifies the infinite-dimensional model h_0 to its finite-dimensional counterpart suitable for estimation. Here we also adhere to the sieve methodology. However, the currently available relevant general theory papers dealing with sieve M-estimation (Wong and Severini, 1991; Shen and Wong, 1994; Shen, 1997; Chen and Shen, 1998) consider only one set of exogenous variables without endogenous regressors and hence we can not apply these results directly in our case. Therefore, in the asymptotic analysis we combine them with several results of Ai and Chen (2003) and our own new results necessitated by the specific nature of SLWCEL under (1). In particular, among other issues we derive an extension of Shen's (1997) theorem on asymptotic normality of general simultaneous sieve estimators for the case of endogenous regressors under strong conditions and then apply it to the SLWCEL case under weak primitive conditions.

The rest of the paper is organized as follows. In Section 2 we develop the new LWCEL estimator and its dual MD counterpart for conditional moment restrictions (2) containing a finite dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ and contrast the LWCEL's finite sample properties to KTA's CEL. Section 3 extends the LWCEL to the semiparametric environment of model (1) containing both θ and a vector

of infinite dimensional unknown functions $h(\cdot)$ in $\alpha \equiv (\theta', h')'$. In Section 4 we derive consistency of the Sieve-based LWCEL $\widehat{\alpha}_n$ under a general metric. In Section 5 we show that $\widehat{\alpha}_n$ converges to α_0 at the rate $n^{-1/4}$ under the Fisher metric, which is a sufficient rate result for asymptotic normality of SLWCEL's parametric component $\widehat{\theta}_n$ derived in Section 6. Section 7 presents the results of a small-scale pilot Monte Carlo simulation study and shows favorable performance of the LWCEL estimator $\widehat{\theta}_n$ compared to KTA's CEL. Section 8 concludes. All technical proofs are presented in the Appendices.

2 The LWCEL Estimator

2.1 Existing Methods

2.1.1 Information-theoretic Approaches to Estimation

We will now develop some intuition useful for subsequent analysis by briefly introducing the heuristic background behind GMM estimation and information-theoretic alternatives such as empirical likelihood. In general terms, suppose that theory is represented by the unconditional prediction $E_Q[g(X, \theta_0)] = 0$. GMM-type estimators are defined by setting the sample moments as close as possible to the zero vector of population moments fixed by the probability measure Q.

In contrast, the information-theoretic approach focuses on a change of measure $dQ/d\Pi$ which enables $\theta \neq \theta_0$ to satisfy the transformed condition $E_\Pi[g(X, \theta)] = 0$. The estimator of θ_0 then sets the probability measure Π as close as possible to Q. Such approach thus uses closeness of probability measures, rather than moments, to estimate θ_0.

More specifically, define by $\mathcal{P}(\theta)$ the set of probability measures Π that satisfy a given condition, such as $E_\Pi[g(X, \theta)] = 0$. In order to find the most suitable Π for each $\theta \in \Theta$, the information-theoretic approach suggests the use of the convex optimization problem

$$\min_{\Pi \in \mathcal{P}(\theta)} D(\Pi, Q) \quad \text{s.t.} \quad E_\Pi[g(Z, \theta)] = 0 \tag{3}$$

where $D(\Pi, Q)$ is a measure of divergence between Π and Q,

$$D(\Pi, Q) = \int \phi\left(\frac{d\Pi}{dQ}\right) dQ \tag{4}$$

(Csiszar, 1967). For a finite sample distributed according to Q, the resulting estimator of θ_0 minimizes the finite-sample counterpart of (3) over Θ. In practice, this involves "re-weighting" the sample data to fit the given restriction. The information-theoretic approach has a long history in mathematical statistics. Its theoretical basis includes maximum entropy principle (Jaynes, 1957) and the principle of minimum discrimination information (Kullback and Leibler, 1951), (Kullback, 1997).

2.1.2 Unconditional Moment Restrictions

A substantial body of literature has been devoted to estimation under the *unconditional* moment restriction

$$E\left[g\left(X,\theta_0\right)\right]=0 \tag{5}$$

In contrast to the conditional case (2), under the unconditional framework all data is treated as exogenous which results in significant simplifications in subsequent analysis. Most notably, Qin and Lawless (1994), Hansen et al. (1996), Kitamura and Stutzer (1997), Imbens et al. (1998), Newey and Smith (2004), and Schennach (2006) belong to this category. In a comprehensive manuscript, Kitamura (2006) elaborates on the use of duality theory from convex analysis in construction of a general class of unconditional GMC estimators. This elegant framework enables one to derive a computationally friendly saddle-point GMC estimator from a dual optimization problem directly related to a primal unfeasible optimization problem that is based on an information-theoretic population specification. This approach, which we build on herein, is tantamount to a generic version of the Lagrange multiplier derivation of GEL estimators utilized in earlier literature.

2.1.3 Conditional Moment Restrictions

Estimation techniques based directly on the *conditional* moment restrictions (2) have so far been analyzed for special cases of the finite-sample conditional counterpart of the divergence measure (4): the Conditional Empirical Likelihood (CEL) with

$$\phi\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)=-\log\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)$$

by KTA, the Conditional Euclidean Likelihood with

$$\phi(x)=\frac{1}{2}\left[\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)^{2}-1\right]$$

by ABR, and the Cressie-Read parametric family with

$$\phi\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)=\frac{2}{\gamma(\gamma+1)}\left[\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)^{-\gamma}-1\right]$$

where $\gamma\in\mathbb{R}$ by Smith (2006). These estimators are all derived from local kernel smoothing based on the unconditional model (5).

2.2 Alternative Estimation Methods for Conditional Moments

The theoretical foundations of our new class of estimators extend the dual GMC approach of Kitamura (2006) to account specifically for the conditional moment restrictions. In contrast to a single GMC optimization problem utilized in Kitamura (2006) suitable for the unconditional moments (5), though, we consider a continuum of GMC optimization problems - one at each X. The resulting estimator then minimizes the expected value of the primal or dual GMC value functions, the expectation being taken with respect to the marginal distribution of the exogenous variables X.

2.2.1 Stochastic Environment

Suppose that the observations $\{(x_i, y_i) : i = 1, ..., n\}$ are drawn independently from the joint distribution $Q(x, y)$ with support $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is a compact subset of $\mathbb{R}^{d_X}$ and $\mathcal{Y}$ is a subset of $\mathbb{R}^{d_Y}$. Suppose that the unknown distribution $Q(x, y)$ satisfies the conditional moment restrictions given by (2), where $g : Z \times \Theta \to \mathbb{R}^{d_g}$ is a known mapping, up to an unknown vector of parameters $\theta_0 \in \Theta$, and $Z \equiv (Y', X_z')' \in Y \times X_Z \equiv Z \subseteq \mathbb{R}^{d_Z}$ where $X_Z \subseteq X$. The restriction (2) can then be reformulated as

$$\int g(Z, \theta_0)\, dQ(y|x) = 0$$

where $Q(y|x)$ is the "true" conditional distribution of Y given X.

Denote by $\pi(y|x)$, $q(y|x)$, $\pi(x, y)$, $q(x, y)$, $\pi(x)$, $q(x)$ the Radon-Nikodym derivatives of the probability measures $\Pi(y|x)$, $Q(y|x)$, $\Pi(x, y)$, $Q(x, y)$, $\Pi(x)$, $Q(x)$ with respect to the Lebesgue measure $m(\cdot)$, respectively.

2.2.2 Information-theoretic GMC Model

Let $\mathcal{M}_Y$ denote the set of all probability measures on $\mathbb{R}^{d_Y}$ and let

$$\mathcal{P}(X; \theta) \equiv \left\{ \Pi(y|x) \in \mathcal{M}_Y : \int g(Z, \theta)\, d\Pi(y|x) = 0;\ X \in \mathcal{X} \right\}$$

Define the set of all probability densities that are compatible with the conditional moment restriction (2) by

$$\mathcal{P}(X) \equiv \cup_{\theta \in \Theta} \mathcal{P}(X; \theta) \tag{6}$$

The set $\mathcal{P}(X)$ indexed by X represents a statistical model that is correctly specified if $q(y|x) \in \mathcal{P}(X)$. Consider the measure of conditional divergence[6]

$$D(\Pi(y|x), Q(y|x)) = \int \phi\left(\frac{d\Pi(y|x)}{dQ(y|x)}\right) dQ(y|x) \tag{7}$$

[6] This conditional measure of divergence is a natural extension of the conditional discrepancy measure formulated by Shannon (1948) for the special case of conditional entropy with $\phi(x) = x \log(x)$.

where, ϕ is a convex function and $\Pi(y|x)$ is absolutely continuous with respect to $Q(y|x)$.

For given $\theta \in \Theta$ and $X \in \mathcal{X}$, define the conditional contrast function $\rho(\theta, dQ(y|x))$ as the infimum of the discrepancy (7) between $\Pi(y|x)$ and $Q(y|x)$

$$\rho(\theta, dQ(y|x)) \equiv \inf_{\Pi(y|x)\in\mathcal{P}(X)} D(\Pi(y|x), Q(y|x)) \tag{8}$$

Assuming model identification conditions are satisfied, for $\theta \neq \theta_0$, $\pi(y|x) \neq q(y|x)$ a.s. Since by definition $D(\cdot, Q(y|x))$ attains its minimum at $Q(y|x)$, it follows from (8) and (6) that the true population parameter value θ_0 uniquely solves the population GMC optimization problem

$$\theta_0 = \arg\inf_{\theta\in\Theta} E_{Q(x)}[\rho(\theta, dQ(y|x))|X] \tag{9}$$

Taking the expectation with respect to the probability measure $Q(x)$ in (9) according to which the exogenous X were distributed is the key to our formulation of the population GMC optimization problem. As Lemma 1 in the Appendix shows, under this specification the expectation of the conditional contrast function with respect to $Q(x)$ minimizes the divergence between the two *joint* distributions $\Pi(x, y)$ and $Q(x, y)$.

2.2.3 Dual Formulation

To facilitate a computationally feasible estimator of θ_0, it is beneficial to express the GMC optimization problem (9) in terms of the arguments θ and X only, stating explicitly the constraints. Define $p(y|x) = \frac{d\Pi(y|x)}{dQ(y|x)} \in \mathbb{R}_+$ and $p(x, y) = \frac{d\Pi(x,y)}{dQ(x,y)} \in \mathbb{R}_+$. For a given $\theta \in \Theta$ and $X \in \mathcal{X}$, using $p(y|x)$ in (7), the conditional contrast function (8) can be expressed as a value function

$$v(\theta, X) = \inf_{p(y|x)\in\mathbb{R}_+} \int \phi(p(y|x))\, dQ(y|x) \;\text{ s.t. }\; \int g(Z, \theta)\, p(y|x) dQ(y|x) = 0, \;\int p(y|x) dQ(y|x) = 1 \tag{10}$$

Using results from convex analysis (see e.g. Luenberger, 1969; Borwein and Lewis, 2006), the numerically unfeasible *primal problem* (10) has an equivalent expression as a *dual problem*

$$v^*(X; \theta) = \max_{\lambda(X)\in\mathbb{R}^{d_g}, \mu(X)\in\mathbb{R}} \left[\mu(X) - \int \phi^*(\mu(X) + \lambda(X)' g(Z, \theta))\, dQ(y|x)\right]$$

where $\phi^*(\cdot)$ is the convex conjugate (or Legendre transformation) of $\phi(\cdot)$. This is a finite-dimensional unconstrained convex maximization problem that will further provide the basis for numerical optimization. By Fenchel duality,

$$v(X; \theta) = v^*(X; \theta) \tag{11}$$

It is beneficial for the construction of the estimator in the next section to express the value-function formulation (10) of the GMC optimization problem (9) in terms of the Lebesgue measure

$$E_{Q(x)}[v(X; \theta)] = \int q(x, y)\phi\left(\frac{\pi(y|x)}{q(y|x)}\right) dm(x, y) \;\text{ s.t. }\; \int \pi(y|x) g(z, \theta)\, dm(y|x) = 0, \;\int \pi(y|x) dm(y|x) = 1 \tag{12}$$

Using (11), (12) is equivalent to

$$E_{Q(x)}\left[v^*(X;\theta)\right] = \max_{\lambda\in\mathbb{R}^{d_g},\mu\in\mathbb{R}}\left[\int q(x)\mu(X)dm(x) - \int q(x,y)\phi^*\left(\mu(X)+\lambda(X)'g\left(Z,\theta\right)\right)dm(x,y)\right] \tag{13}$$

A feasible estimator formulated in the next section minimizes the unconstrained finite-dimensional optimization problem (13) over the parameter space Θ.

2.2.4 The Estimator

Given a sample $\{(x_i,y_i) : i = 1,...,n\}$ distributed according to $Q(x,y)$, the population criteria described above provide a basis for statistical inference wherein we replace the unknown probability measures $Q(x,y)$ and $Q(y|x)$ with their empirical counterparts $Q(x_i,y_j)$ and $Q(y_j|x_i)$, respectively. The densities $q(x,y)$ and $q(y|x)$ need to be estimated nonparametrically as probability mass functions $q(x_i,y_j)$ and $q(y_j|x_i)$ with support on the data. Numerous methods have been suggested in the literature to obtain such estimates with various desirable properties using e.g. kernels, series or nearest neighbors to name just a few (see e.g. Pagan and Ullah, 1999, and references therein).

A sample version of (12) is

$$\begin{aligned}\widehat{v}(\theta) &= \widehat{E}_{Q(x)}\left[v(X;\theta)\right] \\ &= \left\{\sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j)\phi\left(\frac{\pi(y_j|x_i)}{q(y_j|x_i)}\right) : \sum_{j=1}^{n}\pi(y_j|x_i)g\left(z_j,\theta\right) = 0, \sum_{j=1}^{n}\pi(y_j|x_i) = 1\right\}\end{aligned} \tag{14}$$

and of its dual formulation (13)

$$\begin{aligned}\widehat{v}^*(\theta) &= \widehat{E}_{Q(x)}\left[v^*(X;\theta)\right] \\ &= \max_{\lambda\in\mathbb{R}^{d_g},\mu\in\mathbb{R}}\left[\sum_{i=1}^{n} q(x_i)\mu(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j)\phi^*\left(\mu(x_i)+\lambda(x_i)'g\left(z_j,\theta\right)\right)\right]\end{aligned} \tag{15}$$

This leads to the Locally Weighted Conditional GMC estimator for θ

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\widehat{v}(\theta) \tag{16}$$

This estimator corresponds to the *conditional* locally weighted forms of the "Minimum Discrepancy Statistic" of Corcoran (1998) and the "Minimum Distance Estimator" of Newey and Smith (2004). Its computationally convenient dual formulation based on (15) is expressed as

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\widehat{v}^*(\theta) \tag{17}$$

2.2.5 Localization Features

For a sample $\{(y_i, x_i) : i = 1, ..., n\}$ estimation of $q(y|x)$ and $q(x, y)$ amounts to the use of localization methods (Tibshirani and Hastie, 1987). In the stream of literature most relevant to this paper, localization schemes have been used in the conditional moment context in LeBlanc and Crowley (1995), Zhang and Gijbels (2003), KTA for CEL, ABR for the EuL, and Smith (2003, 2005) for GEL. Information on $Q(y|x)$ is inferred from the nearby observations if we assume that $Q(y|x)$ is continuous with respect to X. In other words, in a neighborhood around x_i we approximate $Q(y|x)$ by $Q(y|x) \approx Q(y|x_i)$. This implies that all the z_j with x_j lying in this neighborhood can be roughly viewed as observations from $Q(y|x_i)$. Note that, unlike in the unconditional moment case (5) where $q(x_i) = 1/n$, now the $q(x_i, y_j)$ and $q(y_j|x_i)$ are not derived directly from observed data, since only one realization of the random vector y_j was actually observed at x_i. Rather, these probability masses are inferred from neighboring observations. The data-determined $q(x_i, y_j)$ and $q(y_j|x_i)$ are then used as a benchmark in the value function of the GMC optimization problem in derivations of $\widehat{\theta}$.

2.3 Locally Weighted Conditional Empirical Likelihood

Various choices for the discrepancy measure $\phi(\cdot)$ lead to various special cases of the Dual Locally Weighted Conditional GMC estimator. Setting $\phi(x) = -\log(x)$ corresponds to Locally Weighted Conditional Empirical Likelihood (LWCEL). The unfeasible GMC estimator of (9) becomes

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \widehat{v}(\theta) \equiv \left\{ -\sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i, y_j) \log\left(\frac{\pi(y_j|x_i)}{q(y_j|x_i)}\right) : \sum_{j=1}^{n} \pi(y_j|x_i) g(z_j, \theta) = 0, \sum_{j=1}^{n} \pi(y_j|x_i) = 1 \right\} \tag{18}$$

The convex conjugate of $\phi(x) = -\log(x)$ is $\phi^*(y) = -1 - \log(-y)$. Using this expression in the feasible dual formulation (17) we obtain

$$\widehat{\theta}_{LWCEL} = \arg\min_{\theta\in\Theta} \widehat{v}^*(\theta) \equiv \max_{\lambda\in\mathbb{R}^{d_g}, \mu\in\mathbb{R}} \left[\sum_{i=1}^{n} q(x_i)\mu(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i, y_j) \log\left(-\mu(x_i) - \lambda(x_i)' g(z_j, \theta)\right) \right]$$

It is worth noting that on the population level, the LWCEL minimizes the discrepancy measure

$$\begin{aligned} D\left(\Pi(x, y), Q(x, y)\right) &= \int \log\left(\frac{dQ(x, y)}{d\Pi(x, y)}\right) dQ(x, y) \\ &= K(Q(x, y), \Pi(x, y)) \end{aligned}$$

where $K(Q(x, y), \Pi(x, y))$ is the Kullback-Leibler (KL) divergence between the *joint* probability measures $Q(x, y)$ and $\Pi(x, y)$ with $Q(x, y)$ being the true probability measure according to which

the data are distributed. The $\widehat{\theta}_{LWCEL}$ then solves the minimization problem

$$\inf_{\theta\in\Theta}\inf_{\boldsymbol{\pi}(x,y):\pi(x,y)\in\{\mathbf{M}_Y:X\in\mathcal{X}\}} K(Q_n(x,y),\Pi(x,y))$$

where $Q_n(x,y)$ is the empirical measure and $\Pi(x,y)$ represents the moment conditions model.

Note that this estimator contains two important modifications in comparison to the Conditional Empirical Likelihood (CEL) analyzed by KTA specified in our notation as

$$\widehat{\theta}_{CEL} = \arg\min_{\theta\in\Theta}\max_{\lambda\in\mathbb{R}^{d_g}}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} q(y_j|x_i)\log\left(1+\lambda(x_i)'g\left(z_j,\theta\right)\right)\right]$$

First, the weight of the logarithmic function in $\widehat{\theta}_{CEL}$ is $q(y_j|x_i)$ as opposed to $q(x_i,y_j)$ in $\widehat{\theta}_{LWCEL}$. This is a consequence by taking simple summation of the local discrepancies at x_i in derivation of $\widehat{\theta}_{CEL}$ as opposed to a weighted sum that would capture the relative importance of each local discrepancy in the global objective function. Thus, in the population version of the GMC optimization problem with $E_{m(X)}\left[v(X;\theta)\right]$ the $\widehat{\theta}_{CEL}$ minimizes $D\left(\Pi(y|x),U(X)Q(y|x)\right)$ as opposed to $D\left(\Pi(x,y),Q(x,y)\right)$ for $\widehat{\theta}_{LWCEL}$, where $U(x)$ is the uniform probability measure over X. However, $Q(x,y)\neq U(x)q(y|x)$, almost surely. Second, $\widehat{\theta}_{CEL}$ sets $\mu(x_i)=1$ which is an artefact of using a specific kernel estimation method where individual weights sum up to 1. In general, however, $\mu(x_i)\neq 1$ a.s.

