Semiparametric Estimation of Invertible Models

Andres Santos

Department of Economics

University of California, San Diego

e-mail: a2santos@ucsd.edu

July, 2011

Abstract

This paper proposes a simple estimator for semiparametric nonseparable models of the form $Y_i = m(X_i, h_0(X_i, U_i), \theta_0)$, where U_i is independent of X_i, θ_0 is a finite dimensional parameter and h_0 an unknown function. Under the assumption that the model is strictly monotonic in the scalar error term U_i, we propose an M-estimator $(\hat{\theta}, \hat{h})$ such that: (i) $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normally distributed, and (ii) $\|\hat{h} - h\|_{L^2} = o_p(n^{-\frac{1}{4}})$. Additionally, a weighted bootstrap procedure is shown to be consistent for the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$, hence allowing for straightforward inferential procedures. The estimator is simple to implement, fast to compute and its regularity assumptions are easily verified for a wide class of models. A Monte Carlo study examines small sample performance.

KEYWORDS: Semiparametric estimation, nonseparable models, weighted bootstrap.

1 Introduction

Estimation problems in which unobservables enter in an additive separable manner have long been recognized to impose strong requirements on the underlying economic model. An attractive alternative is thus provided by specifications that only demand invertibility in the latent variables – a property that is sometimes implied by primitives of the model (Heckman et al., 2010).

An extensive literature has examined identification in invertible models under the assumption that disturbances are independent of the exogenous variables (Brown, 1983; Roehrig, 1988; Benkard and Berry, 2006). The identification arguments for these models are constructive, and often lead to natural "plug-in" nonparametric estimators (Matzkin, 2003, 2008). However, the applicability of such estimators is limited to large datasets or low dimensional models, with empirical work often relying on parametric approaches instead (Berry et al., 1995). On the other hand, parametric approaches are not only less robust, but are also often based on a finite number of moment restrictions, which may unfortunately not suffice for guaranteeing that the model is identified.[1]

In this paper, we propose a semiparametric estimator that improves on the robustness of parametric approaches yet may still be applicable in small samples. Specifically, for an exogenous variable X_i with support $\mathcal{X} \subset \mathbf{R}^{d_x}$, and a known function $m : \mathcal{X} \times \mathbf{R} \times \mathbf{R}^{d_\theta} \to \mathbf{R}$, we consider:

$$Y_i = m(X_i, h_0(X_i, U_i), \theta_0) \ , \tag{1}$$

where U_i is a scalar unobservable random variable, and θ_0 and h_0 are unspecified finite and infinite dimensional parameters respectively. In order to guarantee nonparametric identification, (X_i, U_i) are assumed to be independent, the model to be strictly increasing in U_i, and the distribution of U_i is normalized to be uniform on $[0, 1]$.[2] Two important restrictions of this specification are the lack of dependence of h in θ, and the latent variable U_i being scalar. The former requirement enables us to provide easily verifiably primitive regularity conditions, while the latter allows us to obtain a computationally simple estimator that does not require inverting (1) in U_i. Despite these restrictions, specification (1) is fairly flexible, encompassing certain random coefficient (Chamberlain, 1992) and semiparametric transformation models (Linton et al., 2008) as special cases. Fully parametric nonseparable models may also be expressed as in (1) by letting h_0 be the quantile function of the disturbance. In this manner, our procedure enables for the joint estimation of the finite dimensional parameter θ_0, and the distribution of the unobservable in a single step.

As previously noted in the literature, scalar invertible models are fully characterized by the restrictions they impose on the conditional quantiles. In particular, $m(X_i, h_0(X_i, \tau), \theta_0)$ is the τ^{th} quantile of Y_i conditional on X_i, which implies that almost surely in X_i and for all $0 \leq \tau \leq 1$:

$$P(Y_i \leq m(X_i, h_0(X_i, \tau), \theta_0) | X_i) = \tau \ . \tag{2}$$

[1] A notable exception is Brown and Wegkamp (2002) who develop a parametric estimator that exploits the assumed full independence between the exogenous and latent variables employed in the identification argument.

[2] If V_i has strictly increasing cdf F_V, then almost surely $\tilde{h}(X_i, V_i) = h(X_i, F^{-1}(U_i))$ with $U_i = F_V(V_i)$, and therefore $U_i \sim U[0, 1]$. This transformation may of course not be appropriate if the object of interest is $\tilde{h}$ itself.

Our estimator is based on the insight that while (2) provides a continuum of conditional moment restrictions, it may alternatively be rewritten as a single conditional moment restriction with the help of an auxiliary random variable. Specifically, if T_i is independent of (Y_i, X_i) and has full support on $[0, 1]$, then we may instead express the continuum of restrictions in (2) as:

$$E[1\{Y_i \leq m(X_i, h_0(X_i, T_i), \theta_0)\} - T_i | X_i, T_i] = 0 \ . \tag{3}$$

It is worth emphasizing that unlike parametric nonseparable models estimated from unconditional moment restrictions, the single conditional moment restriction in (3) incorporates all the identifying information of (1). In this manner, we ensure nonparametric identification continues to hold and that estimation and identification are both based on the independence of X_i and U_i.

Since relationship (3) must hold for *any* random variable T_i independent of (Y_i, X_i) with full support on $[0, 1]$, we generate a suitable i.i.d. sample $\{Y_i, X_i, T_i\}_{i=1}^n$ by combining the data $\{Y_i, X_i\}_{i=1}^n$ with a randomly generated sample $\{T_i\}_{i=1}^n$. Exploiting the conditional moment restriction in (3) and the generated data $\{Y_i, X_i, T_i\}_{i=1}^n$ we are then able to estimate (θ_0, h_0) thorough traditional minimum distance methods (Ai and Chen, 2003). Specifically, employing results in Chen and Pouzo (2009), we establish that a minimum distance estimator $(\hat{\theta}, \hat{h})$ satisfies: (i) $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normally distributed, (ii) $\|\hat{h} - h\|_{L^2} = o_p(n^{-\frac{1}{4}})$, and (iii) the weighted bootstrap is consistent for the limiting law of $\sqrt{n}(\hat{\theta} - \theta_0)$. Our regularity conditions are more primitive than those in Chen and Pouzo (2009), as we are able to exploit the considerable additional structure of (3) that is not present in the more general model they study. Additionally, we also show that a large class of submodels of (2) are not regular in the sense of Bickel et al. (1993), implying a semiparametric efficiency bound for estimating θ_0 cannot be derived for this problem.

In closely related work, Komunjer and Santos (2010) develop an alternative semiparametric estimator for invertible nonseparable models with scalar latent variables. Their procedure, however, imposes high level assumptions that are hard to verify in many applications. Additional work on estimation also includes Manski (1983) and Altonji and Matzkin (2005), who propose nonparametric estimators with and without endogeneity respectively. Chesher (2003) examines local conditions for identification, while Imbens and Newey (2009) and Torgovitsky (2011) study triangular systems with endogeneity. Chernozhukov and Hansen (2005) and Chernozhukov et al. (2007) explore identification and estimation of quantile effects with endogeneity and without a triangular structure. A related literature has also examined identification and estimation of average derivatives for nonseparable models that are not invertible in the latent variables; see for example Blundell and Powell (2003), Hoderlein and Mammen (2007), Hoderlein and White (2009) and references therein.

The remainder of the paper is organized as follows. Section 2 introduces the estimator, while section 3 establishes consistency and derives the rate of convergence for the nonparametric component. Section 4 includes the asymptotic normality result and the consistency of a bootstrap procedure for conducting inference. A Monte Carlo study is included in section 5, and section 6 briefly concludes. All proofs are relegated to the Appendix.

2 The Estimator

In this section we formally introduce the general model and our proposed estimator. We begin with Assumption 2.1, which summarizes the requirements on the specifications we consider.