A closer look on the structure of the optimization problem behind $\widehat{\theta}_{LWCEL}$ reveals interesting comparisons with the form of empirical likelihood established in the literature for unconditional moment restrictions. Taking first-order conditions of the GMC Lagrangian

$$\begin{aligned} L(\theta,\lambda,\mu,\pi) \quad = \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j)\ln\left(\frac{\pi(y_j|x_i)}{q(y_j|x_i)}\right) - \sum_{i=1}^{n}\lambda(x_i)'\sum_{j=1}^{n}\pi(y_j|x_i)g\left(z_j,\theta\right) \\ & -\sum_{i=1}^{n}\mu(x_i)\left(\sum_{j=1}^{n}\pi(y_j|x_i)-1\right) \end{aligned} \tag{19}$$

corresponding to the GMC objective function (18) yields

$$\frac{\widehat{q}(x_i,y_j)}{\widehat{\pi}(y_j|x_i)} = \widehat{\lambda}(x_i)'g\left(z_j,\widehat{\theta}\right)+\widehat{\mu}_i \quad , \ \forall i,j \tag{20}$$

$$\sum_{j=1}^{n}\widehat{\pi}(y_j|x_i)g\left(z_j,\widehat{\theta}\right)=0 \quad , \ \forall i \tag{21}$$

$$\sum_{j=1}^{n}\widehat{\pi}(y_j|x_i)=1 \tag{22}$$

Summing (20) over j and using (21) yields, for each i,

$$\begin{aligned}\sigma(x_i) &\equiv \sum_{j=1}^{n} \widehat{q}(x_i, y_j) \\ &= \widehat{\lambda}(x_i)' \sum_{j=1}^{n} \widehat{\pi}(y_j|x_i) g\left(z_j, \widehat{\theta}\right) + \widehat{\mu}(x_i) \sum_{j=1}^{n} \widehat{\pi}_{ij} \\ &= \widehat{\mu}(x_i) \end{aligned} \tag{23}$$

Substituting (23) into (20) gives, for each i and j,

$$\widehat{\pi}(y_j|x_i) = \frac{\widehat{q}(x_i, y_j)}{\sigma(x_i) + \widehat{\lambda}(x_i)' g\left(z_j, \widehat{\theta}\right)} \tag{24}$$

Substituting (24) into the Lagrangian (19), and using (21) and (22), yields

$$L(\theta, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \ln \left(\frac{\widehat{q}(x_i)}{\sigma(x_i) + \widehat{\lambda}(x_i)' g\left(z_j, \widehat{\theta}\right)} \right) \tag{25}$$

Then the Locally Weighted Conditional Empirical Likelihood estimator with the new weighting scheme is defined as

$$\widehat{\theta}_{LWCEL} = \arg \max_{\theta \in \Theta} L(\theta, \lambda_i) \tag{26}$$

where $\widehat{\lambda}_i$ solves[7]

$$\sum_{j=1}^{n} \frac{\widehat{q}(x_i, y_j) g\left(z_j, \widehat{\theta}\right)}{\sigma_i + \widehat{\lambda}_i' g\left(z_j, \widehat{\theta}\right)} = 0$$

obtained from (21) and (24). As discussed above, in general $\sigma_i \neq 1$. The presence of σ_i is the hallmark of LWCEL compared to the previous literature where, invariably, $\sigma_i = 1$.

The $\widehat{\theta}_{LWCEL}$ estimator defined in (26) is a special case of a corresponding estimator derived under semiparametric conditional moment restrictions in the next Chapter. For this reason, we will perform the asymptotic analysis pertaining to both estimators in the next chapter. The MD estimator analyzed by Smith (2003, 2005) as well as the CEL estimator elaborated in KTA achieve the semiparametric efficiency lower bound (see Chamberlain, 1987). The weighting introduced for $\widehat{\theta}_{LWCEL}$ in this paper postulates more flexible weights that improve on the fixed-bandwidth kernel weights in finite samples in terms of MSE. We conclude that our new forms of the MD and CEL estimators exhibit first-order asymptotic equivalence in terms of consistency and asymptotic normality with the ones formulated in the previous literature, and hence also achieve the first-order asymptotic semiparametric efficiency lower bound. However, our $\widehat{\theta}_{LWCEL}$ improves on its previously analyzed

[7] In line with KTA we adopt the notation $\widehat{\lambda}_i$ as shorthand for $\widehat{\lambda}(x_i, \widehat{\theta})$. In the same spirit, we denote $\sigma(x_i)$ with σ_i in the sequel. When necessary, we explicitly write the full form to ensure that our arguments are unambiguous.

forms in terms of finite sample performance.

Given the general GMC setup above, the extension of the estimation procedure from LWCEL to a more generic functional form of ϕ appears relatively straightforward and is currently subject to our research.

3 Semiparametric Conditional Moment Restrictions

In this Section we extend the LWCEL estimator (25) to the semiparametric environment defined by (1). In doing so, we will use series estimation (see e.g. Newey, 1997) as a particular form of linear sieves in both approximating h and determining the weights w_{ij}. Series estimators are known to contain functional bases that are superior in terms of MSE criteria to fixed-bandwidth kernel estimators, especially in the presence of spatial inhomogeneities in the data (see e.g. Ramsey, 1999). Silverman (1984) showed that series estimators with spline basis functions behave approximately like the variable-bandwidth kernel estimator which improves on its fixed-bandwidth version in terms of MSE by the virtue of local adaptation. Another advantage of working with the LWCEL estimator based on series approximation is that truncation arguments in regions with small data density are not required in contrast to kernel weights.

3.1 Sieve-based Conditional Empirical Likelihood

The environment setup parallels the one of Newey and Powell (2003) and Ai and Chen (2003). Suppose that the observations $\{(Y_i, X_i) : i = 1, ..., n\}$ are drawn independently from the distribution of (Y, X) with support $\mathcal{Y} \times \mathcal{X}$, where $\mathcal{Y}$ is a subset of $\mathbb{R}^{d_Y}$ and $\mathcal{X}$ is a compact subset of $\mathbb{R}^{d_X}$. Suppose that the unknown distribution of (Y, X) satisfies the semiparametric conditional moment restrictions given by (1), where $g : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}^{d_g}$ is a known mapping, up to an unknown vector of parameters, $\alpha_0 \equiv (\theta_0', h_0')' \in \mathcal{A} \equiv \Theta \times \mathcal{H}$, and $Z \equiv (Y', X_z')' \in \mathcal{Y} \times \mathcal{X}_Z \equiv \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ where $\mathcal{X}_Z \subseteq \mathcal{X}$. We assume that $\Theta \subseteq \mathbb{R}^{d_\theta}$ is compact with non-empty interior and that $\mathcal{H} \equiv \mathcal{H}^1 \times ... \times \mathcal{H}^{d_h}$ is a space of continuous functions. Since $\mathcal{H}$ is infinite-dimensional, in constructing a feasible estimator we follow the sieve literature (Grenander, 1981; Chen, 2005) by replacing $\mathcal{H}$ with a sieve space $\mathcal{H}_n \equiv \mathcal{H}_n^1 \times ... \times \mathcal{H}_n^{d_h}$ which is a computable and finite-dimensional compact parameter space that becomes dense in $\mathcal{H}$ as n increases.

Next, we introduce the series estimator used in the analysis (see Newey, 1997; Ai and Chen, 2003). For each $l = 1, ..., d_g$, and for a given α, let $\{p_{0j}(X),\ j = 1, 2, ...k_n\}$ denote a sequence of known basis functions (power series, splines, wavelets, etc.) and let $p^{k_n}(X) \equiv (p_{01}(X), ..., p_{0k_n}(X))'$. Let further $p^{k_n}(X)$ be a tensor-product linear sieve basis, which is a product of univariate sieves over d_X (see Ai and Chen, 2003, for details). Let $P = (p^{k_n}(x_1), ..., p^{k_n}(x_n))'$ be an $(n \times k_n)$ matrix. Consider the model (1) and denote the conditional mean function

$$\begin{aligned} m(X, \alpha) &\equiv E\left[g\left(Z, \alpha\right)|X\right] \\ &= \int g\left(Z, \alpha\right) dF_{Y|X} \end{aligned} \tag{27}$$

Let $\widehat{m}(X, \alpha) \equiv (\widehat{m}_1(X, \alpha), ..., \widehat{m}_{d_\rho}(X, \alpha))'$. A consistent nonparametric linear sieve estimator of

$m_l(X,\alpha)$ is given by

$$\widehat{m}_l(X,\alpha) = p^{k_n}(X)'\widehat{\kappa}_l$$

where h in $\alpha = (\theta', h')'$ is restricted to the sieve space $\mathcal{H}_n$ and $\widehat{\kappa}_l$ is an OLS estimate obtained by regressing $g_l(Y, X_z, \alpha)$ on $p^{k_n}(X)$,

$$\begin{aligned}\widehat{\kappa}_l &= (P'P)^{-1} P' g_l(Z,\alpha) \\ &= \sum_{j=1}^{n} p^{k_n}(x_j)' (P'P)^{-1} g_l(z_j,\alpha) \end{aligned} \tag{28}$$

and hence

$$\begin{aligned}\widehat{m}_l(x_i,\alpha) &= \widehat{E}_{Z|X}\left[g_l(Z,\alpha)\,|X = x_i\right] \\ &= p^{k_n}(x_i)'\widehat{\kappa}_l \\ &= \sum_{j=1}^{n} p^{k_n}(x_j)' (P'P)^{-1} p^{k_n}(x_i) g_l(z_j,\alpha) \\ &= \sum_{j=1}^{n} w_{ij} g_l(z_j,\alpha)\end{aligned}$$

after substituting from (28), $l = \{1, ..., d_g\}$. In the vector form

$$\widehat{m}(x_i,\alpha) = \sum_{j=1}^{n} w_{ij} g(z_j,\alpha)$$

The weights are given by

$$w_{ij} = p^{k_n}(x_j)' (P'P)^{-1} p^{k_n}(x_i) \tag{29}$$

and

$$\begin{aligned}\sigma_i &= \sum_{j=1}^{n} w_{ij} \\ &= \sum_{j=1}^{n} p^{k_n}(x_j)' (P'P)^{-1} p^{k_n}(x_i) \\ &= \mathbf{i}' P (P'P)^{-1} p^{k_n}(x_i)\end{aligned}$$

where $\mathbf{i}$ is a $(n \times 1)-$vector of ones.

We now turn to the derivation of LWCEL under (1). The Lagrangian[8] for the local semipara-

[8] As discussed above, omission of q_{ij} from the denominator of $\ln(\pi_{ij}/q_{ij})$ is inconsequential in the case of LWCEL.

metric EL estimator is

$$\max_{p_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \ln \pi_{ij} \quad \text{s.t.} \quad \pi_{ij} \geq 0, \ \sum_{j=1}^{n} \pi_{ij} = 1, \ \sum_{j=1}^{n} g\left(z_j, \alpha_n\right) \pi_{ij} = 0, \ \text{for } i, j = 1, ..., n$$

where α_n is α restricted to the sieve space $\mathcal{A}_n$. Then,

$$\widehat{\pi}_{ij} = \frac{w_{ij}}{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)} \tag{30}$$

and for each $\alpha_n \in \mathcal{A}_n$, λ_i solves

$$\sum_{j=1}^{n} \frac{w_{ij} g\left(z_j, \alpha_n\right)}{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)} = 0 \tag{31}$$

The Sieve-based Locally Weighted Conditional Empirical Likelihood (SLWCEL) evaluated at α_n is defined as

$$L_{SLWCEL}(\alpha_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \ln \left\{ \frac{w_{ij}}{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)} \right\}$$

where λ_i solves (31). The estimator of α_0 is defined as

$$\widehat{\alpha}_n = \arg \max_{\alpha_n \in \mathcal{A}_n} L_{SLWCEL}(\alpha_n) \tag{32}$$

Solving (32) is equivalent to solving

$$\widehat{\alpha}_n = \arg \max_{\alpha_n \in \mathcal{A}_n} G_n(\alpha_n) \tag{33}$$

where

$$G_n(\alpha_n) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \ln \left\{ \sigma_i + \lambda_i' g\left(z_j, \alpha_n\right) \right\} \tag{34}$$

Implementing our estimator is straightforward. One advantage of the sieve approach is that once $h \in \mathcal{H}$ is replaced by $h_n \in \mathcal{H}_n$, the estimation problem effectively becomes a parametric one. Commonly used statistical and econometric packages can then be used to compute the estimate. From (31) it follows that

$$\lambda_i = \arg \max_{\rho \in \mathbb{R}^{d_g}} \sum_{j=1}^{n} w_{ij} \ln \left\{ \sigma_i + \rho' g\left(z_j, \alpha_n\right) \right\} \tag{35}$$

This is a well-behaved optimization problem since the objective function is globally concave and can be solved by a Newton-Raphson numerical procedure. The outer loop (33) can be carried out using a numerical optimization procedure. For a relevant discussion of computational issues, see for example Kitamura (2006, section 8.1).

4 Consistency

In this section we present some asymptotic results for the smoothed empirical likelihood estimator as defined in (32). The general approach follows closely the one developed in KTA. The following definitions, adopted from Ai and Chen (2003), are introduced:

Definition 4.1 *A real-valued measurable function $g(Z,\alpha)$ is* Hölder continuous *in $\alpha \in \mathcal{A}$ if there exist a constant $\overline{\kappa} \in (0,1]$ and a measurable function $c_2(Z)$ with $E\left[c_2(Z)^2|X\right]$ bounded, such that $|g(Z,\alpha_1) - g(Z,\alpha_2)| \leq c_2(Z)\left\|\alpha_1 - \alpha_2\right\|^{\overline{\kappa}}$ for all $Z \in \mathcal{Z}$, $\alpha_1, \alpha_2 \in \mathcal{A}$.*

The Hölder space of smooth functions $\Lambda^{\overline{\gamma}}(\mathcal{X})$ of order $\overline{\gamma} > 0$ and the corresponding Hölder ball $\Lambda_c^{\overline{\gamma}}(\mathcal{X}) \equiv \{g \in \Lambda^{\overline{\gamma}}(\mathcal{X}) : \|g\|_{\Lambda^{\overline{\gamma}}} \leq c < \infty\}$ with radius c are defined in Ai and Chen (2003), p. 1800.

Definition 4.2 *A real-valued measurable function $g(Z,\alpha)$* satisfies an envelope condition *over $\alpha \in \mathcal{A}$ if there exists a measurable function $c_1(Z)$ with $E\left\{c_1(Z)^4\right\} < \infty$ such that $|g(Z,\alpha)| \leq c_1(Z)$ for all $Z \in \mathcal{Z}$ and $\alpha \in \mathcal{A}$.*

The following Assumptions are made to facilitate the analysis:

Assumption 4.1 *For each $\alpha \neq \alpha_0$ there exists a set $\mathcal{X}_\alpha$ such that $\Pr\{x \in \mathcal{X}_\alpha\}$, and $E\left[g\left(z,\alpha\right)|x\right] \neq 0$ for every $x \in \mathcal{X}_\alpha$.*

Assumption 4.2 *(i) The data $\{(Y_i, X_i)_{i=1}^n\}$ are i.i.d.; (ii) $\mathcal{X}$ is compact with nonempty interior; (iii) the density of X is bounded and bounded away from zero.*

Assumption 4.3 *(i) The smallest and the largest eigenvalues of $E\left[p^{k_n}(X) \times p^{k_n}(X)'\right]$ are bounded and bounded away from zero for all k_n; (ii) for any $g(\cdot)$ with $E\left[g(X)^2\right] < \infty$, there exists $p^{k_n}(X)'\kappa$ such that $E\left[\left\{g(X) - p^{k_n}(X)'\kappa\right\}^2\right] = o(1)$.*

Assumption 4.4 *(i) There is a metric $\|\cdot\|$ such that $\mathcal{A} \equiv \Theta \times \mathcal{H}$ is compact under $\|\cdot\|$; (ii) for any $\alpha \in \mathcal{A}$, there exists $\Pi_n\alpha \in \mathcal{A}_n \equiv \Theta \times \mathcal{H}_n$ such that $\|\Pi_n\alpha - \alpha\| = o(1)$.*

Assumption 4.5 *(i) $E\left[\left|g\left(Z,\alpha_0\right)\right|^2|X\right]$ is bounded; (ii) $g\left(Z,\alpha\right)$ is Hölder continuous in $\alpha \in \mathcal{A}$.*

Let $k_{1n} \equiv \dim(\mathcal{H}_n)$ denote the number of unknown sieve parameters in $h_n \in \mathcal{H}_n$.

Assumption 4.6 *$k_{1n} \to \infty$, $k_n \to \infty$, $k_n/n \to 0$ and $d_g k_n \geq d_\theta + k_{1n}$.*

Assumption 4.7 *$E\left\|x\right\|^{1+\varrho} < \infty$ for some $\varrho < \infty$.*

Assumption 4.8 *$E\left\{\sup_{\alpha\in\mathcal{A}}\left\|g\left(Z,\alpha\right)\right\|^m\right\} < \infty$ for some $m \geq 8$.*

Assumption 4.1 is Assumption 3.1 in KTA that guarantees identification of θ_0. Assumptions 4.2–4.6 are essentially the same conditions imposed in Newey and Powell (2003) and Ai and Chen (2003). Assumption 4.2 rules out time series observations. Assumptions 4.3–4.6 are typical conditions imposed for series (or linear sieve) estimation of conditional mean functions. Assumption 4.4(i) restricts the parameter space as well as the choice of the metric $\|\cdot\|$. It is a commonly imposed condition in the semiparametric econometrics literature, and is satisfied when the infinite-dimensional parameter space $\mathcal{H}$ consists of bounded and smooth functions (see Gallant and Nychka, 1987). Assumption 4.4(ii) is the definition of a sieve space. Assumption 4.5 is typically imposed on the residual function in the literature on parametric nonlinear estimation. Assumption 4.6 restricts the growth rate of the number of basis functions in the series approximation. Assumption 4.7 is Assumption 3.4(ii) in KTA, used in Lemma A.1. Assumption 4.8 is Assumption 3.2 in KTA used in Lemma A8.

The following Theorem provides a consistency result:

Theorem 4.1 *Let the Assumptions 4.1–4.7 hold. Then* $\|\widehat{\alpha}_n - \alpha_0\| = o_p(1)$.

The proof is derived in the Appendix. The proof proceeds along the lines of KTA. However, the fact that the sieve parameter space $\mathcal{H}_n$ grows dense in an infinite-dimensional space $\mathcal{H}$ now needs to be addressed. The inclusion of σ_i in the LWCEL objective function compared to KTA's CEL also complicates matters. We achieve some simplifications arising from not having to make use of truncation arguments for kernels. Since we are not dealing with kernels, unlike KTA we can not use Lemma B.1 of Ai (1997) to determine uniform convergence rates. For this purpose, we specialize Lemma A.1(A) of Ai and Chen (2003), derived for the combined space $\mathcal{X} \times \mathcal{A}$, to the space $\mathcal{X}$ only, with $g(z_j, \alpha)$ evaluated at a given fixed α. Since we do not have to account for growth restrictions on the parameter space in this Lemma, we are able to obtain faster convergence rate $\widetilde{\delta}_{1n}$ than Ai and Chen (2003).

5 Convergence Rates

Theorem 4.1 established consistency of $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$ under a general metric $\|\cdot\|$ constrained only by Assumption 4.4(i). In order to ascertain asymptotic normality of $\widehat{\theta}_n$, one typically needs that $\widehat{\alpha}_n$ converge to α_0 at a rate faster than $n^{-1/4}$ (see e.g. Newey, 1994). As noted by Newey and Powell (2003), for model (1) where the unknown h_0 can depend on endogenous variables Y, it is generally difficult to obtain fast convergence rate under $\|\cdot\|$. Nonetheless, as demonstrated by Ai and Chen (2003), in simultaneous estimation of $(\widehat{\theta}_n, \widehat{h}_n)$ it is sufficient to show fast convergence rate of $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$ for only a special case of $\|\cdot\|$ to derive asymptotic normality of $\widehat{\theta}_n$. Naturally, we will also follow this approach. However, since the objective function of the problem analyzed in Ai and Chen (2003) is different from ours, our metric also differs. While Ai and Chen (2003) used

a quadratic form type metric, we perform the analysis under the Fisher metric $\|\cdot\|_F$ which is the natural choice for a likelihood-based scenario.