Assumption 2.1. *(i)* X_i *has support* $\mathcal{X} \subset \mathbf{R}^{d_x}$*, and* (Y_i, X_i) *satisfy* (1) *for some known continuous function* $m : \mathcal{X} \times \mathbf{R} \times \mathbf{R}^{d_\theta} \to \mathbf{R}$*, with* $X_i \perp U_i$ *and* $U_i \sim U[0,1]$*; (ii)* $\theta_0 \in \Theta \subset \mathbf{R}^{d_\theta}$ *and* $h_0 \in \mathcal{H} \subset \mathbf{H}$*; (iii)* $\frac{dm(x,e,\theta)}{de} > 0$ *for all* $(x, e, \theta) \in \mathcal{X} \times \mathbf{R} \times \Theta$*; (iv)* $\frac{dh_0(x,u)}{du} > 0$ *for all* $(x, u) \in \mathcal{X} \times (0, 1)$*.*

In Assumption 2.1(ii), $\mathcal{H}$ denotes a nonparametric set of smooth functions and $\mathbf{H}$ a generic Banach space. The parameter space $\Theta \times \mathcal{H}$ will be additionally restricted when we introduce sufficient conditions for consistency. Assumptions 2.1(iii)-(iv) together imply the true model is strictly monotonic in the unobservable U_i. It is worth noting that strict monotonicity is imposed only on the true parameter $h_0 \in \mathcal{H}$, but not on all parameters $h \in \mathcal{H}$. This allows us the flexibility to consider estimators that are not restricted to be monotonic, which are often simpler to compute. On the other hand, the theoretical results also hold if $\mathcal{H}$ is required to contain only monotonic functions – i.e. if we demand our estimator be itself monotonic, which sometimes leads to better finite sample performance when the true model is monotonic as well (Chernozhukov et al., 2010).

Under Assumption 2.1, the function $m(\cdot, h_0(\cdot, \tau), \theta_0)$ is the τ^{th} quantile of Y_i conditional on X_i. This is easily established by observing that if $U_i \perp X_i$ is uniformly distributed on $[0, 1]$, then:

$$\begin{aligned} U_i \perp X_i &\Leftrightarrow P(U_i \leq \tau | X_i) = \tau && \forall\ \tau \in [0,1] \\ &\Leftrightarrow P(Y_i \leq m(X_i, h_0(X,\tau), \theta_0) | X_i) = \tau && \forall\ \tau \in [0,1] \end{aligned} \tag{4}$$

where the second result follows from the strict monotonicity in U_i implied by Assumptions 2.1(iii)-2.1(iv). Thus, identification of (θ_0, h_0) is immediate as long as the parametrization is not redundant – i.e. there exists a unique $(\theta, h) \in \Theta \times \mathcal{H}$ for which $(x, \tau) \mapsto m(x, h(x, \tau), \theta)$ equals the conditional quantile function of Y_i given X_i. We impose this requirement in the following assumption:

Assumption 2.2. $P(m(X_i, h_0(X_i, U_i), \theta_0) = m(X_i, \tilde{h}(X_i, U_i), \tilde{\theta})) < 1\ \forall (\theta_0, h_0) \neq (\tilde{\theta}, \tilde{h}) \in \Theta \times \mathcal{H}$*.*

Demanding that Assumption 2.2 hold for all $(\tilde{\theta}, \tilde{h}) \neq (\theta_0, h_0)$ rather than for all $(\tilde{\theta}, \tilde{h}) \neq (\theta, h)$ and all $(\theta, h) \in \Theta \times \mathcal{H}$ is not necessarily vacuous. This results from having assumed that the function $u \mapsto h_0(x, u)$ is increasing at all $x \in \mathcal{X}$, but not having imposed such requirement on all $h \in \mathcal{H}$. Consequently, being able to distinguish between the specifications implied by different parameters is only necessary insofar as they are both monotonic in the unobservable U_i.

2.1 Transformation Models

To fix ideas, we illustrate our assumptions in the special case of a nonseparable semiparametric transformation model that may not be inverted into an additive separable specification.

Let $X_i = (X_{i1}, X_{i2})$ with $X_{i1} \in \mathcal{X}_1 \subset \mathbf{R}^{d_{x_1}}$, $X_{i2} \in \mathcal{X}_2 \subset \mathbf{R}^{d_{x_2}}$ and define $\Lambda : \mathbf{R} \times \mathbf{R}^{d_\beta} \to \mathbf{R}$ to be a parametric link function. For an unknown vector $(\beta_0, \gamma_0) \in \mathbf{R}^{d_\beta + d_{x_1}}$ and function $h_0 : \mathcal{X}_2 \times [0,1] \to \mathbf{R}$, we then assume the dependent variable Y_i is generated according to the transformation model:

$$Y_i = \Lambda(X_{1i}'\gamma_0 + h_0(X_{2i}, U_i), \beta_0) \ , \tag{M}$$

where U_i is unobservable, independent of X_i and uniformly distributed on $[0,1]$. In this context, the finite dimensional parameter of interest is then given by $\theta_0 = (\beta_0, \gamma_0)$, while the unknown function h_0 is the infinite dimensional nuisance parameter. Moreover, it is easily verified that Assumptions 2.1(iii)-(iv) are satisfied provided $\lambda(a, \beta) \equiv \frac{d\Lambda(a,\beta)}{da}$ is strictly positive for all $(a, \beta) \in \mathbf{R}^{1+d_\beta}$, and the function $u \mapsto h_0(x_2, u)$ is strictly increasing for all $x_2 \in \mathcal{X}_2$.

As Examples 2.1-2.3 show, several particular specifications of (M) are of special interest.

Example 2.1. The canonical parametric transformation model employs the link function of Box and Cox (1964) ($\Lambda^{-1}(y, \beta) = \frac{y^\beta - 1}{\beta}$ $(\beta \neq 0)$ and $\Lambda^{-1}(y, \beta) = \log(y)$ $(\beta = 0)$), and specifies:

$$Y_i = \Lambda(X_{i1}'\gamma_0 + \xi_i, \beta_0) \ , \tag{5}$$

where ξ_i is assumed independent of X_i but not necessarily uniform on $[0,1]$. If ξ_i is continuously distributed with strictly increasing cdf F_ξ, then we may express (5) as in (M) by letting $X_i = X_{i1}$, $h_0(u) = F_\xi^{-1}(u)$ and $U_i = F_\xi(\xi_i)$. As a result, estimation of (M) amounts to the simultaneous estimation of the parametric component (γ_0, β_0) and the quantile function of ξ_i (h_0). ■

Example 2.2. Letting $\Lambda(a, \beta) = a$ for all $(a, \beta) \in \mathbf{R}^{1+d_\beta}$ leads to a partially linear specification:

$$Y_i = X_{i1}'\gamma_0 + h_0(X_{i2}, U_i) \ , \tag{6}$$

which generalizes Robinson (1988) to allow for a more flexible specification that is nonseparable in the error term and the secondary regressor X_{i2}. Unlike Robinson (1988) who only requires mean independence, however, U_i is assumed fully independent of X_i in order to ensure identification. ■

Example 2.3. Suppose $X_{i2} \in \mathbf{R}_+$ almost surely, and consider the random coefficient model:

$$Y_i = \Lambda(X_{i1}'\gamma_0 + X_{i2}\xi_i, \beta_0) \ , \tag{7}$$

where ξ_i is assumed independent of $X_i = (X_{i1}, X_{i2})$. Proceeding as in Example 2.1, we may then rewrite (7) in the form (M) by letting $h_0(x_2, u) = x_2 F_\xi^{-1}(u)$ for F_ξ the cdf of ξ_i. In estimating (M), we therefore simultaneously recover the parametric component $\theta_0 = (\gamma_0, \beta_0)$ and the quantile function for the random coefficients. General versions of (7) were considered in, for example, hedonic models by Bajari and Benkard (2005) and Heckman et al. (2010). ■

We conclude by providing sufficient conditions for the identification requirement of Assumption 2.2 to hold in model (M). As Proposition 2.1 shows, if X_i is continuously distributed, then model (M) is identified provided that for any $\beta \neq \beta'$ the inverse link function $\Lambda^{-1}(\cdot, \beta)$ is not an affine transformation of $\Lambda^{-1}(\cdot, \beta')$. This condition is easily verified for the Box and Cox (1964) model or may be ensured through a scale and location normalization on $\Lambda^{-1}(\cdot, \beta_0)$.

Proposition 2.1. *Suppose (i) $X_i \equiv (X_{1i}, X_{2i})$ is continuously distributed with support $\mathcal{X}$, (ii) (Y_i, X_i) satisfy* (M) *with $U_i \perp X_i$ and $U_i \sim U[0,1]$, (iii) h is differentiable in (x_2, u) for all $h \in \mathcal{H}$, (iv) $\lambda(a,\beta) \equiv \frac{d}{da}\Lambda(a,\beta) > 0$, for all (a,β), (v) $P(\Lambda^{-1}(Y_i,\beta) = c_0\Lambda^{-1}(Y_i,\beta') + c_1) < 1$ for all $\beta \neq \beta'$ and $(c_0, c_1) \in \mathbf{R} \times \mathbf{R}$. It then follows that Assumption 2.2 holds for model* (M).