Some additional notation is necessary to introduce the Fisher metric. The properties of $\mathcal{A}$ and the notation for pathwise derivatives established in this paragraph borrows from Ai and Chen (2003). Suppose the parameter space $\mathcal{A}$ is connected in the sense that for any two points α_1, $\alpha_2 \in \mathcal{A}$ there exists a continuous path $\{\alpha(t) : t \in [0,1]\}$ in $\mathcal{A}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. Also, suppose that $\mathcal{A}$ is convex at the true value α_0 in the sense that, for any $\alpha \in \mathcal{A}$, $(1-t)\alpha_0 + t\alpha \in$ for small $t > 0$. Furthermore, suppose that for almost all Z, $g(Z, (1-t)\alpha_0 + t\alpha)$ is continuously differentiable at $t = 0$. Denote the first pathwise derivative at the direction $[\alpha - \alpha_0]$ evaluated at α_0 by

$$\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \left.\frac{dg(Z,(1-t)\alpha_0 + t\alpha)}{dt}\right|_{t=0} \quad \text{a.s. } Z$$

and for any α_1, $\alpha_2 \in \mathcal{A}$ denote

$$\begin{aligned} \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] &\equiv \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_0] - \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_2 - \alpha_0] \\ \frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] &\equiv E\left\{ \left. \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \right| X \right\} \end{aligned} \tag{36}$$

Furthermore, let

$$\varphi(X, Z, \alpha) \equiv \ln\left\{\sigma_x + \lambda'(X,\alpha) g(Z,\alpha)\right\} \tag{37}$$

$$\psi(X,\alpha) \equiv E\left[\varphi(X,Z,\alpha) \,|X\right] \tag{38}$$

where σ_x stands for σ_i evaluated at a generic $X = x$. For any α_1, $\alpha_2 \in \mathcal{A}$ the Fisher norm $\|\cdot\|_F$ (see e.g. Wong and Severini, 1991, p. 607) is defined[9] as

$$\|\alpha_1 - \alpha_2\|_F = \sqrt{E\left\{ E\left[\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \right)' \frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \middle| X \right] \right\}} \tag{39}$$

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A} - \{\alpha_0\}$ under the metric $\|\cdot\|_F$. Then $\left(\overline{\mathbf{V}}, \|\cdot\|_F\right)$ is a Hilbert space with the inner product

$$\langle v_1, v_2 \rangle_F = \|v_1 - v_2\|_F^2$$

We will now show that our metric $\|\alpha_1 - \alpha_2\|_F$ is equivalent to a *conditional version* of the metric

[9] We use the inner product notation for the pathwise derivatives to explicitly account for the special case when $\alpha \equiv \theta \in \mathbb{R}^{d_\theta}$.

used in Ai and Chen (2003). Let

$$\begin{aligned} s(X,Z,\alpha) &\equiv \lambda'(\alpha,X)g\left(Z,\alpha\right) \\ \varpi\left(X,Z,\alpha\right) &\equiv \frac{d\varphi\left(X,Z,\alpha_0\right)}{ds(X,Z,\alpha)} \\ &= \frac{1}{\sigma_x + s(X,Z,\alpha)} \end{aligned}$$

where $s(X,Z,\alpha)$ and $\varpi\left(X,Z,\alpha\right)$ is scalars. Note that from the conditional moment restriction (1), under the expectation taken over Z conditional on X

$$\lambda(X,\alpha_0) = 0 \tag{40}$$

which means that the constraints on $F_{Y|X}$ imposed by (1) are satisfied with equality and the Lagrange multiplier $\lambda(X,\alpha_0)$ takes on the value 0. This is also apparent from Lemma A.8. We have

$$\begin{aligned} & E\left[\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[\alpha_1-\alpha_2]\right)'\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\ = & E\left[\varpi\left(X,Z,\alpha_0\right)^2\left(\frac{ds(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'\frac{ds(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\ = & E\left[\begin{array}{c}\varpi\left(X,Z,\alpha_0\right)^2\left(\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]+g(Z,\alpha_0)\frac{d\lambda'(X,\alpha_0)}{d\alpha}\right)' \\ \times\left(\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]+g(Z,\alpha_0)\frac{d\lambda'(X,\alpha_0)}{d\alpha}\right)\end{array}\middle| X\right] \\ = & A_1+A_2+A_3+A_4 \end{aligned} \tag{41}$$

where

$$\begin{aligned} A_1 &= E\left[\varpi\left(X,Z,\alpha_0\right)^2\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'\lambda(X,\alpha_0)\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\ A_2 &= E\left[\varpi\left(X,Z,\alpha_0\right)^2\left(\frac{d\lambda(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'g(Z,\alpha_0)\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\ A_3 &= E\left[\varpi\left(X,Z,\alpha_0\right)^2\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'\lambda'(X,\alpha_0)g(Z,\alpha_0)\frac{d\lambda'(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\ A_4 &= E\left[\varpi\left(X,Z,\alpha_0\right)^2\left(\frac{d\lambda(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'g(Z,\alpha_0)g'(Z,\alpha_0)\frac{d\lambda'(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \end{aligned} \tag{42}$$

Using (40) yields $A_1 = A_2 = A_3 = 0$. By the definition of $\lambda(X,\alpha)$ in (35), $\lambda(X,\alpha)$ is a function of $g(Z,\alpha)$ which is a function of α. Moreover, $\lambda(X,\alpha)$ is a function of α *only* via $g(Z,\alpha)$. Hence, under the expectation taken over Z conditional on X

$$\frac{d\lambda(X,\alpha)}{d\alpha}[\alpha_1-\alpha_2] = \frac{d\lambda(X,\alpha)}{dg(Z,\alpha)}\frac{dg(Z,\alpha)}{d\alpha}[\alpha_1-\alpha_2] \tag{43}$$

In particular, under the expectation over Z conditional on X, $\lambda(X,\alpha)$ is defined implicitly as a function of $g(Z,\alpha)$ by the relation

$$F(\lambda,g) = E\left[\frac{g(Z,\alpha)}{\sigma_x+\lambda'(X,\alpha)g(Z,\alpha)}\middle| X\right] = 0$$

By the Implicit Function Theorem

$$\begin{aligned}\frac{d\lambda(X,\alpha)}{dg(Z,\alpha)} &= \frac{\partial F(\lambda,g)/\partial g(Z,\alpha)}{\partial F(\lambda,g)/\partial \lambda(X,\alpha)} \\ &= E\left[\frac{(\sigma_x+\lambda'(X,\alpha)g(Z,\alpha)-\lambda'(X,\alpha)g(Z,\alpha))/\left(\sigma_x+\lambda'(X,\alpha)g(Z,\alpha)\right)^2}{-g(Z,\alpha)g'(Z,\alpha)/\left(\sigma_x+\lambda'(X,\alpha)g(Z,\alpha)\right)^2}\middle| X\right] \\ &= -\sigma_x\left\{E\left[g(Z,\alpha)g'(Z,\alpha)|\,X\right]\right\}^{-1} \\ &= -\sigma_x\Sigma(X,\alpha)^{-1} \end{aligned} \tag{44}$$

Substituting (44) into (43) we obtain

$$\frac{d\lambda(\alpha,X,Z)}{d\alpha}[\alpha_1-\alpha_2] = -\sigma_x\Sigma(X,\alpha)^{-1}\frac{dg(Z,\alpha)}{d\alpha}[\alpha_1-\alpha_2] \tag{45}$$

Substituting (45) into (42) yields

$$A_4 = \sigma_x^2 E\left[\varpi\left(X,Z,\alpha_0\right)^2\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)' W_0(X,Z)^{-1}\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right]$$

where

$$W_0(X,Z)^{-1} \equiv \Sigma(X,\alpha_0)^{-1}g(Z,\alpha_0)g'(Z,\alpha_0)\Sigma(X,\alpha_0)^{-1}$$

Using (40) in $\varpi\left(X,Z,\alpha_0\right)$ results in

$$A_4 = E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)' W_0(X,Z)^{-1}\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \tag{46}$$

Substituting (46) into (42) and (39) yields

$$\|\alpha_1-\alpha_2\|_F = \sqrt{E\left\{E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)' W_0(X,Z)^{-1}\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right]\right\}} \tag{47}$$

The expression (47) can be viewed as a conditional version of the metric used in Ai and Chen (2003). In particular, if $\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]$ and $g(Z,\alpha_0)$ are independent conditional on X, then (47) reduces to $\sqrt{E\left\{\left(\frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'\Sigma(X,\alpha_0)^{-1}\frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right\}}$ which is the metric used in Ai and Chen (2003) with the efficient weighting matrix.

Note that by (40)

$$\begin{aligned} E\left[\left.\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[\alpha_1-\alpha_2]\right| X\right] &= \lambda'(X,\alpha_0)E\left[\left.\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right| X\right] \\ &\quad +\frac{d\lambda'(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]E\left[g(Z,\alpha_0)|\,X\right] \\ &= 0 \end{aligned}$$

and hence

$$E\left[\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[\alpha_1-\alpha_2]\right)'\left.\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[\alpha_1-\alpha_2]\right| X\right] = Var\left(\left.\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[\alpha_1-\alpha_2]\right| X\right)$$

implying

$$\begin{aligned} \left\|\alpha_1-\alpha_2\right\|_F &= \sqrt{E\left\{Var\left(\left.\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[\alpha_1-\alpha_2]\right| X\right)\right\}} \\ \left\langle v,v\right\rangle_F &= E\left\{Var\left(\left.\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[v]\right| X\right)\right\} \end{aligned}$$

We will now introduce the conditions under which the desired convergence rates are derived.

Assumption 5.1 *(i) $\mathcal{A}$ is convex in α_0, and $g(Z,\alpha)$ is pathwise differentiable at α_0; (ii) for some $c_1, c_2 > 0$,*

$$\begin{aligned} c_1 E\left\{m(X,\alpha_n)'W_0(X)^{-1}m(X,\alpha_n)\right\} &\leq \left\|\alpha_n-\alpha_0\right\|_F^2 \\ &\leq c_2 E\left\{m(X,\alpha_n)'W_0(X)^{-1}m(X,\alpha_n)\right\} \end{aligned}$$

holds for all $\alpha_n \in \mathcal{A}_n$ with $\|\alpha_n-\alpha_0\| = o(1)$.

Assumption 5.2 *For any $\widetilde{g}(\cdot)$ in $\Lambda_c^{\overline{\gamma}}(\mathcal{X})$ with $\overline{\gamma} > d_x/2$, there exists $p^{k_n}(\cdot)'\kappa \in \Lambda_c^{\overline{\gamma}}(\mathcal{X})$ such that $\sup_{X\in\mathcal{X}}\left|\widetilde{g}(X)-p^{k_n}(X)'\kappa\right| = O(k_n^{-\overline{\gamma}/d_x})$, and $k_n^{-\overline{\gamma}/d_x} = o(n^{-1/4})$.*

Assumption 5.3 *(i) Each element of $g(Z,\alpha)$ satisfies an envelope condition in $\alpha_n \in \mathcal{A}_n$; (ii) each element of $m(X,\alpha) \in \Lambda_c^{\overline{\gamma}}(\mathcal{X})$ with $\overline{\gamma} > d_x/2$, for all $\alpha_n \in \mathcal{A}_n$.*

In line with Ai and Chen (2003), let $\xi_{0n} \equiv \sup_{X\in\mathcal{X}}\left\|p^{k_n}(X)\right\|_E$, which is nondecreasing in k_n. Denote $N(\delta,\mathcal{A}_n,\|\cdot\|)$ as the minimal number of radius δ covering balls of $\mathcal{A}_n$ under the $\|\cdot\|$ metric.

Assumption 5.4 $k_{1n} \times \ln n \times \xi_{0n}^2 \times n^{-1/2} = o(1)$.

Assumption 5.5 $\ln\left[N(\varepsilon^{1/\kappa},\mathcal{A}_n,\|\cdot\|)\right] \leq const. \times k_{1n} \times \ln(k_{1n}/\varepsilon)$.

Assumption 5.6 $\Sigma_0(X) \equiv Var\left[g(Z,\alpha_0)|X\right]$ *is positive definite for all $X \in \mathcal{X}$.*

The following result gives the convergence rate of the SLWCEL estimator under the Fisher metric. The proof is provided in the Appendix.

Theorem 5.1 *Under Assumptions 4.1 - 5.6, we have* $\|\widehat{\alpha}_n - \alpha_0\|_F = o_p(n^{-1/4})$.

6 Asymptotic Normality

To derive the asymptotic distribution of $\widehat{\theta}_n$, it suffices to derive the asymptotic distribution of $f(\widehat{\alpha}_n) \equiv \tau'\widehat{\theta}_n$ for any fixed non-zero $\tau \in R^{d_\theta}$. The difference $f(\widehat{\alpha}_n) - f(\alpha_0)$ is linked to the pathwise directional derivatives of the sample criterion function via the inner product involving a Riesz representor v^*. Application of a Central Limit Theorem for triangular arrays of functions indexed by a finite-dimensional parameter then shows the desired result. In this Section we introduce the necessary notation, compute the Riesz representor v^* and state the Theorem of $\sqrt{n}$-normality of $\widehat{\theta}_n$.

Since $f(\alpha) \equiv \tau'\theta$ is a linear functional on $\overline{\mathbf{V}}$, it is bounded (i.e. continuous) if and only if

$$\sup_{0 \neq \alpha - \alpha_0 \in \overline{\mathbf{V}}} \frac{|f(\alpha) - f(\alpha_0)|}{\|\alpha - \alpha_0\|_F} < \infty$$

The Riesz Representation Theorem states that there exists a representor $v^* \in \overline{\mathbf{V}}$ satisfying

$$\|v^*\|_F \equiv \sup_{0 \neq \alpha - \alpha_0 \in \overline{\mathbf{V}}} \frac{|f(\alpha) - f(\alpha_0)|}{\|\alpha - \alpha_0\|_F} \tag{48}$$

and

$$f(\alpha) = f(\alpha_0) + \langle v^*, \alpha - \alpha_0 \rangle_F$$

Hence,

$$f(\widehat{\alpha}_n) - f(\alpha_0) = \langle v^*, \widehat{\alpha}_n - \alpha_0 \rangle_F$$

Let

$$\frac{dg(Z, \alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \frac{dg(Z, \alpha_0)}{d\theta'}(\theta - \theta_0) + \frac{dg(Z, \alpha_0)}{dh}[h - h_0] \tag{49}$$

For any $h \in \overline{\mathcal{H}}$, there exists $w_j(\cdot) \in \overline{\mathcal{W}}$ for $j = 1, ..., d_\theta$ such that

$$h - h_0 = -(w_1, ..., w_{d_\theta})(\theta - \theta_0) = -w(\theta - \theta_0)$$

Define

$$\begin{aligned}\frac{dg(Z,\alpha_0)}{dh}[w] &\equiv \left(\frac{dg(Z,\alpha_0)}{dh}[w_1],...,\frac{dg(Z,\alpha_0)}{dh}[w_{d_\theta}]\right)\\ D_w(Z) &\equiv \frac{dg(Z,\alpha_0)}{d\theta'}-\frac{dg(Z,\alpha_0)}{dh}[w]\end{aligned} \tag{50}$$

where $D_w(Z)$ is a $d_g \times d_\theta-$matrix valued function. Definitions (49) and (50) imply

$$\frac{dg(Z,\alpha_0)}{dh}[h-h_0] = -\frac{dg(Z,\alpha_0)}{dh}[w](\theta-\theta_0)$$

and hence

$$\begin{aligned}D_w(Z)(\theta-\theta_0) &= \frac{dg(Z,\alpha_0)}{d\theta'}(\theta-\theta_0)-\frac{dg(Z,\alpha_0)}{dh}[w](\theta-\theta_0)\\ &= \frac{dg(Z,\alpha_0)}{d\theta'}(\theta-\theta_0)+\frac{dg(Z,\alpha_0)}{dh}[h-h_0]\\ &= \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha-\alpha_0]\end{aligned} \tag{51}$$

By definition of $\|\cdot\|_F$ this implies

$$\begin{aligned}\|\alpha-\alpha_0\|_F^2 &= E\left\{E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha-\alpha_0]\right)' W_0(Z,X)^{-1}\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha-\alpha_0]\right)\middle| X\right]\right\}\\ &= E\left\{E\left[(\theta-\theta_0)' D_w(Z)' W_0(Z,X)^{-1} D_w(Z)(\theta-\theta_0)\middle| X\right]\right\}\end{aligned} \tag{52}$$

Let $w^* = \left(w_1^*,...,w_{d_\theta}^*\right)$ be the solution to

$$\inf_{w_j\in\overline{\mathcal{W}},j=1,...,d_\theta} E\left\{E\left[(\theta-\theta_0)' D_w(Z)' W_0(Z,X)^{-1} D_w(Z)(\theta-\theta_0)\middle| X\right]\right\} \tag{53}$$

where "inf" is in positive semidefinite matrix sense. Using the definitions of w^*, $f(\alpha)$, (48) and (52)

$$\begin{aligned}\|v^*\|_F^2 &\equiv \sup_{0\neq\alpha-\alpha_0\in\overline{\mathbf{V}}}\frac{|f(\alpha)-f(\alpha_0)|^2}{\|\alpha-\alpha_0\|_F^2}\\ &= \frac{(\theta-\theta_0)'\tau\tau'(\theta-\theta_0)}{(\theta-\theta_0)' E\{E[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)|X]\}(\theta-\theta_0)}\\ &= \tau'\left[E\left\{E\left[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau\end{aligned} \tag{54}$$

where $v^* \equiv (v_\theta^*, v_h^*) \in \overline{\mathbf{V}}$. By the definition of w^*, $v_h^* = -w^* \times v_\theta^*$. From this and (51) we have

$$\frac{dg(Z,\alpha_0)}{d\alpha}[v^*] = D_{w^*}(Z)v_\theta^* \tag{55}$$

Let

$$v_\theta^* = \left[E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau \tag{56}$$

Substituting (56) into the definition of $\|\cdot\|_F^2$ in (39) via the expression for (55) yields

$$\begin{aligned}
\|v^*\|_F^2 &= E\left\{E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[v^*]\right)' W_0(Z,X)^{-1}\left(\frac{dg(Z,\alpha_0)}{d\alpha}[v^*]\right)\middle| X\right]\right\} \\
&= E\left\{E\left[\left(D_{w^*}(Z)v_\theta^*\right)' W_0(Z,X)^{-1}\left(D_{w^*}(Z)v_\theta^*\right)\middle| X\right]\right\} \\
&= v_\theta^{*\prime}E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle| X\right]\right\}v_\theta^* \\
&= \tau'\left[E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\}\right]^{-1} \\
&\quad \times E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle| X\right]\right\} \\
&\quad \times \left[E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau \\
&= \tau'\left[E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau
\end{aligned}$$

which matches (54) and thus validates (56) shown unique by the Riesz Representation Theorem.

The following additional conditions correspond to Assumptions 4.1-4.3 in Ai and Chen (2003) and are sufficient for the $\sqrt{n}$-normality of $\widehat{\theta}_n$:

Assumption 6.1 *(i)* $E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\}$ *is positive definite; (ii)* $\theta_0 \in int(\Theta)$; *(iii)* $\Sigma_0(X) \equiv Var[g(Z,\alpha_0)|X]$ *is positive definite for all* $X \in \mathcal{X}$.

Assumption 6.2 *There is a* $v_n^* = (v_\theta^*, -\Pi_n w^* \times v_\theta^*) \in \mathcal{A}_n - \alpha_0$ *such that* $\|v_n^* - v^*\|_F = O(n^{-1/4})$.

Following Ai and Chen (2003), let $\mathcal{N}_{0n} \equiv \{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| = o(1), \|\alpha_n - \alpha_0\|_F = o(n^{-1/4})\}$ and define $\mathcal{N}_0$ the same way with $\mathcal{A}_n$ replaced by $\mathcal{A}$. Also, for any $v \in \overline{\mathbf{V}}$, denote

$$\frac{dg(Z,\alpha)}{d\alpha}[v] \equiv \left.\frac{dg(Z,\alpha+tv)}{dt}\right|_{t=0} \quad \text{a.s. } Z$$

and

$$\frac{dm(Z,\alpha)}{d\alpha}[v] \equiv E\left\{\left.\frac{dg(Z,\alpha)}{d\alpha}[v]\ \right| X\right\} \quad \text{a.s. } Z$$

Assumption 6.3 *For all* $\alpha \in \mathcal{N}_0$, *the pathwise first derivative* $(dg(Z,\alpha(t))/d\alpha)[v]$ *exists a.s.* $Z \in \mathcal{Z}$. *Moreover, (i) each element of* $(dg(Z,\alpha(t))/d\alpha)[v_n^*]$ *satisfies the envelope condition and is Hölder continuous in* $\alpha \in \mathcal{N}_{0n}$; *(ii) each element of* $(dm(Z,\alpha(t))/d\alpha)[v_n^*]$ *is in* $\Lambda_c^\gamma(\mathcal{X})$, $\gamma > d_x/2$ *for all* $\alpha \in \mathcal{N}_0$.