2.2 Criterion Function

Assumption 2.2 ensures the parameter (θ_0, h_0) is identified as the unique element of $\Theta \times \mathcal{H}$ for which $(x,\tau) \mapsto m(x, h(x,\tau), \theta)$ equals the conditional quantile function of Y_i given X_i. In proceeding with estimation, it is therefore natural to construct an extremum estimator based on this observation.

For notational convenience, we combine the finite (θ) and infinite (h) dimensional components into a single parameter $\alpha \equiv (\theta, h)$, let $\alpha_0 = (\theta_0, h_0)$ and denote the parameter space by $\mathcal{A} \equiv \Theta \times \mathcal{H}$. Defining $m(x,u,\alpha) = m(x, h(x,u), \theta)$, we then obtain that the conditions identifying α_0 are:

$$E[1\{Y_i \leq m(X_i, \tau, \alpha_0)\} - \tau | X_i] = 0 \qquad \forall\ \tau \in [0,1]\ , \tag{8}$$

which constitute a continuum of conditional moment restrictions indexed by τ. Alternatively, we may reinterpret (8) as a single conditional moment restriction with the introduction of an auxiliary random variable T_i. Specifically, if T_i is independent of (Y_i, X_i) we may then rewrite (8) as:

$$E[1\{Y_i \leq m(X_i, T_i, \alpha_0)\} - T_i | X_i, T_i] = 0\ , \tag{9}$$

provided T_i is continuously distributed with support $[0,1]$. Thus transformed, we see the original nonseparable model is identified through a single conditional moment restriction.

Exploiting this equivalence, we can rely on procedures for the semiparametric estimation of conditional moment models to obtain a suitable estimator for α_0. Towards this end, let us denote:

$$W(X_i, T_i, \alpha) \equiv E[1\{Y_i \leq m(X_i, T_i, \alpha)\} - T_i | X_i, T_i]\ , \tag{10}$$

and following Ai and Chen (2003, 2007) and Chen and Pouzo (2009), define the criterion function:

$$Q(\alpha) \equiv E[W^2(X_i, T_i, \alpha)]\ . \tag{11}$$

By construction, α_0 is the unique minimum of Q on $\mathcal{A}$, and we hence estimate it by the minimizer of a suitable sample analogue Q_n over an approximating sieve $\mathcal{A}_n$.

In order to construct Q_n, we require an i.i.d. sample $\{Y_i, X_i, T_i\}_{i=1}^n$ of (Y_i, X_i, T_i). However, since the main assumption on T_i is that it be independent of (Y_i, X_i), we can simply create $\{T_i\}_{i=1}^n$ through a random number generator and then combine it with the original data $\{Y_i, X_i\}_{i=1}^n$ to obtain the desired sample. In turn, employing the sample $\{Y_i, X_i, T_i\}_{i=1}^n$ we build a nonparametric estimator for the conditional expectation $W(X_i, T_i, \alpha)$ using a linear series estimator. In particular, let $\{p_k\}_{k=1}^\infty$ be a family of basis functions in (x,t), and denote the vector of first k_n terms evaluated at $(x,t) \in \mathcal{X} \times [0,1]$ by $p^{k_n}(x,t) = (p_1(x,t), \ldots, p_{k_n}(x,t))'$. For $P = (p^{k_n}(X_1, T_1), \ldots, p^{k_n}(X_n, T_n))'$,

the nonparametric series estimator for $W(X_i, T_i, \alpha)$ is then given by:

$$\hat{W}(x,t,\alpha) \equiv p^{k'_n}(x,t)(P'P)^{-1}\sum_{i=1}^{n} p^{k_n}(X_i,T_i)(1\{Y_i \leq m(X_i,T_i,\alpha)\} - T_i) \ , \tag{12}$$

which is just the predicted values obtained from an ordinary least square regression. Provided linear combinations of $\{p_k\}_{k=1}^{\infty}$ can approximate the true conditional expectation arbitrarily well, $\hat{W}(x,t,\alpha)$ is then a consistent estimator for $W(x,t,\alpha)$ as $k_n \uparrow \infty$ (Newey, 1997).

Given these definitions, a suitable finite sample analogue to the criterion function Q is then:

$$Q_n(\alpha) \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{W}^2(X_i,T_i,\alpha) \ , \tag{13}$$

and a corresponding estimator for α_0 is given by the minimizer of Q_n. Due to the presence of the infinite dimensional component $h \in \mathcal{H}$, however, minimizing Q_n over $\mathcal{A}$ may prove both impractical and undesirable. We therefore instead employ a finite dimensional linear sieve $\mathcal{H}_n$ required to grow dense in $\mathcal{H}$. Specifically, for $\{q_j\}_{j=1}^{\infty}$ a family of basis functions in (x,t), we denote the vector of the first j_n terms evaluated at $(x,t) \in \mathcal{X} \times [0,1]$ by $q^{j_n}(x,t) = (q_1(x,t), \ldots, q_{j_n}(x,t))'$, and let:

$$\mathcal{H}_n \equiv \{h \in \mathcal{H} : h(x,t) = \pi' q^{j_n}(x,t) \text{ for some } \pi \in \mathbf{R}^{j_n}\} \ . \tag{14}$$

Denoting the sieve for the parameter space $\mathcal{A}$ by $\mathcal{A}_n \equiv \Theta \times \mathcal{H}_n$, we then let our estimator be:

$$\hat{\alpha} \in \arg\min_{\alpha \in \mathcal{A}_n} Q_n(\alpha) \ . \tag{15}$$

The estimator $\hat{\alpha}$ is equivalent to the one proposed in Chen and Pouzo (2009) for semiparametric estimation of nonsmooth conditional moment models. Our asymptotic analysis therefore relies on their general results, though we are able to provide simpler sufficient conditions due to the additional structure present in our problem – mainly exogeneity and monotonicity.

Remark 2.1. Employing a series estimator for the conditional probability, we could have considered

$$\tilde{Q}_n(\alpha) \equiv \int_0^1 \Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{P}(Y_i \leq m(X_i,\tau,\alpha)|X_i) - \tau\Big)^2\Big\} w(\tau) d\tau \ , \tag{16}$$

for some weight function $w : [0,1] \to \mathbf{R}_+$, as a criterion function for estimating α_0. Setting w to equal the density of T_i, we see that $\tilde{Q}_n$ is also a sample analogue for Q, but where the uncertainty over the distribution of T_i is not present. The evaluation of $\tilde{Q}(\alpha)$, however, requires integrating a series of regressions over $\tau \in [0,1]$ – a computational cost that is exacerbated during minimization and bootstrap calculations. On the other hand, the simplicity of Q_n comes at a cost in the asymptotic variance as the fact that the distribution of T_i is known is not exploited in estimation. ■

3 Consistency and Rates

We first introduce the norm under which we obtain consistency for the nonparametric component. Let $\lambda \in \mathbf{R}^{d_x+1}$ be a vector of nonnegative integers with degree $|\lambda| = \sum_i \lambda^{(i)}$, where $\lambda^{(i)}$ is the i^{th} coordinate of the vector λ. Further let $D^\lambda h(x,u) = \partial^{|\lambda|} h(x,u)/\partial x_1^{\lambda_1} \dots \partial x_{d_x}^{\lambda_{d_x}} \partial u^{\lambda^{d_x+1}}$ and define:

$$\|h\|_{\infty,\kappa} \equiv \max_{|\lambda|\leq\kappa} \sup_{(x,u)\in\mathcal{X}\times[0,1]} |D^\lambda h(x,u)| \ , \tag{17}$$

where κ is an integer. Convergence of the estimator for h_0 under $\|\cdot\|_{\infty,\kappa}$ implies uniform convergence not only in level but also in level of derivatives up to order κ. In turn, the norm $\|\cdot\|_{\infty,\kappa}$ can be used to construct a norm for the whole parameter space $\mathcal{A} = \Theta \times \mathcal{H}$ in the natural way by defining:

$$\|\alpha\|_{c,\kappa} \equiv \|\theta\| + \|h\|_{\infty,\kappa} \ . \tag{18}$$

In order to establish the consistency under $\|\cdot\|_{c,\kappa}$ we impose the following additional assumptions:

Assumption 3.1. *(i) $\{Y_i, X_i, T_i\}$ is an i.i.d. sample; (ii) $T_i \perp (Y_i, X_i)$ has support $[0,1]$, with density bounded from above and away from zero; (iii) X_i has compact connected support $\mathcal{X}$ with Lipschitz continuous boundary and density f_X bounded from above and away from zero.*

Assumption 3.2. *(i) The eigenvalues of $E[p^{k_n}(X_i,T_i)p^{k_n}(X_i,T_i)]$ are bounded above and away from zero; (ii) There is $\pi_n : \mathcal{A}_n \to \mathbf{R}^{k_n}$ with $\sup_{\alpha\in\mathcal{A}_n} |W(X_i,T_i,\alpha) - \pi_n(\alpha)p^{k_n}(X_i,T_i)| = O(k_n^{-\frac{r_p}{d_x+1}})$ a.s.; (iii) $j_n \uparrow \infty$, $k_n \uparrow \infty$, $k_n = o(n)$ and $\xi_{p,n}^2 \times k_n = o(n)$ for $\xi_{p,n} \equiv \sup_{(x,t)\in\mathcal{X}\times[0,1]} \|p^{k_n}(x,t)\|$.*

Assumption 3.3. *(i) $\mathcal{A}$ is compact in $\|\cdot\|_{c,\kappa}$ with $\kappa > \frac{d_x+1}{2}$; (ii) $\mathcal{A}_n \subseteq \mathcal{A}_{n+1} \subseteq \mathcal{A}$ with $\mathcal{A}_n, \mathcal{A}$ convex and closed in $\|\cdot\|_{c,\kappa}$; (iii) For every $\alpha \in \mathcal{A}$ there is $\Pi_n\alpha \in \mathcal{A}_n$ with $\|\Pi_n\alpha - \alpha\|_{c,\kappa} = o(1)$; (iv) $(x,y) \mapsto f_{Y|X}(y|x)$ is bounded from above, away from zero, and for all $(x,u) \in \mathcal{X}\times[0,1]$ and $\alpha, \tilde{\alpha} \in \mathcal{A}$, $|m(x,u,\alpha) - m(x,u,\tilde{\alpha})| \leq G(x,u)\|\alpha - \tilde{\alpha}\|_{c,\kappa}$ with $E[G^2(X_i,T_i)] < \infty$.*

Assumption 3.1(ii) imposes the requirements on T_i, while Assumptions 3.1(iii) and 3.2(i)-(iii) are standard when employing series estimators for conditional expectations (Huang, 1998, 2003); see also Remark 3.1. Assumption 3.3(i) imposes compactness on $\mathcal{A}$ under $\|\cdot\|_{c,\kappa}$, which plays an important role in our proof of consistency; see Remark 3.2. Alternatively, utilizing a penalized extremum estimator as in Chen and Pouzo (2011) would enable us to dispense with this requirement, but for simplicity we do not pursue this approach here. Assumption 3.3(iii) requires $\mathcal{A}_n$ to grow appropriately dense in $\mathcal{A}$ – a requirement satisfied by standard sieves such as polynomials and B-splines (Newey, 1997). Finally, the Lipschitz condition imposed in Assumption 3.3(iv) is model specific, though often easily verified due to the strength of the norm $\|\cdot\|_{c,\kappa}$; see Proposition 3.1 for sufficient conditions in model (M).

Remark 3.1. The approximation requirement of Assumption 3.2(ii) can be verified by imposing the conditional probabilities be sufficiently smooth. For example, if we require:

$$\sup_{\alpha\in\mathcal{A}} \|W(\cdot,\cdot,\alpha)\|_{\infty,\kappa} < \infty \ , \tag{19}$$

and $\{p_k\}_{k=1}^{\infty}$ are polynomials or tensor product univariate splines, then Assumption 3.2(ii) holds with $r_p = \kappa$; see Chen (2007). In verifying Assumption 3.2(iii), it is useful to note the relationship between k_n and $\xi_{p,n}$, which may be sieve specific. For example, for $\{p_k\}_{k=1}^{\infty}$ polynomial or tensor product univariate splines we have $\xi_{p,n} \asymp k_n^{d_x+1}$ and $\xi_{p,n} \asymp k_n^{\frac{d_x+1}{2}}$ respectively; see Newey (1997). ■

Remark 3.2. The compactness of $\mathcal{A}$ in $\|\cdot\|_{c,\kappa}$ can be guaranteed by assuming Θ is compact in $\|\cdot\|$ and $\mathcal{H}$ in $\|\cdot\|_{\infty,\kappa}$. A simple way to impose the latter requirement is to define the norm:

$$\|h\|_{2,\kappa_0}^2 \equiv \sum_{|\lambda|\leq\kappa_0} \int_{\mathcal{X}\times[0,1]} [D^\lambda h(x,u)]^2 dxdu \tag{20}$$

and let $\mathcal{H}$ denote the closure of $\{h : \mathcal{X} \times [0,1] \to \mathbf{R} : \|h\|_{2,\kappa_0}^2 \leq B\}$ under the $\|\cdot\|_{\infty,\kappa}$ for some chosen bound $B > 0$. If $\mathcal{X}$ is sufficiently regular and $\kappa_0 > \kappa + (d_x+1)/2$, then this choice for $\mathcal{H}$ is compact in $\|\cdot\|_{\infty,\kappa}$.[3] This smoothness assumption is easily implemented in linear sieves $\mathcal{H}_n$ for which it is equivalent to a quadratic constraint on the coefficients (Newey and Powell, 2003). ■

In Lemma 3.1, we establish the consistency of the estimator $\hat{\alpha}$ under the given assumptions.

Lemma 3.1. *Under Assumptions 2.1, 2.2, 3.1, 3.2 and 3.3, it follows that* $\|\hat{\alpha} - \alpha_0\|_{c,\kappa} = o_p(1)$.

We illustrate Lemma 3.1 by establishing the consistency of the estimator for the semiparametric transformation model (M). Given the discussion of Remarks 3.1 and 3.2 we focus in providing primitives for Assumptions 3.2(ii) and 3.3(iv), which are the only ones that are model specific. The assumptions of Proposition 3.1 are stronger than necessary for ease of exposition.

Proposition 3.1. *Suppose (i)* $\lambda(a,\beta) \equiv \frac{d}{da}\Lambda(a,\beta)$ *is bounded in* (a,β)*, (ii)* $(a,\beta) \mapsto \Lambda(a,\beta)$ *is differentiable in* β *with derivatives bounded in* (a,β)*, (iii)* $(y,x) \mapsto f_{Y|X}(y|x)$ *is bounded above and away from zero, (iv) The domain of* Λ *is convex, and (v)* $\kappa \geq r_p$*,* $(a,x) \mapsto P(Y_i \leq a|X_i = x)$ *and* $a \mapsto \Lambda(a,\beta)$ *are* r_p*-times differentiable for all* β *with bounded derivatives. If Assumptions 3.1(i)-(iii), 3.2(i), 3.2(iii), 3.3(i)-(ii), the conditions of Proposition 2.1 hold, and* $\{p_k\}_{k=1}^{\infty}$ *and* $\{q_j\}_{j=1}^{\infty}$ *are tensor product univariate splines, or polynomials, then it follows that:*

$$\|\hat{\beta} - \beta_0\| \xrightarrow{p} 0 \qquad \|\hat{\theta} - \theta\| \xrightarrow{p} 0 \qquad \|\hat{h} - h_0\|_{\infty,\kappa} \xrightarrow{p} 0 \ .$$

Remark 3.3. An interesting implication of Lemma 3.1, is that if $\kappa \geq 2$, then the derivative of $\hat{h}$ must converge uniformly to the derivative of h_0. In particular, since $u \mapsto h_0(x,u)$ is strictly monotonic at all $x \in \mathcal{X}$, it follows that $\hat{h}$ is strictly monotonic as well with probability tending to one, even if monotonicity is not imposed on the parameter space $\mathcal{H}$. ■

3.1 Rate of Convergence

In this section, we proceed to bound the rate of convergence of $\hat{\alpha}$. The result is not only of independent interest, but also instrumental in establishing the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta)$.

[3] See for example Theorem 6.3 in Adams and Fournier (2003). A sufficient condition for $\mathcal{X}$ to be regular is that it satisfy a "cone condition", see page 82 in Adams and Fournier (2003). Heuristically this requires that there exist a fixed finite cone whose vertex can be placed at any point $x \in \mathcal{X}$ in such a way that the cone lies in $\mathcal{X}$.

Before stating the necessary assumptions, however, we need to introduce some additional notation.

The norm under which we obtain a rate of convergence is the mean square error norm given by:

$$\|\alpha\|_{c,L^2} \equiv \|\theta\| + \|h\|_{L^2} \qquad\qquad \|h\|_{L^2} \equiv \sqrt{E[h^2(X_i, T_i)]} \ . \tag{21}$$

We also let $\bar{\mathcal{H}}$ and $\bar{\mathcal{A}}$ denote the completion of the linear spans of $\mathcal{H}$ and $\mathcal{A}$ under $\|\cdot\|_{L^2}$ and $\|\cdot\|_{c,L^2}$ respectively. As is standard in the literature, the analysis of the rate of convergence exploits the local behavior of the criterion function Q at the point α_0. For this purpose, we assume $\alpha \mapsto m(x, u, \alpha)$ is differentiable, with pathwise derivative at $\bar{\alpha} \equiv (\bar{\theta}, \bar{h})$ in the direction $\alpha \equiv (\theta, h)$ given by:

$$\frac{dm(x, u, \bar{\alpha} + \tau\alpha)}{d\tau}\Big|_{\tau=0} = \Big(\frac{dm(x, \bar{h}(x, u), \theta)}{d\theta'}\Big|_{\theta=\bar{\theta}}\Big)\theta + \Big(\frac{dm(x, e, \bar{\theta})}{de}\Big|_{e=\bar{h}(x,u)}\Big)h(x, u) \ . \tag{22}$$

Alternatively, for ease of notation we also employ the following more compact expression for (22):

$$\frac{dm(x, u, \bar{\alpha})}{d\alpha}[\alpha] \equiv \frac{dm(x, u, \bar{\alpha})}{d\theta}[\theta] + \frac{dm(x, u, \bar{\alpha})}{dh}[h] \ . \tag{23}$$

We will further impose a local identification condition for α_0, which requires the parametric component (θ_0) to be "distinct" from the nonparametric component (h_0) (Chen et al., 2011). Formally, for each element $\theta^{(j)}$ of the vector $\theta \in \mathbf{R}^{d_\theta}$ and $1 \leq j \leq d_\theta$, let h_j^* be defined by:

$$h_j^* = \arg\min_{h \in \bar{\mathcal{H}}} E\Big[\Big(\frac{dm(X_i, T_i, \alpha_0)}{d\theta^{(j)}} - \frac{dm(X_i, T_i, \alpha_0)}{dh}[h]\Big)^2 f^2_{Y|X}(m(X_i, T_i, \alpha_0)|X_i)\Big] \ . \tag{24}$$

Heuristically, each h_j^* is the nonparametric component that makes $\theta_0^{(j)}$ the "hardest" to locally identify. We additionally define the vector of functions $h^* \equiv (h_1^*, \ldots, h_{d_\theta}^*)$ and similarly denote:

$$\frac{dm(x, u, \alpha_0)}{dh}[h^*] \equiv \Big(\frac{dm(x, u, \alpha_0)}{dh}[h_1^*], \ldots, \frac{dm(x, u, \alpha_0)}{dh}[h_{d_\theta}^*]\Big) \ . \tag{25}$$

Finally, we introduce the vector of implied projection residuals, which is pointwise given by:

$$R_{h^*}(x, u) \equiv \left(\frac{dm(x, u, \alpha_0)}{d\theta} - \frac{dm(x, u, \alpha_0)}{dh}[h^*]\right) f_{Y|X}(m(x, u, \alpha_0)|x) \ . \tag{26}$$

Given the above notation, we introduce the following additional assumptions.

Assumption 3.4. *(i) $(x, e, \theta) \mapsto m(x, e, \theta)$ is twice differentiable in (e, θ) with derivatives up to order two continuous in (x, e, θ); (ii) The matrix $E[R_{h^*}(X_i, T_i)R'_{h^*}(X_i, T_i)]$ is positive definite; (iii) For every $\alpha \in \mathcal{A}$ there exists $\Pi_n\alpha \in \mathcal{A}_n$ such that $\|\alpha - \Pi_n\alpha\|_{c,0} = O(j_n^{-\frac{r_q}{d_x+1}})$.*

Assumption 3.4(i) requires $(e, \theta) \mapsto m(x, e, \theta)$ to be differentiable in (e, θ) at all $x \in \mathcal{X}$, but not necessarily differentiable in x itself. The local identification requirement on α_0 is imposed in Assumption 3.4(ii); see also Remark 3.4 below for primitive conditions. Assumption 3.4(iii) controls the bias in estimating α_0 introduced by employing a sieve ($\mathcal{A}_n$) rather than the entire space ($\mathcal{A}$). The discussion in Remark 3.1 is applicable for providing primitives for this rate requirement.

Remark 3.4. Assumption 3.4(ii) holds if and only if $\lambda' E[R_{h^*}(X_i, T_i)R'_{h^*}(X_i, T_i)]\lambda > 0$ for all

$0 \neq \lambda \in \mathbf{R}^{d_\theta}$. Equivalently, since $f_{Y|X}$ is bounded away from zero by Assumption 3.3(iv), the local identification condition is satisfied if for all $0 \neq \lambda \in \mathbf{R}^{d_\theta}$ we have that:

$$\min_{h \in \bar{\mathcal{H}}} E\Big[\Big(\frac{dm(X_i, T_i, \alpha_0)}{d\theta}[\lambda] - \frac{dm(X_i, T_i, \alpha_0)}{dh}[h]\Big)^2\Big] > 0 \ . \tag{27}$$

However, by Assumption 2.1(iii) $\frac{dm(x,e,\theta_0)}{de} > 0$ for all $e \in \mathbf{R}$, and hence condition (27) holds if:

$$P\Big(\Big(\frac{dm(X_i, T_i, \alpha_0)}{dh}\Big)^{-1} \times \frac{dm(X_i, T_i, \alpha_0)}{d\theta}[\lambda] = h(X_i, T_i)\Big) < 1 \ , \tag{28}$$

for all $0 \neq \lambda \in \mathbf{R}^{d_\theta}$ and $h \in \bar{\mathcal{H}}$. For instance, if $X_i = (X_{i1}, X_{i2})$, then Assumption 3.4(ii) may be readily verified if $\frac{dm(X_i,T_i,\alpha_0)}{d\theta}$ depends on X_{i1}, but X_{i1} does not enter h; see Proposition 3.2. ■

Remark 3.5. Since $(x, u) \mapsto h_0(x, u)$ is continuous, Assumption 3.4(i) and compactness of $\mathcal{X}$ imply

$$E\Big[\Big(\frac{dm(X_i, T_i, \alpha_0)}{dh}[h]\Big)^2\Big] < \infty \ , \tag{29}$$

for all $h \in \bar{\mathcal{H}}$. Hence, the set $\mathcal{F} \equiv \{f : \mathcal{X} \times [0,1] \to \mathbf{R}$ with $f(x,u) = \frac{dm(x,u,\alpha_0)}{dh}[h]$ for some $h \in \mathcal{H}\}$ forms a closed subspace of $L^2(X_i, T_i) \equiv \{f : \mathcal{X} \times [0,1] \to \mathbf{R}$ with $\|h\|_{L^2} < \infty\}$. It follows that the minimum in (24) is indeed attained and unique due to the Projection Theorem in Hilbert Spaces; see for example Theorem 3.3.2 in Luenberger (1969). ■

The conditions of Lemma 3.1 and Assumption 3.4 are sufficient for bounding the rate of convergence of the proposed estimator.

Theorem 3.1. *If Assumptions 2.1, 2.2, 3.1, 3.2, 3.3 and 3.4 hold, then it follows that*

$$\|\hat{\alpha} - \alpha_0\|_{c,L^2} = O_p\Big(\frac{\sqrt{k_n}}{\sqrt{n}} + k_n^{-\frac{r_p}{d_x+1}} + j_n^{-\frac{r_q}{d_x+1}}\Big) \ .$$

We conclude this section by providing primitive conditions for the semiparametric transformation model (M) to meet the assumptions of Theorem 3.1. Once again we focus on the assumptions that are model specific, particularly the local identification requirement of Assumption 3.4(ii).