The following result is proved in the Appendix.

Theorem 6.1 *Under Assumptions 4.1-4.8, 5.1-5.6 and 6.1-6.3,* $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Omega)$ *where*

$$\begin{aligned} \Omega &= E\left[Var\left(\frac{d\varphi(X, Z, \alpha_0)}{dg(Z, \alpha)} D_{w^*}(Z)\middle| X\right)\right] \\ &= \left[E\left\{E\left[D_{w^*}(Z)' W_0(Z, X)^{-1} D_{w^*}(Z)\middle| X\right]\right\}\right]^{-1} \end{aligned} \tag{57}$$

Note that if $D_w(Z)$ and $g(Z, \alpha_0)$ are independent conditional on X then the expression (57) reduces to the asymptotic variance-covariance formula (22) in Ai and Chen (2003) that is shown to be asymptotically efficient by these authors. A consistent estimator of Ω can be obtained in the following way: First estimate $W_0(x_i, z_j)^{-1}$ with

$$\begin{aligned} w_{ij} &= p^{k_n}(x_j)' \left(P'P\right)^{-1} p^{k_n}(x_i) \\ \widehat{\Sigma}(x_i, \widehat{\alpha}_n) &= \sum_{j=1}^{n} w_{ij} g(z_j, \widehat{\alpha}_n) g'(z_j, \widehat{\alpha}_n) \\ \widehat{W_0}(x_i, z_j)^{-1} &= \widehat{\Sigma}(x_i, \widehat{\alpha}_n)^{-1} g(z_j, \widehat{\alpha}_n) g'(z_j, \widehat{\alpha}_n) \widehat{\Sigma}(x_i, \widehat{\alpha}_n)^{-1} \end{aligned} \tag{58}$$

Then for each $s = 1, ..., d_\theta$ estimate w_s^* with $\widehat{w}_s^*$ which is a solution to the minimization problem

$$\begin{aligned} \min_{w_s \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} &\left(\frac{dg(z_j, \widehat{\alpha}_n)}{d\theta_s} - \frac{dg(z_j, \widehat{\alpha}_n)}{dh}[w_s]\right)' \widehat{W_0}(z_j, x_i)^{-1} \\ &\times \left(\frac{dg(z_j, \widehat{\alpha}_n)}{d\theta_s} - \frac{dg(z_j, \widehat{\alpha}_n)}{dh}[w_s]\right) \end{aligned}$$

and let $\widehat{w}^* = (\widehat{w}_1^*, ..., \widehat{w}_{d_\theta}^*)$ implying

$$\widehat{D}_{\widehat{w}^*}(z_j) = \frac{dg(z_j, \widehat{\alpha}_n)}{d\theta_s} - \frac{dg(z_j, \widehat{\alpha}_n)}{dh}[\widehat{w}^*] \tag{59}$$

Finally, use (58) and (59) in a finite-sample analog of (57) to obtain

$$\widehat{\Omega} = \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}' \widehat{D}_{\widehat{w}^*}(z_j)' \widehat{W_0}(x_i, z_j)^{-1} \widehat{D}_{\widehat{w}^*}(z_j)\right]^{-1}$$

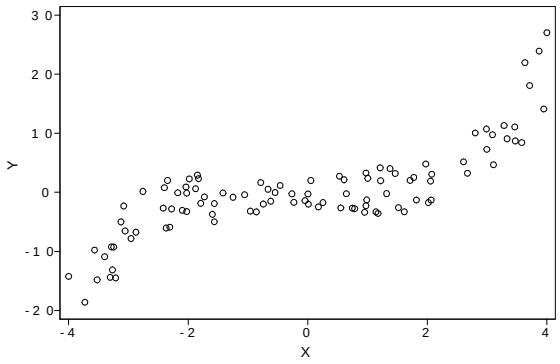
We note that for linear sieves computing $\widehat{w}_s^*$ does not require nonlinear optimization and thus the covariance estimator is easy to compute.

7 Simulation

To evaluate the finite sample performance of the estimator $\widehat{\theta}_{LWCEL}$ defined in (26) against KTA's $\widehat{\theta}_{CEL}$ we have conducted a small scale pilot Monte Carlo (MC) simulation study aimed at maximum

simplicity of the simulation design. More extensive MC analysis assessing the performance of LWCEL and SLWCEL is currently being conducted and will be included in further updates of the paper. We set $Z = X$ and $Y = \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$ with heteroskedastic $e = 0.5u|X|$, $u = U(-5, 5)$. A random sample $N = 100$ of $X \tilde{} N(0, 2)$ was truncated at -1 and 1 and spread over the interval $[-4, 4]$ to avoid far outliers. The true parameter values were set at $\beta_1 = -0.2$, $\beta_2 = 0.1$, $\beta_3 = 0.3$. A typical data draw looks as follows:

Figure 1: Sample Simulated Data



In order to deal with possible negative arguments in the log function, we followed the approach suggested by Owen (2001) cited in Kitamura (2006) (p. 51): for a small number $\delta = 0.2$ we used the objective function

$$\log_* y = \begin{cases} \log(y) & \text{if } y > \delta \\ \log(\delta) - 1.5 + 2y/\delta - \delta^2/2\delta^2 & \text{if } y \leq \delta \end{cases}$$

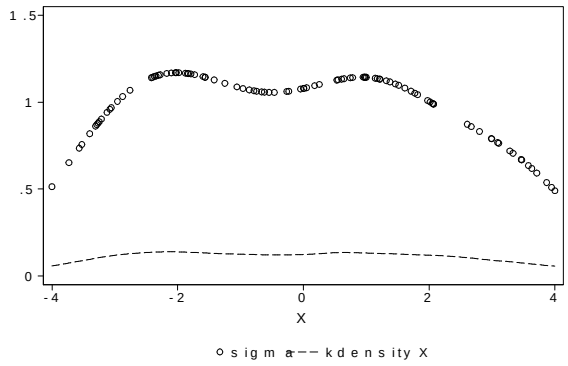
Indeed, the proportion of $y \leq \delta$ in the overall sample was 6.6×10^{-3} and 4.7×10^{-3} for $\widehat{\theta}_{LWCEL}$ and $\widehat{\theta}_{CEL}$, respectively. The Nadaraya-Watson kernel estimator (Pagan and Ullah, 1999, p.86) with the Gaussian kernel, employing the Silverman's rule of thumb for the bandwidth determination (Silverman, 1986, p.45), was used to calculate w_{ij} the case of $\widehat{\theta}_{CEL}$. Thus each i-th local conditional empirical likelihood of $\widehat{\theta}_{CEL}$ was normalized with its corresponding $\sum_{j=1}^{N} w_{ij}$ in the denominator of the Nadaraya-Watson kernel estimator. In contrast, the denominator of the Nadaraya-Watson kernel estimator was replaced with $n^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$ for the case of $\widehat{\theta}_{LWCEL}$. This is equivalent (up to a constant of proportionality) to weighting each i-th local conditional empirical likelihood of $\widehat{\theta}_{LWCEL}$ with σ_i as defined in (**??**). We compared bias, variance and mean-square error over 100 MC iterations on the three estimated coefficients β_1, β_2 and β_3. The results are as follows:

Table 1: Simulation Results

Criterion	*Estimate*	*CEL*	*LWCEL*
Bias	$\widehat{\beta}_1$	-9.100×10^{-2}	-8.619×10^{-2}
	$\widehat{\beta}_2$	1.436×10^{-2}	1.471×10^{-2}
	$\widehat{\beta}_3$	1.050×10^{-2}	9.416×10^{-3}
Variance	$\widehat{\beta}_1$	8.297×10^{-3}	6.189×10^{-3}
	$\widehat{\beta}_2$	2.474×10^{-3}	2.351×10^{-3}
	$\widehat{\beta}_3$	4.202×10^{-4}	3.916×10^{-4}
MSE	$\widehat{\beta}_1$	1.652×10^{-2}	1.362×10^{-2}
	$\widehat{\beta}_2$	2.681×10^{-3}	2.568×10^{-3}
	$\widehat{\beta}_3$	5.304×10^{-4}	4.802×10^{-4}

Both estimators performed relatively well under the simulation scenario which can be attributed to the relatively well-behaved nature of the data. Nonetheless, the $\widehat{\theta}_{LWCEL}$ improved on the $\widehat{\theta}_{CEL}$ in all cases, barring one bias term. The values of σ_i were also retained as an interesting byproduct of the $\widehat{\theta}_{LWCEL}$ estimation procedure, weighting individual local conditional empirical log likelihoods. Naturally, their magnitude follows the density of the data juxtaposed against σ_i in Figure 2:

Figure 2: Plot of σ_i against x_i



8 Conclusion

In this paper we propose a new form of the Conditional Empirical Likelihood (CEL), the Locally Weighted CEL (LWCEL) estimator for models of conditional moment restrictions that contain finite dimensional unknown parameters θ. This estimator extends the CEL analyzed by Kitamura et al. (2004). In contrast to previous literature, we consider an information-theoretic dual locally

weighted GMC optimization problem built directly on conditional moments that minimizes a discrepancy from a probability measure according to which the data was distributed. In a Monte Carlo study, we show that the resulting estimator exhibits better finite-sample properties in the finite-dimensional case $E\left[g\left(Z,\theta_0\right)|X\right]=0$ than found in the previous literature. We further extend the LWCEL estimator to the semiparametric environment defined by models of conditional moment restrictions $E\left[g\left(Z,\alpha_0\right)|X\right]=0$ containing both θ and infinite dimensional unknown functions h. We establish consistency of the new estimator $\widehat{\alpha}_n$, convergence rates of $\widehat{\alpha}_n$ under the Fisher norm, and asymptotic normality of the finite-dimensional component $\widehat{\theta}_n$. The new Sieve-based LWCEL estimator (SLWCEL) is a direct alternative to the Sieve Minimum Distance estimators considered by Ai and Chen (2003) and Newey and Powell (2003). As shown by Newey and Smith (2004), GEL-type estimators, such as EL, outperform the GMM estimator in terms of higher-order properties in parametric models $E\left[g\left(Z,\theta_0\right)|X\right]=0$. We conjecture that a similar type of improvements is likely to occur also in the semiparametric context of $E\left[g\left(Z,\alpha_0\right)|X\right]=0$.

Appendix 1: Proofs of Main Results

LWCEL

Note: just bare bones results stated - needs tidying up.

Lemma 1 $E_{Q(x)}\left[\rho\left(\theta, dQ(y|x)\right)|X\right] = \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} D\left(\Pi(x,y), Q(x,y)\right)$

Proof.

Let $\mathbf{M}_Y$ denote the set of all probability densities on $\mathbb{R}^{d_Y}$ and let

$$\boldsymbol{\pi}(X;\theta) \equiv \left\{\pi(y|x) \in \mathbf{M}_Y : \int \pi(y|x) g\left(Z,\theta\right) dm(y|x) = 0;\ X \in \mathcal{X}\right\}$$

Define the set of all probability densities that are compatible with the conditional moment restriction (**??**) by

$$\boldsymbol{\pi}(X) \equiv \cup_{\theta\in\Theta}\boldsymbol{\pi}(X;\theta)$$

This result can be conveniently derived by converting the optimization problem into one in which all integral operators are taken with respect to the Lebesgue measure. In (9) multiply the argument inside $\phi\left(\cdot\right)$ by $\frac{d\Pi(x)}{dQ(x)} = \frac{\pi(x)}{q(x)} = 1$ to obtain

$$\begin{aligned}
\inf_{\theta\in\Theta} E_{Q(x)}\left[\rho\left(\theta, dQ(y|x)\right)|X\right] &= \int_{\mathcal{Y}} \phi\left(\frac{d\Pi(y|x)}{dQ(y|x)}\right) dQ(y|x) \\
&= \inf_{\theta\in\Theta} E_{Q(x)}\left[\inf_{\pi(y|x)\in\boldsymbol{\pi}(X)} D\left(\pi(y|x), q(y|x)\right)\right] \\
&= \inf_{\theta\in\Theta} \int q(x)\left[\inf_{\pi(y|x)\in\boldsymbol{\pi}(X)} \int q(y|x)\phi\left(\frac{\pi(y|x)}{q(y|x)}\right) dm(y|x)\right] dm(x) \\
&= \inf_{\theta\in\Theta} \inf_{\{\pi(y|x)\in\mathbf{M}_Y : X\in\mathcal{X}\}} \int\int q(x)q(y|x)\phi\left(\frac{\pi(y|x)}{q(y|x)}\frac{\pi(x)}{q(x)}\right) dm(y|x)dm(x) \\
&= \inf_{\theta\in\Theta} \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} \int q(x,y)\phi\left(\frac{\pi(x,y)}{q(x,y)}\right) dm(x,y) \\
&= \inf_{\theta\in\Theta} \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} D\left(\Pi(x,y), Q(x,y)\right) \qquad (60)
\end{aligned}$$

The marginal density $q(x)$ of X is independent of the parameter θ and hence the former can be estimated directly from the data. The same holds for the "choice" marginal density $\pi(x)$ in the optimization problem and hence $\pi(x) = q(x)$ which was used in deriving the expression above. ■

Discussion of Consistency

In outlining our consistency proof, we follow the discussion as given by KTA and extend it to our case of infinite dimensional parameter space. For a standard extremum estimation procedure (for example via maximization), consistency can be shown by considering the sample objective function and its population counterpart and arguing in the following manner. Consider an arbitrary neighborhood of the true parameter value. Check that:

(A) Outside the neighborhood, the sample objective function is bounded away from the maximum of the population objective function achieved at the true parameter value, w.p.a. 1.

(B) The maximum of the sample objective function is by definition not smaller than its value at the true parameter value. The latter converges to the population objective function evaluated at the true value, due to the LLN.

By (A) and (B) the maximum of the sample objective function is unlikely to occur outside the (arbitrarily defined) neighborhood for large samples. This shows the consistency.

While Newey and Powell (2003) were able to recast their estimator as an argmin of a quadratic form delivering (A), in Chen (2005) (Theorem 3.1) (A) is assumed. In our problem, however, such approach cannot be applied directly. Specifically, showing (A) is problematic here, since the objective function G_n defined in (34) contains the Lagrange multiplier $\lambda(\alpha_n)$ which is endogenously determined at each α_n. Therefore, in our proof we follow the KTA approach binding G_n with a dominating function and then check (A) for the latter by comparing the convergence rates of G_n at α_0 and outside a $\delta-$neighborhood of α_0. The convergence rate of $G_n(\alpha_0)$ is a new result which differs from the one of KTA since the definition of our G_n contains an additional term σ_i arising from the use of a different weighting scheme and due to our estimator being based on series rather than kernel weights. In our proof, a Uniform Law of Large Numbers (ULLN) for the dominating function is used only for α_n outside the $\delta-$neighborhood of α_0.

Regarding the complications incurred by considering an infinite dimensional parameter space α, we note that our consistency proof differs from the ones used in Newey and Powell (2003) (Theorem 1) and Chen (2005) (Theorem 3.1) for M-estimators with α. Using a ULLN over the sieve space, these authors show that the sample objective function G_n and its expectation are, w.p.a 1, within a $\delta-$neighborhood of each other when evaluated at a parameter $\widetilde{\alpha}_n$ in the sieve space that converges to the true parameter value α_0. Existence of such parameter $\widetilde{\alpha}_n$ is guaranteed by the definition of the sieve space. This approach, however, would necessitate evaluating the convergence rates of $G_n(\widetilde{\alpha}_n)$ to its expectation which is problematic in our saddle-point case since it is difficult to capture the behavior of the endogenous $\lambda_i(\alpha)$ away from α_0. Recall that $\widehat{\alpha}_n$ is defined as maximizing $G_n(\alpha_n)$ over the sieve space $\mathcal{A}_n$ and thus using $G_n(\alpha)$, $\alpha \in \mathcal{A}$ for estimation purposes would yield an unfeasible estimator. Nonetheless, the function $g(z_j, \alpha)$ and hence the functions $G_n(\alpha)$ and $\Sigma_n(x_i, \alpha)$ can theoretically be evaluated at a generic parameter value $\alpha \in \mathcal{A}$ not restricted to the sieve space. Hence the asymptotic rate of convergence of $G_n(\alpha_0)$ at the true parameter value can be derived to facilitate asymptotic analysis.

Further Notation

Let us introduce some additional notation. Let $\|\cdot\|_E$ denote the Euclidean norm. Define

$$\begin{aligned} a_i &\equiv \sigma_i - 1 \\ &= \sum_{j=1}^{n} w_{ij} - 1 \\ &= \mathbf{i}' P \left(P'P\right)^{-1} p^{k_n}(x_i) - 1 \end{aligned}$$

For generic n vectors z and a vector x we drop the subscript i and use

$$a_x \equiv \mathbf{i}' P \left(P'P\right)^{-1} p^{k_n}(x) - 1 \tag{61}$$

Further define $B(\alpha_0, \delta)$ and $B_n(\alpha_0, \delta)$ as $\delta-$neighborhoods around α_0 with $B(\alpha_0, \delta) \subset \mathcal{A}$ and $B_n(\alpha_0, \delta) \subset \mathcal{A}_n$, respectively. Consider the function $\psi(X, \alpha)$ as defined in (38). Denote

$$\begin{aligned} \psi_n(x_i, \alpha) &\equiv \sum_{j=1}^{n} w_{ij} \varphi(x_i, z_j, \alpha) \\ &= \sum_{j=1}^{n} w_{ij} \ln\left\{\sigma_i + \lambda_i' g(z_j, \alpha)\right\} \end{aligned} \tag{62}$$

$$\begin{aligned} G_n(\alpha_n) &\equiv -\frac{1}{n} \sum_{i=1}^{n} \psi_n(x_i, \alpha) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \varphi(x_i, z_j, \alpha) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \ln\left\{\sigma_i + \lambda_i' g(z_j, \alpha_n)\right\} \end{aligned} \tag{63}$$

$$\begin{aligned} \Sigma_n(x_i, \alpha) &\equiv \sum_{j=1}^{n} w_{ij} g(z_j, \alpha) g'(z_j, \alpha) \\ \Sigma(X, \alpha) &\equiv E_Z\left[\Sigma_n(X, \alpha)\right] \end{aligned} \tag{64}$$

and recall the definition of $\Sigma_0(X) \equiv Var[g(Z, \alpha_0)|X]$ in Assumption 6.1 (iii).

Main Proofs

Proof of Theorem 4.1. Following KTA, in the asymptotic analysis we will replace $\lambda_i(\alpha)$ by

$$u(x_i, \alpha) = \frac{E\left[g(z, \alpha)|x_i\right]}{(1 + \|E\left[g(z, \alpha)|x_i\right]\|)}$$

For a constant $\widetilde{c} \in (0, 1)$ define a sequence of truncation sets

$$C_n = \left\{ z : \sup_{\alpha \in \mathcal{A}} \left| a_x + u'(x, \alpha_n) g(z, \alpha_n) \right| \leq \widetilde{c} n^{1/m} \right\} \tag{65}$$

and let

$$s_n \equiv n^{-1/m} \left[a_x + u'(x, \alpha_n) g(z, \alpha_n)\right] \mathbb{I}\{z \in C_n\} \tag{66}$$

Let

$$\begin{aligned} q_n(x, z, \alpha_n) &= -\log\left(1 + n^{-1/m} \left[a_x + u'(x, \alpha_n) g(z, \alpha_n)\right] \mathbb{I}\{z \in C_n\}\right) \\ &= -\log(1 + s_n) \end{aligned}$$

The modified objective function is

$$Q_n(\alpha_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} q_n(x_i, z_j, \alpha_n) \tag{67}$$

Note that

$$G_n(\alpha_n) \leq Q_n(\alpha_n) \tag{68}$$

for all $\alpha_n \in \mathcal{A}_n$ by the optimality of λ_i.

Then by the Taylor series expansion for logarithms

$$\begin{aligned} q_n(x,z,\alpha_n) &= -\log(1+s_n) \\ &= -s_n + \frac{\widetilde{s}_n^2}{2} \\ &= -s_n + \frac{s_n^2}{2(1-ts_n)} \\ &= -n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right]\mathbb{I}\{z \in C_n\} + \frac{s_n^2}{2(1-ts_n)} \end{aligned}$$

$$\begin{aligned} &= n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right] - n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right] \\ &\quad -n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right]\mathbb{I}\{z \in C_n\} + \frac{s_n^2}{2(1-ts_n)} \\ &= -n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right] \\ &\quad +n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right](1-\mathbb{I}\{z \in C_n\}) + \frac{s_n^2}{2(1-ts_n)} \\ &= -n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right] + R_n(t,a_x,\alpha_n) \end{aligned} \tag{69}$$

where

$$\begin{aligned} R_n(t,a_x,\alpha_n) &= n^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right](1-\mathbb{I}\{z \in C_n\}) \\ &\quad + \frac{n^{-2/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right]^2 \mathbb{I}\{z \in C_n\}}{2(1-tn^{-1/m}\left[a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right]\mathbb{I}\{z \in C_n\})^2} \end{aligned}$$

Note that, by the triangle and Cauchy-Schwarz inequalities

$$\begin{aligned} |R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x| + \left\|u'(x,\alpha_n)\right\| \|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z \in C_n\}) \\ &\quad + \frac{n^{-2/m}\left[a_x^2 + 2\|a_x\| \|u'(x,\alpha_n)\| \|g(z,\alpha_n)\| + \|u'(x,\alpha_n)\|^2 \|g(z,\alpha_n)\|^2\right]\mathbb{I}\{z \in C_n\}}{2(1-tn^{-1/m}\left[a_x + u'(x,\alpha_n)\, g_n(z,\alpha_n)\right])^2} \end{aligned}$$

and by $\|u'(x,\alpha_n)\| < 1$ we obtain

$$\begin{aligned} |R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x| + \|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z \in C_n\}) \\ &\quad + \frac{n^{-2/m}\left[a_x^2 + 2a_x\|g(z,\alpha_n)\| + \|g(z,\alpha_n)\|^2\right]}{2(1-tn^{-1/m}\left[a_x + u'(x,\alpha_n)\, g_n(z,\alpha_n)\right])^2} \end{aligned}$$

From (65) it follows that

$$\begin{aligned} \widetilde{c} &\geq n^{-1/m}\sup_{\alpha\in\mathcal{A}}\left|a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right| \\ &\geq n^{-1/m}\left|a_x + u'(x,\alpha_n)\, g(z,\alpha_n)\right| \\ &\geq tn^{-1/m}\left|a_x + u'(x,\alpha_n)\, g_n(z,\alpha_n)\right| \end{aligned}$$

and hence

$$\begin{aligned} |R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x| + \|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z \in C_n\}) \\ &\quad + \frac{n^{-2/m}\left[a_x^2 + 2a_x\|g(z,\alpha_n)\| + \|g(z,\alpha_n)\|^2\right]}{2(1-\widetilde{c})^2} \\ &= n^{-1/m}\left[|a_x| + \|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z \in C_n\}) \\ &\quad + n^{-2/m}\frac{a_x^2}{2(1-\widetilde{c})^2} + \frac{n^{-2/m}\left[2a_x\|g(z,\alpha_n)\| + \|g(z,\alpha_n)\|^2\right]}{2(1-\widetilde{c})^2} \end{aligned}$$

taking sup over $\mathcal{A}$ we obtain

$$\begin{aligned}\sup_{\alpha\in\mathcal{A}}|R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x|+\sup_{\alpha\in\mathcal{A}}\|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z\in C_n\})+n^{-2/m}\frac{a_x^2}{2(1-\widetilde{c})^2}\\ &\quad+\frac{n^{-2/m}\left[2a_x\sup_{\alpha\in\mathcal{A}}\|g(z,\alpha_n)\|+\sup_{\alpha\in\mathcal{A}}\|g(z,\alpha_n)\|^2\right]}{2(1-\widetilde{c})^2}\end{aligned} \tag{70}$$

In view of (69) and (70) approximate $n^{1/m}Q_n(\alpha_n)$ by $n^{1/m}\overline{Q}_n(\alpha_n)$ where

$$\overline{Q}_n(\alpha_n)=-\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}u'(x_i,\alpha_n)E[g(z,\alpha_n)|x_i] \tag{71}$$

Lemma A.2 shows that

$$n^{1/m}Q_n(\alpha_n)=n^{1/m}\overline{Q}_n(\alpha_n)+o_p(1)\qquad\text{uniformly in }\alpha_n\in\mathcal{A}_n \tag{72}$$

Next, we will apply a uniform law of large numbers to $n^{1/m}\overline{Q}_n(\alpha)$ over the whole parameter space $\mathcal{A}$. Under Assumptions 4.4(i), 4.5, and 4.6 $E[g(z,\alpha)|x_i]$ is continuous in $\alpha\in\mathcal{A}$ by Corollary 4.2 of Newey (1991), and so is

$$-u'(x_i,\alpha)E[g(z,\alpha)|x_i]=-\frac{\|E[g(z,\alpha)|x_i]\|^2}{1+\|E[g(z,\alpha)|x_i]\|}$$

Under Assumption 4.5(i) $E\left[\sup_{\alpha\in\mathcal{A}}|-u'(x_i,\alpha)E[g(z,\alpha)|x_i]|\right]<\infty$. These, together with Assumption 4.4(i) satisfy the conditions of Lemma A2 of Newey and Powell (2003) implying the following uniform law of large numbers:

$$\sup_{\alpha\in\mathcal{A}}\left|n^{1/m}\overline{Q}_n(\alpha)-E\left[-u'(x_i,\alpha)E[g(z,\alpha)|x_i]\right]\right|=o_p(1) \tag{73}$$

where $-E[-u'(x_i,\alpha)E[g(z,\alpha)|x_i]]$ is continuous in $\mathcal{A}$. This function is bounded above by

$$-E\left[u'(x_i,\alpha)E[g(z,\alpha)|x_i]\right]\leq -E\left[\mathbb{I}\{x\in\mathcal{X}_{\mathcal{A}}\}\|E[g(z,\alpha)|x_i]\|^2/(1+\|E[g(z,\alpha)|x_i]\|)\right] \tag{74}$$

By Assumption 4.1, the right-hand side of this inequality is strictly negative at each $\alpha\neq\alpha_0$. Therefore, by continuity of $-E[u'(x_i,\alpha)E[g(z,\alpha)|x_i]]$ and compactness of $\mathcal{A}$, there exists a strictly positive number $H(\delta)$ such that

$$\sup_{\alpha\in\mathcal{A}\setminus B(\alpha_0,\delta)}E\left[-u'(x_i,\alpha)E[g(z,\alpha)|x_i]\right]\leq -H(\delta) \tag{75}$$

By (68), (72), and Assumption 4.4(ii) we have

$$\sup_{\alpha_n\in\mathcal{A}_n}n^{1/m}G_n(\alpha_n)\leq\sup_{\alpha_n\in\mathcal{A}_n}n^{1/m}Q_n(\alpha_n)=\sup_{\alpha_n\in\mathcal{A}_n}n^{1/m}\overline{Q}_n(\alpha_n)+o_p(1) \tag{76}$$

Together (76) with (75) and (73) imply that

$$\Pr\left\{\sup_{\alpha_n\in\mathcal{A}_n\setminus B_n(\alpha_0,\delta)}G_n(\alpha_n)>-n^{-1/m}H(\delta)\right\}<\delta/2\qquad\text{eventually.} \tag{77}$$

Next, we evaluate G_n at the true value α_0 and show that $G_n(\alpha_0)$ converges to its expectation faster than $G_n(\alpha_n)$ with α_n outside a $\delta-$neighborhood of α_0 whose convergence rate is given in (77). Having established this fact the conclusion of the proof is then straightforward. This approach was taken by KTA for the finite-dimensional parameter θ and we extend it to the infinite-dimensional parameter α. Our way of deriving the rate of convergence of $G_n(\alpha_0)$ differs from KTA, though, because we do not make use of kernel-based results. Rather, based on the series literature, we derive a new result for the rate of convergence by specializing Lemma A.1(A) of Ai and Chen (2003) to our case.

Using Lemma A.4 and the fact

$$1+a_i=\sum_{j=1}^{n}w_{ij}>0\quad\text{for each }i$$

we obtain

$$\begin{aligned} G_n(\alpha_0) &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\left(1+a_i+\lambda_i'(\alpha_0)\,g(z_j,\alpha_0)\right) \\ &\geq -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(a_i+\lambda_i'(\alpha_0)\,g(z_j,\alpha_0)\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\lambda_i'(\alpha_0)\sum_{j=1}^{n} w_{ij}g(z_j,\alpha_0) \\ &\geq -\max_{1\leq i\leq n}\|\lambda_i(\alpha_0)\|\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g(z_j,\alpha_0)\right\| \end{aligned}$$

Then by Lemmas A.1 and A8,

$$\begin{aligned} G_n(\alpha_0) &= \left\{o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)\right\}^2 \\ &= o_p(r_n^2) \end{aligned}$$

where

$$r_n \equiv o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

with $\widetilde{\delta}_{1n}$ defined in Lemma A.7 and ϱ defined in 4.7. Therefore, we have the following LLN

$$\Pr\left\{G_n(\alpha_0) < -r_n^2 H(\delta)\right\} < \delta/2 \qquad \text{eventually.} \tag{78}$$

Denote

$$\begin{aligned} \widehat{Q}_1(\alpha) &\equiv n^{1/m}G_n(\alpha) \\ \widehat{Q}_2(\alpha) &\equiv r_n^{-2}G_n(\alpha) \\ Q_1(\alpha) &\equiv -E\left[u'(x,\alpha)E\left[g(z,\alpha)|x\right]\right] \\ Q_2(\alpha) &\equiv E\widehat{Q}_2(\alpha) \end{aligned}$$

where the last expectation is taken with respect to the joint density of (Y,X). Under Assumptions 4.4(i), 4.5, and 4.6 $Q_2(\alpha)$ is continuous in $\alpha\in\mathcal{A}$ by Corollary 4.2 of Newey (1991). Note that since $n^{1/m}r_n^2\to 0$ and $n^{1/m}G_n(\alpha)\leq 0$, by (73) and (76), w.p.a. 1,

$$\begin{aligned} r_n^{-2} &> n^{1/m} \\ \widehat{Q}_2(\alpha) &\leq \widehat{Q}_1(\alpha) \end{aligned} \tag{79}$$

If we retain $\lambda_i(\alpha)$ instead of $u(x,\alpha)$ in the definition of $Q_n(\alpha)$ in (67), using $\lambda_i(\alpha)=O_p(1)$ which follows from (35), we can derive an analog of $\overline{Q}_n(\alpha)$ in (71) as

$$\overline{Q}_{2n}(\alpha) = -\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}\lambda_i'(\alpha)E\left[g(z,\alpha)|x_i\right]$$

By a corresponding analog of (72) and the moment restriction $E\left[g(z,\alpha_0)|x_i\right]=0$ it follows that $\overline{Q}_{2n}(\alpha_0)=0$ and $Q_2(\alpha_0)=0$. Also, by (74) $Q_1(\widehat{\alpha}_n)<0$ for each $\theta\neq\theta_0$ and thus

$$Q_1(\widehat{\alpha}_n)\leq 0 \tag{80}$$

Then, w.p.a. 1,

$$\begin{aligned} Q_1(\widehat{\alpha}_n) &\geq \widehat{Q}_1(\widehat{\alpha}_n) + H(\delta)/2 && (81)\\ &\geq \widehat{Q}_1(\alpha_0) + H(\delta)/2 && (82)\\ &\geq \widehat{Q}_2(\alpha_0) + H(\delta)/2 && (83)\\ &> Q_2(\alpha_0) + H(\delta) && (84)\\ &= H(\delta) && (85) \end{aligned}$$

where (81) holds by (73) and (76), (82) holds by the definition of $\widehat{\alpha}_n$, (83) by (79), (84) by LLN at α_0 (78), and (85) by $Q_2(\alpha_0) = 0$. By (80) and δ being arbitrary, taking $H(\delta) \to 0$,

$$\widehat{Q}_1(\widehat{\alpha}_n) \xrightarrow{p} 0$$

Then, using Assumption 4.4(ii), $\Pr\left(\left|\widehat{Q}_1(\widehat{\alpha}) - Q_2(\alpha_0)\right| \geq H(\delta)\right) \to 0$ and by (77) $\Pr\left(\widehat{\alpha}_n \in \mathcal{A}_n \backslash B_n\left(\theta_0, \delta\right)\right) \to 0$.

■

Proof of Theorem 5.1.

In deriving the convergence rates under the Fisher norm $\|\cdot\|_F$ we will proceed in a way that is similar to the proof of Theorem 3.1 in Ai and Chen (2003). Specifically, we will use their Lemma A.1 and Corollary A.1 that hold for a generic function $m(X,\alpha)$ and the Euclidean metric. However, since our objective function and metric differs from the ones used by these authors, we need to derive the counterparts of their Corollaries A.2 and B.1 for our case.

Recall the definition of $G_n(\alpha_n)$ in (63)

$$G_n(\alpha_n) \equiv -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \ln\left\{\sigma_i + \lambda_i' g\left(z_j,\alpha_n\right)\right\}$$

and define

$$\overline{G}_n(\alpha_n) \equiv -\frac{1}{n}\sum_{i=1}^{n} E\left[\ln\left\{\sigma_i + \lambda_i' g\left(z,\alpha_n\right)\right\}|x_i\right] \tag{86}$$

Let $\delta_{0n} = o(n^{-1/4})$ and denote $\alpha_{n0} = \Pi\alpha_0$ (the orthogonal projection of α_0 onto the sieve space).

$$P\left(\|\widehat{\alpha}_n - \alpha_0\|_F \geq \delta_{0n}\right) = P\left(\sup_{\left\{\|\widehat{\alpha}_n-\alpha_0\|_F \geq \delta_{0n}, \alpha_n \in \mathcal{A}_n\right\}} G_n(\alpha_n) \geq G_n(\alpha_{n0})\right)$$

For the sake of brevity, let "AC" stand for "Ai and Chen (2003)" for the remainder of the proof. Note that Assumptions 3.1-3.2, 3.6-3.8 and 4.1(iii) in AC are equivalent to our Assumptions 4.2, 4.3, 5.2, 4.5, 4.6, 5.3-5.5 and 5.6, respectively. Assumption 3.3 in AC is implied by our Assumption 4.1 and the condition (1). The analog of AC's Assumption 3.4 for our $\Sigma_n(x_i,\alpha)$ defined in (64) is satisfied by AC's Corollary A.1(i). Thus Assumptions of AC's Lemma A.1 and Corollary A.1 are satisfied.

Lemma B.1 states the counterparts of their AC's Corollaries A.2 and B.1 for our case. We note that condition (A) of our consistency proof was shown to hold for $G_n(\alpha_n)$ in Theorem 4.1. Since $\widetilde{G}_n(\alpha_n) \leq G_n(\alpha_n)$, by (76) the condition also holds for $\widetilde{G}_n(\alpha_n)$. Thus the identification condition is satisfied. Satisfying Assumptions of Theorem 1 of Shen and Wong (1994) is also a necessary condition for AC's Theorem 3.1. Since the role of the pseudodistance in Theorem 1 of Shen and Wong (1994) is performed by our metric $\|\cdot\|_F^2$ in a way topologically equivalent to the AC's one, and the remaining AC's Assumptions hold as described above, this condition is also satisfied. Invocation of AC's Theorem 3.1, with their objective function and metric replaced with ours, completes the proof.

■

Proof of Theorem 6.1.

Substituting (56) into (55) yields

$$\frac{dg(Z,\alpha_0)}{d\alpha}[v^*] = D_{w^*}(Z)\left[E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau \tag{87}$$

Note that by the chain rule

$$\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[v^*] = \frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}\frac{dg\left(Z,\alpha_0\right)}{d\alpha}[v^*] \tag{88}$$

Using Lemma C.1 and (87) in (88), we obtain

$$\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}[v^*] = \frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle| X\right]\right\}\right]^{-1}\tau \tag{89}$$

We will now check the conditions for Theorem 7.1 in Appendix 3 that is an extension of Theorem 1 of Shen (1997) to our conditional case. Lemma C.2 shows that under our Assumptions, Conditions A is satisfied. Since $\{g\left(z,\alpha_n\right) : \alpha_n \in \mathcal{A}_n\} \subset \Lambda_c^{\gamma}(\mathcal{X})$, Condition B follows directly from Lemma B.1. Since $\|\widehat{\alpha}_n - \alpha_0\|_F = o_p(n^{-1/4})$, then $\delta_n = n^{-1/4}$ and hence for Condition C we require

$$\begin{aligned}\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n-\alpha_0\| \leq \delta_n\}} \|\varepsilon_n u^* - \varepsilon_n u_n^*\| &= O_p(\delta_n^{-1}\varepsilon_n^2) \\ &= O_p(n^{-1/4})\end{aligned}$$

which is satisfied by Assumption 6.2. Condition D follows from the smoothness of $\frac{d\varphi(x_i,z_j,\alpha_0)}{d\alpha}[\alpha-\alpha_0]$ in $\mathcal{N}_{0n}$. Condition F is satisfied by the definition of $f(\widehat{\alpha}_n) \equiv \tau'\widehat{\theta}_n$, $\omega = 1$, and Assumption 6.2. Condition G is satisfied by Assumption 6.1.

By Theorem 7.1 in Appendix 3, for arbitrarily fixed $\tau \in \mathbb{R}^{d_\theta}$ with $|\tau| \neq 0$,

$$\sqrt{n}\tau'(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma_{v^*})$$

where

$$\begin{aligned} \Sigma_{v^*} &\equiv E\left[Var\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}\middle| X\right)\right] \\ &= \tau'\Omega\tau \end{aligned} \tag{90}$$

and hence

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Omega)$$

Using (89) in (90) we obtain

$$\begin{aligned} \Omega &= \left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle| X\right]\right\}\right]^{-1} \\ &\quad \times E\left[Var\left(\frac{d\varphi(X,Z,\alpha_0)}{dg(Z,\alpha)}D_{w^*}(Z)\middle| X\right)\right] \\ &\quad \times \left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle| X\right]\right\}\right]^{-1} \end{aligned} \tag{91}$$

Using Lemma C.1 and (91)

$$\Omega = \left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle| X\right]\right\}\right]^{-1}$$

■

Appendix 2: Auxiliary Results

A. CONSISTENCY

Lemma A.1 (B.3) *Let Assumptions 4.5 and 4.7 hold. Then, pointwise for a given $\alpha \in \mathcal{A}$,*

$$\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\| = o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

where $\widetilde{\delta}_{1n}$ is defined in Lemma A.7 and ϱ in Assumption 4.7.