Proposition 3.2. *Let $\tilde{Y}_i = X_{1i}'\gamma_0 + h_0(X_{2i}, T_i)$ and $Z_i = (\nabla_{\beta'}\Lambda(\tilde{Y}_i, \beta_0), \lambda(\tilde{Y}_i, \beta_0) \times (1, \tilde{Y}_i)')'$. If (i) $E[Z_i Z_i']$ has full rank, (ii) $\gamma_0 \neq 0$, (iii) $(a, \beta) \mapsto \Lambda(a, \beta)$ is twice continuously differentiable, and all assumptions of Proposition 3.1 hold, then $\|\hat{h} - h_0\|_{L^2} = O_p(\sqrt{k_n}/\sqrt{n} + k_n^{-\frac{r_p}{d_x+1}} + j_n^{-\frac{\kappa}{d_{x_2}+1}})$.*

4 Inference

This section develops the results necessary for performing inference on the parametric component (θ_0). For this purpose we first establish the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$ and then show the validity of a weighted bootstrap procedure for estimating the limiting distribution.

Towards this end, it is convenient to employ an alternative representation for $(\hat{\theta} - \theta_0)$. Let $0 \neq \lambda \in \mathbf{R}^{d_\theta}$ and define the functional $F_\lambda : \bar{\mathcal{A}} \to \mathbf{R}$ by $F_\lambda(\alpha) = \lambda'\theta$. The functional $F_\lambda(\alpha)$ is

linear and continuous under $\|\cdot\|_{c,L^2}$ and hence admits for a simple representation by the Riesz representation theorem. In fact, as originally noted in Ai and Chen (2003), the functions h_j^* and matrix $E[R_{h^*}(X_i,T_i)R'_{h^*}(X_i,T_i)]$ play a crucial role in this alternative representation. Defining:

$$v^\lambda \equiv (v_\theta^\lambda, v_h^\lambda) \qquad v_\theta^\lambda \equiv (E[R_{h^*}(X_i,T_i)R'_{h^*}(X_i,T_i)])^{-1}\lambda \qquad v_h^\lambda \equiv -h^{*\prime} v_\theta^\lambda \ , \tag{30}$$

we show in Corollary 4.1 that v^λ is the Riesz representor for the functional $\lambda \mapsto F_\lambda(\alpha)$.

Corollary 4.1. *Under Assumption 3.4(ii), it follows that for any $\alpha \in \bar{\mathcal{A}}$ we have:*

$$F_\lambda(\alpha-\alpha_0) = E\Big[f^2_{Y|X}(m(X_i,T_i,\alpha_0)|X_i)\Big(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha-\alpha_0]\Big)\Big(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[v^\lambda]\Big)\Big] \ .$$

We introduce the final set of Assumptions required to obtain asymptotic normality.

Assumption 4.1. *(i) $y \mapsto f_{Y|X}(y|x)$ is differentiable for all x, with derivative uniformly bounded in (y,x); (ii) For every $\lambda \in \mathbf{R}^{d_\theta}$ there is $\Pi_n v_h^\lambda \equiv \pi(\lambda)q^{j_n}$ such that $\|\Pi_n v_h^\lambda - v_h^\lambda\|_\infty = O(j_n^{-\frac{rq}{d_x+1}})$ and $\gamma(\lambda)$ such that $|f_{Y|X}(m(X_i,T_i,\alpha_0)|X_i)\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[v^\lambda] - \gamma(\lambda)p^{k_n}(X_i,T_i)| = O(k_n^{-\frac{rp}{d_x+1}})$ a.s.*

Assumption 4.2. *(i) $\theta_0 \in int(\Theta)$; (ii) $\|h_j^*\|_\infty < \infty$ for all $1 \le j \le d_\theta$; (iii) The eigenvalues of $E[q^{j_n}(X_i,T_i)q^{j'_n}(X_i,T_i)]$ are bounded above and away from zero; (iv) $k_n^5\xi_{q,n}^2 \vee k_n^2\xi_{q,n}^4 = o(n)$, $\xi_{q,n}(k_n^{-\frac{5rp}{d_x+1}} \vee j_n^{-\frac{5rq}{d_x+1}}) = o(n^{-2})$, $\xi_{q,n}(k_n^{-\frac{rp}{d_x+1}} \vee j_n^{-\frac{rq}{d_x+1}}) = o(n^{-\frac{1}{4}})$ for $\xi_{q,n} \equiv \sup_{(x,t)\in\mathcal{X}\times[0,1]}\|q^{j_n}(x,t)\|$.*

Our previous conditions only limit h_j^* to satisfy $E[(h_j^*(X_i,T_i))^2] < \infty$ by virtue of $h_j^* \in \bar{\mathcal{H}}$ for all $1 \le j \le d_\theta$. Assumptions 4.1(ii) and 4.2(ii) impose additional smoothness restrictions on h_j^* by requiring that they be bounded and well approximated by the sieve $\{q_j\}_{j=1}^\infty$. In verifying Assumption 4.1(ii), the comments of Remark 3.1 are applicable. Assumption 4.2(iii) is analogous to 3.2(i), and enables us to bound the rate of convergence of $\hat{\alpha}$ under $\|\cdot\|_{c,0}$ employing Theorem 3.1. The rate requirements in Assumption 4.2(iv) in turn imply $\|\hat{\alpha}-\alpha_0\|_{c,0} = o_p(n^{-\frac{1}{4}})$, and they may be verified as discussed in Remark 3.1. These stronger rate requirements are needed because the norm $\|\cdot\|_{c,L^2}$ is too weak to control the linearization error present as a result of $\alpha \mapsto W(x,t,\alpha)$ depending on α in a nonlinear way.

Theorem 4.1 establishes the asymptotic normality of the proposed estimator.

Theorem 4.1. *If Assumptions 2.1, 2.2, 3.1, 3.2, 3.3, 3.4, 4.1 and 4.2 hold, then it follows that:*

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{L} N(0,\Sigma)$$

where $\Sigma = \Gamma^{-1}\Omega\Gamma^{-1}$, $\Gamma \equiv E[R_{h^*}(X_i,T_i)R'_{h^*}(X_i,T_i)]$ and $\Omega \equiv E[R_{h^*}(X_i,T_i)R'_{h^*}(X_i,T_i)T_i(1-T_i)]$.

The asymptotic variance of the estimator implicitly depends on the distribution of the random variable T_i. Deriving the optimal choice of distribution for T_i is a challenging problem beyond the scope of this paper. A complication of the analysis is the lack of a clear lower bound for the asymptotic variance of the proposed estimator. In particular, we show in Lemma 4.1 below, that the model is not regular in the sense of Bickel et al. (1993), and therefore a semiparametric efficiency

bound may not be derived. Heuristically, this is the result of the support of the joint distribution of (Y_i, X_i) depending on the parameter of interest (θ, h) – a setting that often leads to the failure of regularity conditions for maximum likelihood estimation.

Lemma 4.1. *Let Assumptions 2.1, 3.1(iii), 3.3(i)-(ii), 3.4(i), 4.2(i) hold, and for some* $\lambda \in \mathbf{R}^{d_\theta}$

$$\max\Big\{P\Big(\frac{dm(X_i, h_0(X_i,1),\theta_0)}{d\theta}[\lambda] \neq 0\Big), P\Big(\frac{dm(X_i, h_0(X_i,0),\theta_0)}{d\theta}[\lambda] \neq 0\Big)\Big\} > 0 \ . \tag{31}$$

It then follows that the model $Y_i = m(X_i, h_0(X_i, U_i), \theta)$ *is not differentiable in quadratic mean at* θ_0.

We conclude this discussion by employing Theorem 4.1 to establish the asymptotic normality of the proposed estimator in the semiparametric transformation model (M). In addition, we also illustrate Lemma 4.1 by showing this model is not differentiable in quadratic mean.