Proof. Decompose

$$\begin{aligned}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\| \quad \leq \quad & \max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\mathbb{I}_{i,n} \\ & +\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\max_{1\leq i\leq n}\mathbb{I}^{c}_{i,n}\end{aligned}$$

Note that he results of Lemma D.3 and D.5 in KTA hold also for w_{ij} as defined in this paper. Therefore

$$\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\max_{1\leq i\leq n}\mathbb{I}^{c}_{i,n} = o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

Next, pick any $\epsilon>0$, $c_n\downarrow 0$, and observe that

$$\Pr\left\{\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\mathbb{I}_{i,n}>\epsilon c_n\right\}\leq \Pr\left\{\sup_{X\in\mathcal{X}}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|>\epsilon c_n\right\}$$

Using Lemma A.7,

$$\Pr\left\{\sup_{X\in\mathcal{X}}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|>\epsilon c_n\right\}\leq\epsilon$$

if

$$c_n=\widetilde{\delta}_{1n}$$

where $\widetilde{\delta}_{1n}$ is defined in Lemma A.7. Hence

$$\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\mathbb{I}_{i,n} = o_p\left(\widetilde{\delta}_{1n}\right)$$

and the desired result follows. ■

Lemma A.2 (B.8) *Let Assumptions 4.5 and 4.7 hold. Then*

$$\sup_{\alpha_n\in\mathcal{A}_n}\left|Q_n(\alpha_n)-\overline{Q}_n(\alpha_n)\right| = o_p(n^{-1/m})$$

Proof. Substituting from (69) for $q_n\left(x_i,z_j,\alpha_n\right)$ we obtain

$$\begin{aligned} & n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}q_n\left(x_i,z_j,\alpha_n\right)+\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}u'\left(x_i,\alpha_n\right)E\left[g\left(z,\alpha_n\right)|x_i\right]\right| \\ \leq \quad & n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left\{-n^{-1/m}\left[a_i+u'\left(x_i,\alpha_n\right)g\left(z_j,\alpha_n\right)\right]\right\}+\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}u'\left(x_i,\alpha_n\right)E\left[g\left(z,\alpha_n\right)|x_i\right]\right| \\ & +n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}R_n(t,a_i,\alpha_n)\right|\end{aligned}$$

$$
\begin{aligned}
= & \sup_{\alpha_n \in \mathcal{A}_n} \left| -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} a_i + \frac{1}{n} \sum_{i=1}^{n} u'(x_i, \alpha) E[g(z, \alpha_n) | x_i] - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} u'(x_i, \alpha_n) g(z_j, \alpha_n) \right| \\
& + n^{1/m} \sup_{\alpha_n \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} R_n(t, a_i, \alpha_n) \right| \\
\leq & - \sup_{\alpha_n \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} a_i \right| + \sup_{\alpha_n \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^{n} \left\| E[g(z, \alpha_n) | x_i] - \sum_{j=1}^{n} w_{ij} g(z_j, \alpha_n) \right\| \\
& + n^{1/m} \sup_{\alpha_n \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} R_n(t, a_i, \alpha_n) \right|
\end{aligned}
$$

The first term drops out by Lemma A.4, the second term is $o_p(1)$ by Corollary A.1(i) in Ai and Chen (2003), p. 1824, and the third term is $o_p(1)$ by Lemma A.3. ■

Lemma A.3 *Let Assumptions 4.5 and 4.7 hold. Then*

$$
n^{1/m} \sup_{\alpha_n \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} R_n(t, a_i, \alpha_n) \right| = o_p(1)
$$

Proof. Note that by (70)

$$
\begin{aligned}
& n^{1/m} \sup_{\alpha_n \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} R_n(t, a_i, \alpha_n) \right| \\
\leq & \frac{1}{n^{1-1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sup_{\alpha_n \in \mathcal{A}_n} |R_n(t, a_i, \alpha_n)| \\
\leq & \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left[|a_i| + \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| \right] (1 - \mathbb{I}\{z_j \in C_n\}) \\
& + \frac{1}{n^{1+1/m}} \frac{1}{2(1-\widetilde{c})^2} \sum_{i=1}^{n} a_i^2 \sum_{j=1}^{n} w_{ij} \\
& + \frac{1}{n^{1+1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{\left[2a_i \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| + \sup_{\alpha \in \mathcal{A}} \|g(z_j, \alpha_n)\|^2 \right]}{2(1-\widetilde{c})^2} \\
= & D_1 + D_2 + D_3
\end{aligned}
$$

By Assumption 4.5(i) and 4.4(ii), $\sup_{\alpha_n \in \mathcal{A}_n} \|g(z, \alpha_n)\| < \infty$. By Lemma A.5 $|a_i| < \infty$ and hence by Lemma A.6

$$
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left[|a_i| + \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| \right] = O_p(1).
$$

Since $\max_{1 \leq j \leq n} \mathbb{I}\{z_j \notin C_n\} = o_p(1)$, $D_1 = o_p(1)$. By Lemma A.6 $D_2 = o_p(1)$.

$$
\begin{aligned}
D_3 = & \frac{1}{n^{1+1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{\left[2a_i \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| + \sup_{\alpha \in \mathcal{A}} \|g(z_j, \alpha_n)\|^2 \right]}{2(1-\widetilde{c})^2} \\
= & \frac{1}{n^{1+1/m}(1-\widetilde{c})^2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} a_i + \frac{1}{n^{1+1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{\sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\|^2}{2(1-\widetilde{c})^2}
\end{aligned}
$$

where the first part drops out by Lemma A.4 and the second part is $o_p(1)$ by Assumption 4.5(i), 4.4(ii) and Lemma A.6. ■

Lemma A.4 *Under Assumptions 4.3 and 4.4, for w_{ij} defined in (29) and a_i defined in (61), it holds that*

$$
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} a_i = 0
$$

Proof.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}a_i &= \frac{1}{n}\sum_{i=1}^{n} a_i \sum_{j=1}^{n} w_{ij} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} w_{ij} - 1\right]\sum_{j=1}^{n} w_{ij} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i) - \mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)p^{k_n}(x_i)'\left(P'P\right)^{-1}P'\mathbf{i} - \mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)\right] \\
&= \mathbf{i}'P\left(P'P\right)^{-1}\left(P'P\right)\left(P'P\right)^{-1}P'\mathbf{i} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i) \\
&= \frac{1}{n}\mathbf{i}'P\left(P'P\right)^{-1}P'\mathbf{i} - \frac{1}{n}\mathbf{i}'P\left(P'P\right)^{-1}P'\mathbf{i} \\
&= 0
\end{aligned}$$

■

Lemma A.5 *Under Assumptions 4.3 and 4.4, for w_{ij} defined in (29),*

$$\sum_{j=1}^{n} w_{ij} = O(1)$$

for each $X \in \mathcal{X}$.

Proof. By Assumption 4.3, for any $E\left[\rho_l\left(Z,\alpha\right)|x_i\right]$ there exists $p^{k_n}(x_i)'\pi_l = \sum_{j=1}^{n} w_{ij}g_l\left(z_j,\alpha\right)$ such that

$$E\left[E\left[g_l\left(Z,\alpha\right)|x_i\right] - \sum_{j=1}^{n} w_{ij}g_l\left(z_j,\alpha\right)\right] = O(1)$$

The result follows by boundedness of $g_l\left(z_j,\alpha\right)$. ■

Lemma A.6 *Under Assumptions 4.3 and 4.4, for w_{ij} defined in (29),*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = O_p(1)$$

Proof. Follows directly from Lemma A.5. ■

Lemma A.7 *Let*

$$\begin{aligned}
\xi_{0n} &\equiv \sup_{X\in\mathcal{X}}\left\|p^{k_n}(X)\right\|_E \\
\xi_{1n} &\equiv \sup_{X\in\mathcal{X}}\left\|\frac{\partial p^{k_n}(X)}{\partial x'}\right\|_E
\end{aligned}$$

Let $\widetilde{g}:\mathcal{Z}\to\mathbb{R}$ denote a generic measurable function of the data $Z\in\mathcal{Z}$, evaluated at a given fixed parameter α. Define $\varepsilon\left(Z,\alpha\right) = \widetilde{g}(Z,\alpha) - E\left[\widetilde{g}(Z,\alpha)|X\right]$ and $\varepsilon(\alpha) = \left(\varepsilon\left(Z_1,\alpha\right),...,\varepsilon\left(Z_n,\alpha\right)\right)'$.

Suppose that Assumptions 4.2 and 4.3(i) and the following are satisfied:

(i) There exists a constant c_{1n} and a measurable function $c_1(Z):\mathcal{Z}\to[0,\infty)$ with $E[c_1(Z)^p]<\infty$ for some $p\geq 4$ such that $|\widetilde{g}(Z,\alpha)|\leq c_{1n}c_1(Z)$ for all $Z\in\mathcal{Z}$;

(ii) There exists a positive value $\widetilde{\delta}_{1n} = o_p(1)$ such that

$$\frac{n\widetilde{\delta}_{1n}^2}{\ln\left[\left(\frac{\xi_{1n}c_{1n}}{\delta_{1n}}\right)^{d_x}\right]\max\left\{\xi_{0n}^2c_{1n}^2,\xi_{0n}^{2+2/p}\delta_{1n}^{1-2/p}c_{1n}^{1+2/p}\right\}}\to\infty$$

Then

$$p^{k_n}(X)'(P'P)^{-1}P'\varepsilon(\alpha) = o_p(\delta_{1n})$$

uniformly over $X \in \mathcal{X}$.

Proof. This result specializes Lemma A.1(A) in Ai and Chen (2003), derived for the combined space $\mathcal{X} \times \mathcal{A}$ to the space $\mathcal{X}$ only, with $g(z_j, \alpha)$ evaluated at a given fixed α. Since we do not have to account for growth restrictions on the parameter space, we are able to obtain faster convergence rate δ_{1n} than Ai and Chen (2003).

Let c denote a generic constant that may have different values in different expressions. For any pair $X_1 \in \mathcal{X}$ and $X_2 \in \mathcal{X}$

$$\begin{aligned} & \left|p^{k_n}(X_1)'(P'P)^{-1}P'\varepsilon(\alpha) - p^{k_n}(X_2)'(P'P)^{-1}P'\varepsilon(\alpha)\right| \\ = & \left|\left[p^{k_n}(X_1) - p^{k_n}(X_2)\right]'(P'P)^{-1}P'\varepsilon(\alpha)\right| \end{aligned}$$

Note that

$$\left\|p^{k_n}(X_1)' - p^{k_n}(X_2)'\right\|_E^2 \leq \xi_{1n}^2 \|X_1 - X_2\|_E^2$$

It follows that

$$\left|\left[p^{k_n}(X_1) - p^{k_n}(X_2)\right]'(P'P)^{-1}P'\varepsilon(\alpha)\right| \leq \xi_{1n}^2 \|X_1 - X_2\|_E^2 \sqrt{\frac{1}{n\lambda_n}\sum_{i=1}^n \varepsilon(Z_i, \alpha)^2}$$

where λ_n denotes the smallest eigenvalues of $P'P/n$. Condition (i) implies

$$\frac{1}{n}\sum_{i=1}^n \varepsilon(Z_i, \alpha)^2 \leq \frac{c_{1n}^2}{n}\sum_{i=1}^n \left(c_1(Z_i) + E\left[c_1(Z_i)|X_i\right]\right)^2$$

Assumption 4.3(i) implies $\lambda_n = O_p(1)$. Applying the weak law of large numbers and $E\left\{\left(E\left[c_1(Z_i)|X_i\right]\right)^2\right\} \leq E\left\{c_1(Z)^2\right\}$, we obtain

$$\frac{1}{n}\sum_{i=1}^n \left(c_1(Z_i) + E\left[c_1(Z_i)|X_i\right]\right)^2 = O_p(1)$$

Thus there exists a constant c such that

$$\Pr\left(\sqrt{\frac{1}{n\lambda_n}\sum_{i=1}^n \left(c_1(Z_i) + E\left[c_1(Z_i)|X_i\right]\right)^2} > c\right) < \eta$$

for sufficiently large n.

For any small ϵ partition $\mathcal{X}$ into b_n mutually exclusive subsets $\mathcal{X}_m$, $m = 1, ..., b_n$, where $X_1 \in \mathcal{X}_m$ and $X_2 \in \mathcal{X}_m$ imply $\|X_1 - X_2\|_E^2 \leq \epsilon\widetilde{\delta}_{1n}/(c_{1n}\xi_{1n}c)$. Then with probability approaching one we have

$$\left|p^{k_n}(X_1)'(P'P)^{-1}P'\varepsilon(\alpha) - p^{k_n}(X_2)'(P'P)^{-1}P'\varepsilon(\alpha)\right| \leq \epsilon\widetilde{\delta}_{1n}$$

Let X_m denote a fixed point in $\mathcal{X}_m$. For any X there exists an m such that $\|X_1 - X_2\|_E^2 \leq \epsilon\widetilde{\delta}_{1n}/(c_{1n}\xi_{1n}c)$. Then with probability approaching one

$$\sup_{X\in\mathcal{X}}\left|p^{k_n}(X)'(P'P)^{-1}P'\varepsilon(\alpha)\right| \leq \epsilon\widetilde{\delta}_{1n} + \max_m \left|p^{k_n}(X_m)'(P'P)^{-1}P'\varepsilon(\alpha)\right|$$

Hence

$$\begin{aligned} & \Pr\left(\sup_{X\in\mathcal{X}}\left|p^{k_n}(X)'(P'P)^{-1}P'\varepsilon(\alpha)\right| > 2\epsilon\widetilde{\delta}_{1n}\right) \\ < & 2\eta + \Pr\left(\max_m \left|p^{k_n}(X_m)'(P'P)^{-1}P'\varepsilon(\alpha)\right| > 2\epsilon\widetilde{\delta}_{1n}\right) \end{aligned}$$

For some constant c, let

$$M_n = \left(\frac{c\xi_{0n} c_{1n}}{\delta_{1n} \epsilon \eta} \right)^{2/p}$$

Define $d_{in} = \mathbb{I}\{c_1(Z) \leq M_n\}$. Define $g_1(Z_i, \alpha) = d_{in} g_1(Z_i, \alpha)$ and $g_2(Z_i, \alpha) = (1 - d_{in})\, g_1(Z_i, \alpha)$. Define $\varepsilon_1(Z_i, \alpha)$ and $\varepsilon_2(Z_i, \alpha)$ accordingly. It follows that

$$\begin{aligned}
& \Pr\left(\max_m \left| p^{k_n}(X_m)'(P'P)^{-1} P' \varepsilon(\alpha) \right| > 2\epsilon \widetilde{\delta}_{1n} \right) \\
\leq \quad & \Pr\left(\max_m \left| p^{k_n}(X_m)'(P'P)^{-1} \sum_{i=1}^n \varepsilon_1(Z_i, \alpha) \right| > \epsilon \widetilde{\delta}_{1n} \right) \\
& + \Pr\left(\max_m \left| p^{k_n}(X_m)'(P'P)^{-1} \sum_{i=1}^n \varepsilon_2(Z_i, \alpha) \right| > \epsilon \widetilde{\delta}_{1n} \right) \\
\equiv \quad & P_1 + P_2
\end{aligned}$$

Ai and Chen (2003) show that $P_2 \leq \eta$, along with

$$\sigma_m^2 \equiv nE\left\{ \left[p^{k_n}(X_m)'(P'P)^{-1} \sum_{i=1}^n p^{k_n}(X_i) \varepsilon_1(Z_i, \alpha) \right]^2 \right\} = O(c_{1n}^2 \xi_{0n}^2)$$

and

$$\left| p^{k_n}(X_m)'(P'P/n)^{-1} p^{k_n}(X_i) \varepsilon_1(Z_i, \alpha) \right| \leq \frac{M_n \xi_{0n}^2 c_{1n}}{\lambda_n}$$

Noting that

$$\begin{aligned}
& \Pr\left(\left| p^{k_n}(X_m)'(P'P)^{-1} \sum_{i=1}^n \varepsilon_1(Z_i, \alpha) \right| > \epsilon \widetilde{\delta}_{1n} \right) \\
= \quad & E\left[\Pr\left(\left| p^{k_n}(X_m)'(P'P)^{-1} \sum_{i=1}^n \varepsilon_1(Z_i, \alpha) \right| > \epsilon \widetilde{\delta}_{1n} \mid X_1, ..., X_n \right) \right]
\end{aligned}$$

Ai and Chen (2003) apply the Bernstein inequality for independent processes to obtain

$$\begin{aligned}
& \Pr\left(\left| p^{k_n}(X_m)'(P'P)^{-1} \sum_{i=1}^n \varepsilon_1(Z_i, \alpha) \right| > \epsilon \delta_{1n} \right) \\
\leq \quad & 2E\left[\exp\left(-n\varepsilon^2 \widetilde{\delta}_{1n}^2 / \left(c\sigma_m^2 + M_n \xi_{0n}^2 c_{1n}^2 \lambda_n^{-1} \varepsilon \widetilde{\delta}_{1n} \right) \right) \right]
\end{aligned}$$

where $E[\cdot]$ is taken with respect to the joint distribution of $(X_1, ..., X_n)$. Hence

$$P_1 < 2b_n E\left[\exp\left(-n\varepsilon^2 \widetilde{\delta}_{1n}^2 / \left(c\sigma_m^2 + M_n \xi_{0n}^2 c_{1n}^2 \lambda_n^{-1} \varepsilon \widetilde{\delta}_{1n} \right) \right) \right]$$

which is arbitrarily small if

$$\frac{n\widetilde{\delta}_{1n}^2}{\max\left\{ \xi_{0n}^2 c_{1n}^2, M_n \xi_{0n}^2 c_{1n} \widetilde{\delta}_{1n} \right\}} - \ln(b_n) \to \infty$$

Since $\mathcal{X}$ is a compact subset in $\mathbb{R}^d$, we have

$$b_n = O\left(\left(\frac{\widetilde{\delta}_{1n}}{c_{1n} \xi_{1n}} \right)^{-d_x} \right)$$

Substituting for M_n and b_n we obtain

$$\begin{aligned}
&\frac{n\widetilde{\delta}_{1n}^2}{\ln(b_n)\max\left\{\xi_{0n}^2 c_{1n}^2, M_n\xi_{0n}^2 c_{1n}\widetilde{\delta}_{1n}\right\}} \\
&= O\left(\frac{n\widetilde{\delta}_{1n}^2}{\ln\left[\left(\frac{\widetilde{\delta}_{1n}}{c_{1n}\xi_{1n}}\right)^{-d_x}\right]\max\left\{\xi_{0n}^2 c_{1n}^2, \xi_{0n}^{2+2/p}\widetilde{\delta}_{1n}^{1-2/p} c_{1n}^{1+2/p}\right\}}\right)
\end{aligned}$$

Thus, for $P_1 < \eta$ for sufficiently large n by condition (ii). ■

Lemma A.8 (part of B.1) *Let Assumptions 4.2-4.6 and 4.8 hold. Let also $n^{1/m}\widetilde{\delta}_{1n} \downarrow 0$ and $\rho > 2/m$ where $\widetilde{\delta}_{1n}$ is defined in Lemma A.7 and ϱ in Assumption 4.7. Then*

$$\max_{1\leq i\leq n}\|\lambda_i(\alpha_0)\| = o_p(\widetilde{\delta}_{1n}) + o_p\left(\frac{1}{n^{\varrho-1/m}}\right) \tag{92}$$

This Lemma is analogous to Lemma B.1 of KTA. However, the analysis is somewhat complicated due to the extra term σ_i. Moreover, here we do not make use of results related to kernel estimation. Thus, for example, consistency of the variance-covariance matrix $\Sigma_n(x_i, \alpha_0)$ follows from series results of Ai and Chen (2003).