Proposition 4.1. *Suppose (i)* $\|h_j\|_{\kappa,\infty} < \infty$ *for* $1 \leq j \leq d_\theta$, *(ii)* $(y,x) \mapsto f_{Y|X}(y,x), \lambda(\cdot,\beta)$ *and* $\nabla_\beta \Lambda(\cdot, \beta_0)$ *are* r_p*-times differentiable with bounded derivatives. If Assumptions 4.2(i), 4.2(iii), 4.2(iv) and all Assumptions of Proposition 3.2 hold, then it follows that:*

$$\sqrt{n}\begin{pmatrix} \hat\beta - \beta_0 \\ \hat\gamma - \gamma_0 \end{pmatrix} \xrightarrow{L} N(0, \Sigma) \ . \tag{32}$$

Furthermore, the model is not differentiable in quadratic mean at (β_0, γ_0).

4.1 Weighted Bootstrap

Estimating the asymptotic variance of $\sqrt{n}(\hat\theta - \theta_0)$ is a nontrivial problem as it requires estimation of R_{h^*}, implicitly defined through (24)-(26). For this reason, we instead follow Chen and Pouzo (2009) and employ a simpler to implement weighted bootstrap procedure.

Let $\{B_i\}_{i=1}^n$ be an i.i.d. sequence of random weights generated independently of $\{Y_i, X_i, T_i\}_{i=1}^n$. We require $E[B_i] = 1$, $Var(B_i) = 1$, and define the "perturbed" version of $\hat W(x,t,\alpha)$ by:

$$\hat W^*(x,t,\alpha) \equiv p^{k_n'}(x,t)(P'P)^{-1}\sum_{i=1}^n p^{k_n}(X_i,T_i)(1\{Y_i \leq m(X_i,T_i,\alpha)\} - T_i)B_i \ . \tag{33}$$

Similarly, we define $\hat\alpha^*$ to be the estimator based on the "perturbed" criterion function:

$$\hat\alpha^* \in \arg\min_{\alpha\in\mathcal{A}_n} \frac{1}{n}\sum_{i=1}^n \Big(\hat W^*(X_i,T_i,\alpha)\Big)^2 \ . \tag{34}$$

For $(\hat\theta^*, \hat h^*) \equiv \hat\alpha^*$, the asymptotic distribution of $\sqrt{n}(\hat\theta^* - \hat\theta)$ conditional on the data $\{Y_i, X_i, T_i\}_{i=1}^n$ (but not $\{B_i\}_{i=1}^n$), coincides with the unconditional asymptotic distribution of $\sqrt{n}(\hat\theta - \theta_0)$. This conclusion holds under the following additional assumption on the distribution of B_i:

Assumption 4.3. *(i)* $\{B_i\}_{i=1}^n$ *is i.i.d., independent of* $\{Y_i, X_i, T_i\}_{i=1}^n$*; (ii)* $E[B_i] = 1$, $Var(B_i) = 1$.

The requirements of Assumption 4.3 are remarkably weak. Popular choices for B_i include setting $B_i = Z_i + 1$, where Z_i is a standard normal random variable, or $B_i = R_i + 1$ where R_i is a Rademacher random variable.[4] Consistency of the weighted bootstrap holds under essentially the same conditions employed to obtain asymptotic normality. This appealing property has been previously noted, for example, in Barbe and Bertail (1995) and Ma and Kosorok (2005) who employ it in semiparametric unconditional moment models. The following Theorem, a direct application of Theorem 4.1 in Chen and Pouzo (2009), justifies its validity in the present context as well:

Theorem 4.2. *If Assumptions 2.1, 2.2, 3.1-3.4, 4.1, 4.2 and 4.3 hold, then for any* $c \in \mathbf{R}^{d_\theta}$*:*

$$P(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq c \mid \{Y_i, X_i, T_i\}_{i=1}^n) = P(\sqrt{n}(\hat{\theta} - \theta_0) \leq c) + o_p(1) \ .$$

5 Simulation Evidence

We assess the finite sample performance of the proposed estimator through a Monte Carlo study based on the transformation model (M). Specifically, we generate data according to:

$$Y_i = \frac{1}{\beta_0} \arctan\{\beta_0(X_{1i}\gamma_0 + \log(1 + X_{2i}\xi_i))\} \ , \tag{35}$$

where X_{1i}, X_{2i} and ξ_i are mutually independent with X_{1i}, X_{2i} uniformly distributed on $[0, 1]$, ξ_i Beta distributed with parameters $\alpha = \beta = 2$, and in (35) we set $\beta_0 = \gamma_0 = 1$. In estimation, we treat $(x_2, \xi) \mapsto \log(1 + x_2\xi)$ as an unknown function, and hence (35) is a special case of model (M) with $\Lambda(a, \beta) = \beta^{-1} \arctan(a\beta)$ and $h_0(x, u) = \log(1 + x_2 F_\xi^{-1}(u))$ for F_ξ^{-1} the inverse cdf of ξ.

We let the parameter space for h_0 be given by $\mathcal{H} \equiv \{h : [0, 1]^2 \to \mathbf{R} \text{ s.t. } \|h\|_{\infty,1} \leq C\}$, which is compact under $\|\cdot\|_{\infty,0}$ by the Arzela-Ascoli theorem. For the basis functions $\{q_j\}_{j=1}^\infty$, we employ tensor product B-Splines (in (x_2, u)) of order two. Hence, the sieve for $\mathcal{H}$ is given by:

$$\mathcal{H}_n \equiv \{h : h(x_2, u) = \pi' q^{j_n}(x_2, u) \text{ for some } \pi \in \mathbf{R}^{j_n} \text{ and } \|\pi' q^{j_n}\|_{0,\infty} \leq C\} \ . \tag{36}$$

Computationally, we impose $\|\pi' q^{j_n}\|_{0,\infty} \leq C$ by verifying that $|D^\lambda \pi' q^{j_n}(x, u)| \leq C$ for all multi-indices $|\lambda| \leq 1$ and all (x, u) on a grid of $[0, 1]^2$. To estimate conditional expectations, we employ tensor product B-Splines (in (x_1, x_2, t)) of order one. The auxiliary i.i.d. sample $\{T_i\}_{i=1}^n$ was generated independently of $\{Y_i, X_i\}_{i=1}^n$ with T_i uniformly distributed on $[0, 1]$.

Tables 1 and 2 report the performance of the estimators $\hat{\gamma}$ and $\hat{\beta}$ respectively for different sample sizes N and choices of j_n, k_n and C. All results are based on one thousand replications. For both basis functions $\{q_j\}_{j=1}^{j_n}$ and $\{p_k\}_{k=1}^{k_n}$, we employed tensor products of one and two knot B-splines, corresponding to $(j_n, k_n) = (9, 8)$ and $(j_n, k_n) = (16, 22)$ respectively. The smoothness constraint C was set to equal 1 and 5,[5] with larger values of C yielding similar results. We do not report results

[4] A Rademacher random variables satisfies $P(R_i = 1) = P(R_i = -1) = 1/2$

[5] Note that $\|h_0\|_{0,\infty} = 1$, and hence we consider $C = 1$ to be a tight constraint.

Table 1: Simulation Result for Estimate of γ_0

	$N = 200$, $C = 5$					$N = 200$, $C = 1$				
Sieve Choice	Bias	MSE	25%	50%	75%	Bias	MSE	25%	50%	75%
$(k_n, j_n) = (8, 9)$	0.0048	0.0119	0.9426	0.9966	1.0602	-0.0040	0.0099	0.9385	0.9937	1.0525
$(k_n, j_n) = (8, 16)$	-0.0003	0.0054	0.9622	0.9975	1.0367	-0.0067	0.0065	0.9547	0.9951	1.0360
$(k_n, j_n) = (22, 9)$	–	–	–	–	–	–	–	–	–	–
$(k_n, j_n) = (22, 16)$	–	–	–	–	–	–	–	–	–	–
	$N = 500$, $C = 5$					$N = 500$, $C = 1$				
Sieve Choice	Bias	MSE	25%	50%	75%	Bias	MSE	25%	50%	75%
$(k_n, j_n) = (8, 9)$	0.0023	0.0066	0.9562	0.9959	1.0395	-0.0028	0.0050	0.9574	0.9943	1.0365
$(k_n, j_n) = (8, 16)$	-0.0027	0.0037	0.9663	0.9955	1.0267	-0.0036	0.0033	0.9653	0.9953	1.0259
$(k_n, j_n) = (22, 9)$	-0.0004	0.0011	0.9780	1.0010	1.0208	-0.0011	0.0011	0.9778	1.0001	1.0200
$(k_n, j_n) = (22, 16)$	0.0005	0.0037	0.9683	1.0007	1.0332	-0.0070	0.0020	0.9669	0.9954	1.0213
	$N = 1000$, $C = 5$					$N = 1000$, $C = 1$				
Sieve Choice	Bias	MSE	25%	50%	75%	Bias	MSE	25%	50%	75%
$(k_n, j_n) = (8, 9)$	0.0006	0.0032	0.9678	0.9983	1.0328	-0.0001	0.0025	0.9680	0.9980	1.0306
$(k_n, j_n) = (8, 16)$	-0.0001	0.0024	0.9719	0.9989	1.0241	-0.0014	0.0020	0.9728	0.9976	1.0245
$(k_n, j_n) = (22, 9)$	-0.0001	0.0005	0.9855	0.9993	1.0153	-0.0001	0.0005	0.9855	0.9993	1.0153
$(k_n, j_n) = (22, 16)$	0.0018	0.0018	0.9769	0.9998	1.0266	-0.0027	0.0011	0.9770	0.9978	1.0197