Proof. In this Lemma, we will use the F.O.C.s (22) and (24) that combine to

$$\begin{aligned}
\sum_{j=1}^{n}\frac{w_{ij}}{1+a_i+\lambda_i' g\left(x_j,\alpha\right)} &= \sum_{j=1}^{n}\frac{w_{ij}}{\lambda_i' g\left(x_j,\alpha\right)+\sigma_i} \\
&= \sum_{j=1}^{n}\widehat{\pi}_{ij} \\
&= 1
\end{aligned} \tag{93}$$

Let

$$\lambda_i\left(\alpha_0\right) = \rho_i\xi_i \tag{94}$$

where $\rho_i \geq 0$ and $\xi_i \in \mathbb{R}^{d_g}$. It holds that

$$\begin{aligned}
\sum_{j=1}^{n} w_{ij}\frac{\left[a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)\right]^2}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} &= a_i^2\sum_{j=1}^{n}\frac{w_{ij}}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} + \frac{2a_i\rho_i\sum_{j=1}^{n}w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} \\
&\quad + \frac{\rho_i^2\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)}
\end{aligned} \tag{95}$$

For the first term of the RHS sum of (95), using (93), it holds that

$$\begin{aligned}
a_i^2\sum_{j=1}^{n}\frac{w_{ij}}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} &= a_i^2 \\
&= \left(\sigma_i-1\right)^2 \\
&= \sigma_i^2-2\sigma_i+1
\end{aligned} \tag{96}$$

Substituting (96) into (95) yields

$$\begin{aligned}
\sum_{j=1}^{n} w_{ij}\frac{\left[a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)\right]^2}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} &= \sigma_i^2-2\sigma_i+1+\frac{2a_i\rho_i\sum_{j=1}^{n}w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} \\
&\quad + \frac{\rho_i^2\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)}
\end{aligned} \tag{97}$$

Note that for a generic constant c

$$\begin{aligned}\frac{c^2}{1+c} &= \frac{c^2}{1+c} + (1-c) - (1-c) \\ &= \frac{c^2}{1+c} + \frac{(1-c)(1+c)}{1+c} - (1-c) \\ &= \frac{c^2}{1+c} + \frac{1-c^2}{1+c} - (1-c) \\ &= \frac{1}{1+c} - 1 + c\end{aligned}$$

Using this fact, letting $c = a_i + \lambda_i'(\alpha_0) g(z_j, \alpha_0)$, we have

$$\begin{aligned}\sum_{j=1}^n w_{ij} \frac{[a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)]^2}{1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)} &= \sum_{j=1}^n w_{ij} \left\{ \frac{1}{1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)} - 1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0) \right\} \\ &= \sum_{j=1}^n \frac{w_{ij}}{1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)} - \sum_{j=1}^n w_{ij} + \sum_{j=1}^n w_{ij} a_i \\ &\quad + \sum_{j=1}^n w_{ij} \lambda_i'(\alpha_0) g(z_j,\alpha_0) \\ &= 1 - \sum_{j=1}^n w_{ij} + \sum_{j=1}^n w_{ij} a_i + \sum_{j=1}^n w_{ij} \lambda_i'(\alpha_0) g(z_j,\alpha_0) \qquad (98)\end{aligned}$$

By the definition of σ_i,

$$\begin{aligned}1 - \sum_{j=1}^n w_{ij} + a_i \sum_{j=1}^n w_{ij} &= 1 - \sigma_i + (\sigma_i - 1)\sigma_i \\ &= \sigma_i^2 - 2\sigma_i + 1 \qquad (99)\end{aligned}$$

Substituting (99) into (98) gives us

$$\sum_{j=1}^n w_{ij} \frac{[a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)]^2}{1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)} = \sigma_i^2 - 2\sigma_i + 1 + \rho_i \sum_{j=1}^n w_{ij} \xi_i' g(z_j,\alpha_0) \qquad (100)$$

Combining (97) and (100) yields, after canceling $\sigma_i^2 - 2\sigma_i + 1$ from both sides,

$$\frac{2a_i\rho_i \sum_{j=1}^n w_{ij}\xi_i' g(z_j,\alpha_0)}{1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)} + \frac{\rho_i^2 \xi_i' \Sigma_n(x_i,\alpha_0)\xi_i}{1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0)} = \rho_i \sum_{j=1}^n w_{ij} \xi_i' g(z_j,\alpha_0) \qquad (101)$$

Using Assumption 4.8, by Lemma D.2 in KTA,

$$\max_{1 \le j \le n} \|g(z_j,\alpha_0)\| = o_p(n^{1/m}) \qquad (102)$$

and this $o_p(n^{1/m})$ term does not depend on i, j, or $\alpha_n \in \mathcal{A}_n$. By (102) it holds that

$$0 \le 1 + a_i + \lambda_i'(\alpha_0) g(z_j,\alpha_0) \le 1 + a_i + \rho_i \|g(z_j,\alpha_0)\| = 1 + a_i + \rho_i o_p(n^{1/m}) \qquad (103)$$

Using (103) in (101) and canceling ρ_i yields

$$\frac{2a_i \sum_{j=1}^n w_{ij}\xi_i' g(z_j,\alpha_0)}{1 + a_i + \rho_i o_p(n^{1/m})} + \frac{\rho_i \xi_i' \Sigma_n(x_i,\alpha_0)\xi_i}{1 + a_i + \rho_i o_p(n^{1/m})} \le \sum_{j=1}^n w_{ij} \xi_i' g(z_j,\alpha_0) \qquad (104)$$

By Corollary D.1 of Ai and Chen (2003), $\Sigma_n(x_i,\alpha_0) = \Sigma(x_i,\alpha_0) + o_p(1)$ uniformly over $X \in \mathcal{X}$. Using the fact that $\xi_i'\Sigma(x_i,\alpha_0)\xi_i$ is bounded away from zero on $(x_i,\xi_i) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_g}$, we can divide (104) by

$\frac{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}{1+a_i+\rho_i o_p(n^{1/m})}$ and rearrange terms to obtain

$$\begin{aligned}\rho_i &\leq \left[1+a_i+\rho_i o_p(n^{1/m})\right]\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} - 2a_i\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\\ &= (1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} + \rho_i o_p(n^{1/m})\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\end{aligned}$$

and hence

$$\begin{aligned}\rho_i\left(1-o_p(n^{1/m})\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\right) &\leq (1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\\ \rho_i &\leq (1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\\ &\quad\times\left(1-o_p(n^{1/m})\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\right)^{-1} \qquad (105)\end{aligned}$$

For the last term of the RHS of (105), using Lemma A.1 and $\|\xi_i'\|<\infty$ for all i, it holds that

$$\begin{aligned}o_p(n^{1/m})\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} &= o_p(n^{1/m})\left\|\xi_i'\right\|\max_{1\leq i\leq n}\left\|\sum_{j=1}^n w_{ij} g\left(z_j,\alpha_0\right)\right\|\\ &= o_p(n^{1/m})O(1)\left[o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)\right]\\ &= o_p(n^{1/m}\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-2/m}}\right) \qquad (106)\end{aligned}$$

while for the first term of the RHS of (105), using also Lemma A.5,

$$\begin{aligned}(1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} &= O(1)\left\|\xi_i'\right\|\max_{1\leq i\leq n}\left\|\sum_{j=1}^n w_{ij} g\left(z_j,\alpha_0\right)\right\|\\ &= O(1)O(1)\left[o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)\right]\\ &= o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right) \qquad (107)\end{aligned}$$

Under our assumptions, $n^{1/m}\widetilde{\delta}_{1n}\downarrow 0$ and $n^{-\varrho+2/m}\downarrow 0$ in (106). This used in (105) along with (107) and consistency of $\Sigma_n(x_i,\alpha_0)$, implies that

$$\max_{1\leq i\leq n}\|\rho_i\| = o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

which yields the desired result by the definition of ρ_i in (94). ■

B: CONVERGENCE RATES

Lemma B.1 *Consider the functions $G_n(\alpha_n)$ and $\overline{G}_n(\alpha_n)$ defined in (63) and (86), respectively. Assumptions 4.1-4.3, 4.5, 4.6, 5.1-5.6 imply: (i) $G_n(\alpha_n) - \overline{G}_n(\alpha_n) = o_p(n^{-1/4})$ uniformly over $\alpha_n \in \mathcal{A}_n$; and (ii) $G_n(\alpha_n) - G_n(\alpha_0) - \left\{\overline{G}_n(\alpha_n) - \overline{G}_n(\alpha_0)\right\} = o_p(\eta_n n^{-1/4})$ uniformly over $\alpha_n \in \mathcal{A}_n$ with $\|\alpha_n - \alpha_0\|_F \leq o(\eta_n)$, where $\eta_n = n^{-\tau}$ with $\tau \leq 1/4$.*

Proof.

This Lemma shows the counterpart of AC's Corollary B.1 for our case. Since $\lambda_i(\alpha_n)$ solves

$$\sum_{j=1}^{n} \frac{w_{ij} g\left(z_j, \alpha_n\right)}{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)} = 0 \tag{108}$$

denote by $\lambda_{i0}(\alpha_n)$ the solution to

$$E\left[\left.\frac{g\left(z_j, \alpha_n\right)}{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)}\right| x_i\right] = 0$$

For the sake of brevity, let "VW" stand for "Van der Vaart and Wellner (1996)." Lemma A.5 and Assumption 4.5(i) suffice to satisfy the pointwise convergence condition of Lemma 3.3.5 (p. 311) in VW for the objective function (108). Note that $\{g\left(z, \alpha_n\right) : \alpha_n \in \mathcal{A}_n\} \subset \Lambda_c^{\overline{J}}(\mathcal{X})$ and $\Lambda_c^{\overline{J}}(\mathcal{X})$ is a Donsker class by Theorem 2.5.6 in VW. Since $\lambda_i\left(\alpha_n\right) \in R^{d_g}$, $\{\lambda_i\left(\alpha_n\right) : \alpha_n \in \mathcal{A}_n\}$ belongs to the Donsker class. By Example 2.10.8 (p. 192) in VW $\{\lambda_i' g\left(z, \alpha_n\right) : \alpha_n \in \mathcal{A}_n\}$ is Donsker. Since $0 < \sigma_i < \infty$ is a data-determined scalar by Lemma A.5, by Example 2.10.9 (p. 192) in VW (108) is Donsker in $\alpha_n \in \mathcal{A}_n$. Hence the Assumptions of Lemma 3.3.5 (p. 311) in VW are satisfied and we can invoke Theorem 3.3.1 (p. 310) in VW to conclude that $\|\lambda_i(\alpha_n) - \lambda_{i0}(\alpha_n)\|_E = O_p(n^{-1/2})$, uniformly over $\alpha_n \in \mathcal{A}_n$, for each i. Lemma A.1(A) of Ai and Chen (2003) (defining δ_{1n}) states that $\sum_{j=1}^{n} w_{ij} g\left(z_j, \alpha_n\right) - m\left(x_i, \alpha_n\right) = o_p(\delta_{1n})$ uniformly over $\mathcal{X} \times \mathcal{A}_n$. These two rate results for $\lambda_i(\alpha_n)$ and $g\left(z_j, \alpha_n\right)$, simple law of large numbers for σ_i and continuity of the log function satisfy the satisfy the pointwise convergence condition of Lemma 3.3.5 (p. 311) in VW for the objective function $G_n(\alpha_n)$. By Theorem 2.10.6 (p. 192) in VW $\{\ln[\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)] : \alpha_n \in \mathcal{A}_n\}$ is Donsker. By Lemma A.5, $0 < \sigma_i < \infty$ for each i and thus we can renormalize σ_i by dividing by $\sup_{1 \leq i \leq n} \sigma_i$ that guarantees $\sum_{i=1}^{n} \sigma_i < 1$. By Theorem 2.10.3 (p. 190) in VW

$$\begin{aligned} \left|G_n(\alpha_n) - \overline{G}_n(\alpha_n)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \ln\left\{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)\right\} - \frac{1}{n}\sum_{i=1}^{n} E\left[\ln\left\{\sigma_i + \lambda_{i0}' g\left(z, \alpha_n\right)\right\} |x_i\right]\right| \\ &= O_p(n^{-1/2}) \end{aligned}$$

uniformly over $\alpha_n \in \mathcal{A}_n$, which shows the result (i) in this Lemma.

In order to show part (ii) of the proof, we first derive the counterpart of AC's Corollary A.2 that is a building block for their Corollary B.1 (ii). Note that since $m(X, \alpha_0) = 0$, $\|\alpha_n - \alpha_0\|_F = o_p(1)$ and AC's result $(i.1)$ of the proof of their Corollary A.2 holds also for our $\|m(X, \alpha)\|_E^2$, we only need to show the counterpart of their part $(i.2)$. We replace Assumption 3.9 of AC by our Assumption 5.1 which applies to our metric $\|\cdot\|_F$. This Assumption together with Lemma C.1 imply that $E\{\|m(X, \alpha)\|_E^2\}$ and $\|\alpha - \alpha_0\|_F^2$ are (topologically) equivalent. Then by Assumptions 4.1, 5.1, and $5.3(i)$, we have

$$E\left\{\left[\|m(X, \alpha)\|_E^2\right]^2\right\} \leq E\left\{\|m(X, \alpha)\|_E^2\right\} \times \left[\sup_{X, \alpha}\left\{\|m(X, \alpha)\|_E\right\}\right]^2 \leq const. \times \|\alpha_n - \alpha_0\|_F^2$$

satisfying part $(i.2)$. Part (ii) of AC's Corollary A.2 holds for our metric $\|\cdot\|_F$ by replacing their Assumption 3.9 with our Assumption 5.1. This, along with AC's Corollary A.1 shows (ii). ■

C: ASYMPTOTIC NORMALITY

Lemma C.1 *Under Assumptions 4.1-5.6,*

$$\begin{aligned} & E\left[Var\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\middle| X\right)\right] \\ = \; & E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\} \\ = \; & E\left\{E\left[D_w(Z)'\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\right)'D_w(Z)\middle| X\right]\right\} \end{aligned}$$

Proof. Using (51) and (49)

$$\begin{aligned} E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\middle| X\right] &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}\frac{dg(Z,\alpha_0)}{d\alpha}\left[v^*\right]\middle| X\right] \\ &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}\left[v^*\right]\middle| X\right] \\ &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\theta'}\left(u_\theta^*-\theta_0\right)+\frac{d\varphi\left(X,Z,\alpha_0\right)}{dh}\left[u_h^*-h_0\right]\middle| X\right] \\ &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\theta'}\middle| X\right]\left(u_\theta^*-\theta_0\right)+E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{dh}\left[u_h^*-h_0\right]\middle| X\right] \\ &= 0 \end{aligned}$$

by the definition of α_0. Hence

$$Var\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\middle| X\right)=E\left[D_{w^*}(Z)'\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\right)'D_{w^*}(Z)\middle| X\right]$$

Taking expectation over X yields the required result. ■

Lemma C.2 *Consider the notation for $v_n(\cdot)$ and $\widetilde{r}[\cdot]$ defined in Appendix 3. Then, under Assumptions 4.1-5.6,*

$$n^{-1/2}v_n\left(\widetilde{r}[\alpha_n-\alpha_0,X,Y]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,X,Y]\right)=o_p(n^{-1/4})$$

Proof. This Lemma performs a similar function as Lemmas C.1 - C.3 in Ai and Chen (2003). By the definition of $v_n(\cdot)$ and $\widetilde{r}[\cdot]$,

$$\begin{aligned} & n^{-1/2}v_n\left(\widetilde{r}[\alpha_n-\alpha_0,X,Y]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,X,Y]\right) \\ = \; & n^{-1}\sum_{i=1}^n\sum_{j=1}^n\left(\begin{array}{c} w_{ij}\left\{\widetilde{r}[\alpha_n-\alpha_0,x_i,y_j]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,x_i,y_j]\right\} \\ -E\left\{\widetilde{r}[\alpha_n-\alpha_0,X,Y]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,X,Y]\right\}\end{array}\right) \\ = \; & A_1-A_2 \end{aligned}$$

$$\begin{aligned} A_1 &= n^{-1}\sum_{i=1}^n\sum_{j=1}^n w_{ij}\widetilde{r}[\alpha_n-\alpha_0,x_i,y_j]-E\widetilde{r}[\alpha_n-\alpha_0,X,Y] \\ A_2 &= n^{-1}\sum_{i=1}^n\sum_{j=1}^n w_{ij}\widetilde{r}[\alpha_n+\varepsilon_nu_n^*-\alpha_0,x_i,y_j]-E\widetilde{r}[\alpha_n+\varepsilon_nu_n^*-\alpha_0,X,Y] \end{aligned}$$

$$\begin{aligned} A_1 &= A_{11}-A_{12} \\ A_{11} &= n^{-1}\sum_{i=1}^n\sum_{j=1}^n w_{ij}\varphi\left(x_i,z_j,\alpha\right)-E\varphi\left(z,x,\alpha\right) \\ A_{12} &= n^{-1}\sum_{i=1}^n\sum_{j=1}^n w_{ij}\frac{d\varphi\left(x_i,z_j,\alpha_0\right)}{d\alpha}[\alpha-\alpha_0]-E\left\{\frac{d\varphi\left(x,z,\alpha_0\right)}{d\alpha}[\alpha-\alpha_0]\right\} \end{aligned}$$

$$
\begin{aligned}
A_2 &= A_{21} - A_{22} \\
A_{21} &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \varphi\left(x, z, \alpha_n + \varepsilon_n u_n^*\right) - E\varphi\left(x, z, \alpha_n + \varepsilon_n u_n^*\right) \\
A_{22} &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{d\varphi\left(x_i, z_j, \alpha_0\right)}{d\alpha} [\alpha_n + \varepsilon_n u_n^* - \alpha_0] - E\left\{ \frac{d\varphi\left(x, z, \alpha_0\right)}{d\alpha} [\alpha_n + \varepsilon_n u_n^* - \alpha_0] \right\}
\end{aligned}
$$

The goal is to show $A_{11} - A_{12} - A_{21} + A_{22} = O_p(\varepsilon_n^2) = o_p(n^{-1/4})$. Note that $A_{11} = o_p(n^{-1/4})$ and $A_{21} = o_p(n^{-1/4})$ follows from parts A and B of AC's Lemma A.1. $A_{12} = o_p(n^{-1/4})$ and $A_{22} = o_p(n^{-1/4})$ follows from the rate results for A_{11} and A_{21}, respectively, and the continuous mapping theorem. ■

Appendix 3

In this Appendix we extend Theorem 1 of Shen (1997) to our conditional case.[10] Consider the setup as in Shen (1997), with the following modifications. Suppose that the observations $\{(X_i, Y_j) : i, j = 1, ..., n\}$ are drawn independently distributed according to density $p(\alpha_0, X_i, Y_j)$.

Define

$$K(\alpha_0, \alpha) = E_0 l(\alpha_0, X_i, Y_j) - E_0 l(\alpha, X_i, Y_j)$$

Let the empirical criterion be

$$L_n(\alpha) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} l(\alpha, X_i, Y_j)$$

where $l(\alpha, Y_j, X_i)$ is the criterion based on a single observation. Consider $l(\alpha, x, y)$ for which (*analog of Shen's (4.1)*)

$$\widetilde{r}[\alpha - \alpha_0, x, y] = l(\alpha, x, y) - l(\alpha_0, x, y) - l'_{\alpha_0}[\alpha - \alpha_0, x, y] \tag{S 4.1}$$

where $l'_{\alpha_0}[\alpha - \alpha_0, x, y]$ is defined as $\lim_{t \to 0}[l(a + t[\alpha - \alpha_0], x, y) - l(\alpha_0, x, y)]/t$. Denote $\widehat{\alpha}_n$ the maximizer of $L_n(\alpha_n)$ over $\alpha_n \in \mathcal{A}_n$. We estimate a real functional of $\widehat{\alpha}_n$ denoted as $f(\alpha)$. With $\widehat{\alpha}_n$ as defined, $f(\alpha)$ is estimated by a substitution estimate $f(\widehat{\alpha}_n)$. By the definition of $\widehat{\alpha}_n$, we have (*analog of Shen's (2.1)*)

$$L_n(\widehat{\alpha}_n) \geq \sup_{\alpha \in \mathcal{A}_n} L_n(\alpha_n) - O(\varepsilon_n^2) \tag{S 2.1}$$

where $\varepsilon_n^2 \to 0$ as $n \to \infty$. For any generic function $g(X, Y)$ let

$$\nu_n(g) = n^{-1} \sum_{i=1}^{n} n^{1/2} \left\{ \sum_{j=1}^{n} w_{ij} g(X_i, Y_j) - E\left[g(X, Y) | \, X = x_i \right] \right\}$$

be the empirical process induced by g. Let the convergence rate of the sieve estimate under $\|\cdot\|$ be $o_p(\delta_n)$ and let $\varepsilon_n^2 = o_p(n^{-1/2})$.

The following conditions are modified versions of Shen's 1997 (p. 2568) conditions:

Condition A (Stochastic Equicontinuity) *For $\widetilde{r}[\alpha - \alpha_0, x, y]$ defined in (S 4.1),*

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} n^{-1/2} \nu_n \left(\widetilde{r}[\alpha_n - \alpha_0, X, Y] - \widetilde{r}[\alpha_n + \varepsilon_n u_n^* - \alpha_0, X, Y] \right) = O_p(\varepsilon_n^2)$$

Condition B (Expectation of Criterion Difference)

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} \left[K\left(\alpha_0, \alpha_n + \varepsilon_n u_n^*\right) - K\left(\alpha_0, \alpha_n\right) \right] - \frac{1}{2} \left[\left\| \alpha_n + \varepsilon_n u^* - \alpha_0 \right\|^2 - \left\| \alpha_n - \alpha_0 \right\|^2 \right] = O_p(\varepsilon_n^2)$$

Condition C (Approximation Error)

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} \left\| \varepsilon_n u^* - \varepsilon_n u_n^* \right\| = O_p(\delta_n^{-1} \varepsilon_n^2)$$

In addition,

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} n^{-1/2} \nu_n \left(l'_{\alpha_0}[\varepsilon_n u^* - \varepsilon_n u_n^*, X, Y] \right) = O_p(\varepsilon_n^2)$$

Condition D (Gradient)

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} n^{-1/2} \nu_n \left(l'_{\alpha_0}[\alpha_n - \alpha_0, X, Y] \right) = O_p(\varepsilon_n)$$

Condition E (Smoothness)

Suppose the functional f has the following smoothness property: for any $\alpha_n \in \mathcal{A}_n$

$$\left| f_{\alpha_n} - f_{\alpha_0} - f'_{\alpha_0}[\alpha_n - \alpha_0] \right| \leq u_n \left\| \alpha_n - \alpha_0 \right\|_F^{\omega} \tag{S 4.2}$$

[10] Measurability with respect to the underlying probability space is assumed throughout the paper and hence we do not distiguish outer expectation from the usual one.

as $\|\alpha_n - \alpha_0\|_F \to 0$ *where* ω *is the degree of smoothness of* $f'_{\alpha_0}[\alpha_n - \alpha_0]$ *at* α_0.

Condition F (Convergence Rates and Smoothness) $u_n \delta_n^{\omega} = O_p(n^{-1/2})$.

Condition G (Variance) $Var\left(l'_{\alpha_0}[v^*, X, Y]\right) < \infty$ *is positive definite for all* $X \in \mathcal{X}$, $y \in \mathcal{Y}$.

Theorem 7.1 *Let the Conditions A-G hold. Then for the approximate substitution sieve estimate defined in (S 2.1),*

$$n^{-1/2}(f(\widehat{\alpha}_n) - f(\alpha_0)) \xrightarrow{d} N(0, E\left[Var\left(l'_{\alpha_0}[v^*, Y]\right)|X\right])$$

Proof of Theorem 7.1. Rearrange (S 4.1) as

$$l(\alpha, x, y) = \widetilde{r}[\alpha - \alpha_0, x, y] + l(\alpha_0, x, y) + l'_{\alpha_0}[\alpha - \alpha_0, x, y]$$

Subtract from (S 4.1) its expectation (under $P(\theta_0, X_i, Y_j)$ denoted by E_0), for a given (X_i, Y_j) to obtain

$$\begin{aligned} l(\alpha, x_i, y_j) - E_0 l(\alpha, x_i, y_j) &= l(\alpha, x_i, y_j) - E_0 l(\alpha, x_i, y_j) \\ &\quad + l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] - E_0 l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] \\ &\quad + \widetilde{r}[\alpha - \alpha_0, x_i, y_j] - E_0 \widetilde{r}[\alpha - \alpha_0, x_i, y_j] \end{aligned}$$

rearrange

$$\begin{aligned} l(\alpha, x_i, y_j) &= l(\alpha, x_i, y_j) - [E_0 l(\alpha, x_i, y_j) - E_0 l(\alpha, x_i, y_j)] \\ &\quad + l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] - E_0 l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] \\ &\quad + \widetilde{r}[\alpha - \alpha_0, x_i, y_j] - E_0 \widetilde{r}[\alpha - \alpha_0, x_i, y_j] \end{aligned}$$

take a weighted average over i, j with weights w_{ij}

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} l(\alpha, x_i, y_j) &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} l(\alpha_0, x_i, y_j) \\ &\quad - n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left[E_0 l(\alpha_0, x_i, y_j) - E_0 l(\alpha, x_i, y_j)\right] \\ &\quad + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] - E_0 l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j]\right) \\ &\quad + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(\widetilde{r}[\alpha - \alpha_0, x_i, y_j] - E_0 \widetilde{r}[\alpha - \alpha_0, x_i, y_j]\right) \end{aligned}$$

and hence using the notation above, for any $P_n\alpha_n \in \{P_n\alpha_n \in \mathcal{A}_n : \|P_n\alpha_n - \alpha_0\| \leq \delta_n\}$, we have

$$\begin{aligned} L_n(P_n\alpha_n) &= L_n(a_0) - K(\alpha_0, P_n\alpha_n) \\ &\quad + n^{-1/2} \nu_n(l'_{\theta_0}[P_n\alpha_n - \alpha_0, X, Y]) \\ &\quad + n^{-1/2} \nu_n(r[P_n\alpha_n - \alpha_0, X, Y]) \end{aligned} \tag{S 9.1}$$

Substituting $P_n\alpha_n$ by $\widehat{\alpha}_n$ here above, we obtain

$$\begin{aligned} L_n(\widehat{\alpha}_n) &= L_n(a_0) - K(\alpha_0, \widehat{\alpha}_n) \\ &\quad + n^{-1/2} \nu_n(l'_{\theta_0}[\widehat{\alpha}_n - \alpha_0, X, Y]) \\ &\quad + n^{-1/2} \nu_n(r[\widehat{\alpha}_n - \alpha_0, X, Y]) \end{aligned} \tag{S 9.2}$$

Subtracting (S 9.2) from (S 9.1) and substituting α_n by $\alpha^*(\widehat{\alpha}_n, \varepsilon_n)$ in (S 9.1), we have

$$\begin{aligned} & L_n(P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)) - L_n(\widehat{\alpha}_n) \\ = \; & L_n(\alpha_0) - L_n(\alpha_0) \\ & -K(\theta_0, P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n) + K(\alpha_0,\widehat{\alpha}_n) \\ & +n^{-1/2}\nu_n(l'_{\alpha_0}[P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n) - \alpha_0, X, Y]) - n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - \alpha_0, X, Y]) \\ & +n^{-1/2}\nu_n(r[P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n) - \alpha_0, X, Y]) - n^{-1/2}\nu_n(r[\widehat{\alpha}_n - \alpha_0, X, Y]) \end{aligned}$$

which yields

$$\begin{aligned} L_n(\widehat{\alpha}_n) \;=\; & L_n(P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)) \\ & -[K(\alpha_0,\widehat{\alpha}_n) - K(\theta_0, P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)] \\ & +n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n), X, Y]) \\ & +n^{-1/2}\nu_n(r[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n), X, Y]) \end{aligned}$$

By Condition A (second line of the following)

$$\begin{aligned} & n^{-1/2}\nu_n(r[P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n) - \alpha_0, X, Y]) - n^{-1/2}\nu_n(r[\widehat{\alpha}_n - \alpha_0, X, Y]) \\ = \; & n^{-1/2}\nu_n(r[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n), X, Y]) \\ = \; & O_p(\varepsilon_n^2) \end{aligned}$$

Using Condition B on the difference in Ks, we obtain

$$\begin{aligned} L_n(\widehat{\alpha}_n) \;=\; & L_n(P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)) - \frac{1}{2}\left[\|\widehat{\alpha}_n - \alpha_0\|^2 - \|P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n) - \alpha_0\|^2\right] \\ & +n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n), X, Y]) \\ & +O_p(\varepsilon_n^2) \end{aligned}$$

By Condition C (applicable to the second line)

$$\|P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n) - \alpha^*(\widehat{\alpha}_n,\varepsilon_n)\| = O(\delta_n^{-1}\varepsilon_n^2)$$

Hence, using (S 2.1) we have

$$\begin{aligned} -O(\varepsilon_n^2) \;\leq\; & -\frac{1}{2}\left[\|\widehat{\alpha}_n - \alpha_0\|^2 - \|P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n) - \alpha_0\|^2\right] \\ & +n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - \alpha^*(\widehat{\alpha}_n,\varepsilon_n), X, Y]) \\ & +O_p(\varepsilon_n^2) \end{aligned} \tag{S 9.3}$$

We will use the relation

$$\begin{aligned} \widehat{\alpha}_n - \alpha^*(\widehat{\alpha}_n,\varepsilon_n) \;&=\; \widehat{\alpha}_n - \widehat{\alpha}_n + \varepsilon_n\widehat{\alpha}_n - \varepsilon_n u^* - \varepsilon_n\alpha_0 \\ &=\; -\varepsilon_n\left(u^* - (\widehat{\alpha}_n - \alpha_0)\right) \end{aligned}$$

in $\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - \alpha^*(\widehat{\alpha}_n,\varepsilon_n), X, Y])$ to get $-\nu_n(l'_{\alpha_0}[\varepsilon_n\left(u^* - (\widehat{\alpha}_n - \alpha_0)\right), X, Y])$.

In (S 9.3) we have

$$\begin{aligned}
\left\|P_n a^*(\widehat{a}_n,\varepsilon_n)-a_0\right\|^2 &= \left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)+\alpha^*(\widehat{a}_n,\varepsilon_n)-\theta_0\right\|^2\\
&= \left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)+(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)+\varepsilon_n u^*\right\|^2\\
&\leq \left\|(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)\right\|\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)+\varepsilon_n u^*\right\|\\
&\leq \left\|(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)\right\|\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\right\|\\
&\quad +\left\|(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)\right\|\left\|\varepsilon_n u^*\right\|\\
&= (1-\varepsilon_n)\left\|(\widehat{\alpha}_n-\alpha_0)\right\|\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\right\|\\
&\quad +(1-\varepsilon_n)\left\langle\widehat{\alpha}_n-\alpha_0,\varepsilon_n u^*\right\rangle
\end{aligned}$$

We multiply $\|\widehat{a}_n-\alpha_0\|$ by the factor

$$\begin{aligned}
1-(1-\varepsilon_n)^2 &= 1-(1-2\varepsilon_n+\varepsilon_n^2)\\
&= 2\varepsilon_n-\varepsilon_n^2
\end{aligned}$$

which is a positive fraction that preserves the inequality. We also multiply $\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\theta_0\|^2$ by 2 which also preserves the inequality. Hence we obtain

$$\begin{aligned}
-O(\varepsilon_n^2) \leq\ & -\frac{1}{2}\left[1-(1-\varepsilon_n)^2\right]\left\|\widehat{\alpha}_n-\alpha_0\right\|^2\\
&+(1-\varepsilon_n)\left\|(\widehat{\alpha}_n-\alpha_0)\right\|\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\right\|\\
&+(1-\varepsilon_n)\left\langle\widehat{\alpha}_n-\alpha_0,\varepsilon_n u^*\right\rangle\\
&-n^{-1/2}\nu_n(l'_{\alpha_0}[\varepsilon_n\left(u^*-(\widehat{\alpha}_n-\alpha_0)\right),X,Y])\\
&+O_p(\varepsilon_n^2)
\end{aligned}$$

Adding $\varepsilon_n\left\|(\widehat{\alpha}_n-\alpha_0)\right\|\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\right\|$ still preserves the inequality. For the first line, $\varepsilon_n^2\left\|\widehat{\alpha}_n-\alpha_0\right\|^2 = O_p(\varepsilon_n^2)$. Hence

$$\begin{aligned}
-O(\varepsilon_n^2) \leq\ & -\varepsilon_n\left\|\widehat{\alpha}_n-\alpha_0\right\|^2+\left\|(\widehat{\alpha}_n-\alpha_0)\right\|\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\right\|\\
&+(1-\varepsilon_n)\left\langle\widehat{\alpha}_n-\alpha_0,\varepsilon_n u^*\right\rangle-n^{-1/2}\nu_n(l'_{\alpha_0}[\varepsilon_n\left(u^*-(\widehat{\alpha}_n-\alpha_0)\right),X,Y])+O_p(\varepsilon_n^2)
\end{aligned}$$

Note that

$$\begin{aligned}
-\varepsilon_n\left\|\widehat{\alpha}_n-\alpha_0\right\|^2 &= O_p(\varepsilon_n)o_p(\delta^2)\\
&= o_p(\delta^2)
\end{aligned}$$

By Condition C

$$\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\right\| = O_p(\delta^{-1}\varepsilon_n^2)$$

since

$$\left\|\widehat{\alpha}_n-\alpha_0\right\| = o_p(\delta)$$

then

$$\begin{aligned}
\left\|\widehat{\alpha}_n-\alpha_0\right\|\left\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\right\| &= o_p(\delta)O_p(\delta^{-1}\varepsilon_n^2)\\
&= o_p(\varepsilon_n^2)
\end{aligned}$$

and using Conditions C and D

$$n^{-1/2}\nu_n(l'_{\alpha_0}[\varepsilon_n\left(u^*-(\widehat{\alpha}_n-\alpha_0)\right),X,Y]) = n^{-1/2}\nu_n(l'_{\alpha_0}[u^*,X,Y])+O_p(\varepsilon_n^2)+O_p(\varepsilon_n^2)$$

Hence

$$-(1-\varepsilon_n)\left\langle\widehat{\alpha}_n-\alpha_0,u^*\right\rangle+n^{-1/2}\nu_n(l'_{\alpha_0}[u^*,X,Y]) = o_p(n^{-1/2}) \tag{S 9.4}$$

This gives, together with the inequality in (S 9.4) with u^* replaced by $-u^*$,

$$\left|\left\langle\widehat{\alpha}_n-\alpha_0,u^*\right\rangle-n^{-1/2}\nu_n(l'_{\alpha_0}[u^*,X,Y])\right| = o_p(n^{-1/2})$$

so

$$\langle \widehat{\alpha}_n - \alpha_0, v^* \rangle = n^{-1/2} \nu_n(l'_{\alpha_0}[v^*, X, Y]) + o_p(n^{-1/2})$$

Hence, by (S 4.2)

$$\begin{aligned} f_{\alpha_n} - f_{\alpha_0} &= f'_{\alpha_0}[\alpha_n - \alpha_0] + o_p(u_n \|\alpha_n - \alpha_0\|_F^{\omega}) \\ &= \langle \widehat{\alpha}_n - \alpha_0, v^* \rangle + o_p(n^{-1/2}) \\ &= n^{-1/2} \nu_n(l'_{\alpha_0}[u^*, X, Y]) + o_p(n^{-1/2}) \\ &= n^{-1} \sum_{i=1}^{n} n^{1/2} \left\{ \sum_{j=1}^{n} w_{ij} l'_{\alpha_0}[u^*, X_i, Y_j] - E\left[l'_{\alpha_0}[u^*, X, Y] \middle| X = x_i \right] \right\} \end{aligned}$$

The result then follows from the Central Limit Theorem (CLT) for triangular arrays (Proposition) in Andrews (1994, p. 2251). Note that the conditions of the Proposition are satisfied under our assumptions. In particular, $\Theta \subseteq \mathbb{R}^{d_\theta}$ is compact, finite-dimensional convergence of $n^{1/2} \sum_{j=1}^{n} w_{ij} l'_{\alpha_0}[u^*, X_i, Y_j] - E\left[l'_{\alpha_0}[u^*, X, Y] \middle| X = x_i \right]$ holds for each x_i due to the classical Lindeberg-Levy CLT, and Condition A satisfies the stochastic equicontinuity requirement of the Proposition.

■

References

Ai, C. (1997). A semiparametric maximum likelihood estimator. *Econometrica 65*, 933–963.

Ai, C., A. Chatrath, and F. Song (2006). On the comovement of commodity prices. *American Journal of Agricultural Economics 88*(3), 574–588.

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Altonji, J. G. and L. M. Segal (1996, July). Small-sample bias in gmm estimation of covariance structures. *Journal of Business & Economic Statistics 14*(3), 353–66.

Andrews, D. W. K. (1994, May). Empirical process methods in econometrics. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4 of *Handbook of Econometrics*, Chapter 37, pp. 2248–2294. Elsevier.

Antoine, B., H. Bonnal, and E. Renault (2006a). On the efficient use of the informational content of estimating equations: Implied probabilities and maximum euclidean likelihood. forthcoming in the *Journal of Econometrics*.

Antoine, B., H. Bonnal, and E. Renault (2006b). On the efficient use of the informational content of estimating equations: Implied probabilities and maximum euclidean likelihood. forthcoming in the *Journal of Econometrics*.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.

Borwein, J. M. and A. S. Lewis (2006). *Convex Analysis and Nonlinear Optimization: Theory and Examples* (Second ed.). New York, NY: Springer.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*, 305–334.

Chen, X. (2005). Large sample sieve estimation of semi-nonparametric models. Technical report, Department of Economics, New York University.

Chen, X. and S. Ludvigson (2006). Land of addicts? an empirical investigation of habit-based asset pricing models. Technical report, Department of Economics, New York University.

Chen, X. and X. Shen (1998, March). Sieve extremum estimates for weakly dependent data. *Econometrica 66*(2), 289–314.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika 85*(4), 967–972.

Csiszar, I. (1967). On topological properties of f-divergences. *Studia Scientriarum Mathematicarum Hungaria 2*, 329–339.

Domínguez, M. A. and I. N. Lobato (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica 72*(5), 1601–1615.

Duncan, G. M. (1986, June). A semi-parametric censored regression estimator. *Journal of Econometrics 32*(1), 5–34.

Gallant, A. R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica 55*(2), 363–390.

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

Hall, P. and J. L. Horowitz (1996, July). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica 64*(4), 891–916.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–54.

Hansen, L. P., J. Heaton, and A. Yaron (1996, July). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics 14*(3), 262–80.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Econometrica 58*, 71–120.

Imbens, G. H., R. Spady, and P. Johnson (1998). Information theoretic approaches to inference in moment condition models. *Econometrica 66*, 333–357.

Imbens, G. W. (1997, July). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies 64*(3), 359–83.

Jaynes, E. T. (1957). Information theory and statistical mechanics i. *Physical Review 106*(4), 620–630.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice. Cowles Foundation Discussion Paper No. 1569.

Kitamura, Y. and M. Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica 65*(4), 861–874.

Kitamura, Y., G. Tripathi, and H. Ahn (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica 72*(6), 1667–1714.

Kullback, S. (1997). *Information Theory and Statistic* (Dover ed.). Mineola, NY: Dover Publications, Inc.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*(1), 79–86.

LeBlanc, M. and J. Crowley (1995). Semiparametric regression functionals. *Journal of the American Statistical Association 90*, 95–105.

Luenberger, D. G. (1969). *Optimization by vector space methods*. New York, NY: Wiley.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica 59*(4), 1161–1167.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In G. Maddala, C. Rao, and H. Vinod (Eds.), *Handbook of Statistics*, Volume 11, pp. 2111–2245. Amsterdam: Elsevier.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–82.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K. and R. J. Smith (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

Nishiyama, Y., Q. Liu, and N. Sueishi (2005, December). Semiparametric estimators for conditional moment restrictions containing nonparametric functions: Comparison of gmm and empirical likelihood procedures. In A. Zerger and R. Argent (Eds.), *MODSIM 2005 International Congress on Modelling and Simulation*, pp. 170–176. Modelling and Simulation Society of Australia and New Zealand. ISBN: 0-9758400-2-9.

Otsu, T. (2003a). Empirical likelihood for quantile regression. Manuscript.

Otsu, T. (2003b). Penalized empirical likelihood estimation of conditional moment restriction models with unknown functions. Department of Economics, University of Wisconsin-Madison.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–49.

Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall/CRC.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press.

Powell, J. L. (1994, May). Estimation of semiparametric models. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4 of *Handbook of Econometrics*, Chapter 41, pp. 2443–2521. Elsevier.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–30.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*(1), 300–325.

Ramalho, J. (2005). Small sample bias of alternative estimation methods for moment condition models: Monte carlo evidence for covariance structures. *Studies in Nonlinear Dynamics & Econometrics 9*(1).

Ramsey, J. B. (1999). The contribution of wavelets to the analysis of economic and financial data. *Phil. Trans. R. Soc. Lond. A 357*, 2593–2606.

Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica 55*, 875–891.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–54.

Schennach, S. M. (2006). Point estimation with exponentially tilted empirical likelihood. Technical report, University of Chicago. forthcoming in the Annals of Statistics.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*. Reprinted in the ACM SIGMOBILE Mobile Computing and Communications Review, Vol.5(1) (January 2001).

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics 25*(6), 2555–2591.

Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics 22*(2), 580–615.

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

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Annals of Statistics 12*(3), 898–916.

Smith, R. J. (1997, March). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Economic Journal 107*(441), 503–19.

Smith, R. J. (2003). Local gel estimation with conditional moment restrictions. Technical report, University of Warwick.

Smith, R. J. (2005). Local gel methods for conditional moment restrictions. Cemmap working paper cwp15/05, University of Warwick.

Smith, R. J. (2006). Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*. forthcoming.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*(398), 559–567.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York, Inc.

Wong, W. H. and T. A. Severini (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *The Annals of Statistics 19*(2), 603–632.

Zhang, J. and I. Gijbels (2003). Sieve empirical likelihood and extensions of the generalized least squares. *Scandinavian Journal of Statistics 30*, 1–24.