Table 2: Simulation Result for Estimate of β_0

	$N = 200$, $C = 5$					$N = 200$, $C = 1$				
Sieve Choice	Bias	MSE	25%	50%	75%	Bias	MSE	25%	50%	75%
$(k_n, j_n) = (8, 9)$	-0.0145	0.0202	0.9187	0.9992	1.0651	-0.0232	0.0203	0.9144	0.9933	1.0635
$(k_n, j_n) = (8, 16)$	-0.0110	0.0109	0.9350	0.9994	1.0466	-0.0218	0.0157	0.9288	0.9958	1.0465
$(k_n, j_n) = (22, 9)$	–	–	–	–	–	–	–	–	–	–
$(k_n, j_n) = (22, 16)$	–	–	–	–	–	–	–	–	–	–
	$N = 500$, $C = 5$					$N = 500$, $C = 1$				
Sieve Choice	Bias	MSE	25%	50%	75%	Bias	MSE	25%	50%	75%
$(k_n, j_n) = (8, 9)$	-0.0089	0.0102	0.9369	0.9935	1.0473	-0.0139	0.0098	0.9392	0.9930	1.0444
$(k_n, j_n) = (8, 16)$	-0.0104	0.0067	0.9506	0.9940	1.0357	-0.0115	0.0069	0.9506	0.9940	1.0357
$(k_n, j_n) = (22, 9)$	-0.0037	0.0022	0.9672	0.9978	1.0282	-0.0046	0.0021	0.9670	0.9971	1.0277
$(k_n, j_n) = (22, 16)$	-0.0059	0.0063	0.9584	0.9998	1.0394	-0.0129	0.0036	0.9548	0.9923	1.0260
	$N = 1000$, $C = 5$					$N = 1000$, $C = 1$				
Sieve Choice	Bias	MSE	25%	50%	75%	Bias	MSE	25%	50%	75%
$(k_n, j_n) = (8, 9)$	-0.0046	0.0053	0.9567	1.0007	1.0399	-0.0043	0.0041	0.9582	1.0002	1.0379
$(k_n, j_n) = (8, 16)$	-0.0043	0.0036	0.9588	0.9981	1.0325	-0.0054	0.0035	0.9608	0.9966	1.0308
$(k_n, j_n) = (22, 9)$	-0.0019	0.0010	0.9775	0.9986	1.0195	-0.0020	0.0010	0.9775	0.9986	1.0194
$(k_n, j_n) = (22, 16)$	-0.0010	0.0025	0.9686	1.0010	1.0322	-0.0059	0.0017	0.9688	0.9963	1.0220

for $N = 200$ and $k_n = 22$ as the regression routine used to construct $\hat{W}(x, t, \alpha)$ failed in a large number of the simulations.

The performance of both estimators $\hat{\gamma}$ and $\hat{\beta}$ is robust to the different choices of j_n, k_n and C we considered. Across specifications, increasing j_n or k_n from $(j_n, k_n) = (8, 9)$ leads to improvements in mean squared error, but setting j_n and k_n to both be large affects it adversely. For both estimators $\hat{\gamma}$ and $\hat{\beta}$, setting $C = 1$ leads to a general better performance than $C = 5$, which is analogous to

setting a large penalty in the penalized minimum distance estimator of Chen and Pouzo (2011). Overall, we find the performance of the semiparametric estimator to be encouraging.

6 Conclusion

Exploiting the insight that nonseparable models may be rewritten as conditional mean restriction models, we have proposed a simple to implement semiparametric estimator. We have derived its rate of convergence and established the asymptotic normality of the parametric component. A bootstrap procedure may be employed for inference, and a brief simulation study shows good performance.

Optimality in these models remains an open question. As Lemma 4.1 shows, addressing it may require a different analysis than the usual semiparametric efficiency bound calculation. We hope to study this problem in future work.

APPENDIX

The following is a table of the notation and definitions that will be used throughout the appendix.

$a \lesssim b$	$a \leq Mb$ for some constant M which is universal in the context of the proof.
$\|\cdot\|_{\infty,\kappa}$	The norm $\|h\|_{\infty,\kappa} = \max_{\|\lambda\|\leq\kappa} \sup_{x,u} \|D^\lambda h(x,u)\|$.
$\|\cdot\|_{c,\kappa}$	The norm $\|\alpha\|_{c,\kappa} = \|\theta\| + \|h\|_{\infty,\kappa}$.
$N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|)$	The bracketing numbers of size ϵ for $\mathcal{F}$ under the norm $\|\cdot\|$.
$\bar{W}(x,t,\alpha)$	The function $\bar{W}(x,t,\alpha) \equiv p^{k'_n}(x,t)(P'P)^{-1}\sum_i p^{k_n}(x_i,t_i)W(x_i,t_i,\alpha)$.

PROOF OF PROPOSITION 2.1: Suppose by way of contradiction that there exist $(\beta,\gamma,h) \neq (\beta_0,\gamma_0,h_0)$ such that:

$$P\left(\Lambda(X_{1i}'\gamma + h(X_{2i},U_i),\beta) = \Lambda(X_{1i}'\gamma_0 + h_0(X_{2i},U_i),\beta_0)\right) = 1\ . \tag{37}$$

Since X_i is continuously distributed, all $h \in \mathcal{H}$ are continuous, and $a \mapsto \Lambda(a,\beta)$ is also continuous, we may conclude

$$\Lambda(x_1'\gamma + h(x_2,u),\beta) = \Lambda(x_1'\gamma_0 + h_0(x_2,u),\beta_0)\ , \tag{38}$$

for all $(x,u) \in \mathcal{X} \times [0,1]$ and $x \equiv (x_1,x_2)$. Let $\mathcal{X}^o$ denote the interior of $\mathcal{X}$, and observe $P(X_i \in \mathcal{X}^o) = 1$. It then follows by differentiating both sides of (38) with respect to x_1 at any $(x,u) \in \mathcal{X}^o \times (0,1)$ that we must have:

$$\lambda(X_{1i}'\gamma + h(X_{2i},U_i),\beta)\gamma = \lambda(X_{1i}'\gamma_0 + h_0(X_{2i},U_i),\beta_0)\gamma_0 \qquad a.s. \tag{39}$$

Thus, we must have $P(\lambda(X_{1i}'\gamma + h(X_{2i},U_i),\beta) = c_0\lambda(X_{1i}'\gamma_0 + h_0(X_{2i},U_i),\beta_0)) = 1$ for some $c_0 \in \mathbf{R}$. Similarly, let $x_2 \in \mathbf{R}^{d_{x_2}}$ and $x_2^{(j)}$ denote the j^{th} coordinate of the vector x_2. Differentiating (38) again we then obtain:

$$\frac{d}{du}h(X_{2i},U_i) = c_0\frac{d}{du}h_0(X_{2i},U_i) \qquad \frac{d}{dx_2^{(j)}}h(X_{2i},U_i) = c_0\frac{d}{dx_2^{(j)}}h_0(X_{2i},U_i)\ \ \forall\ 1 \leq j \leq d_{x_2}\ a.s. \tag{40}$$

Denote $x_2^{(-j)}$ be the x_2 vector minus its j^{th} component. Integrating all equations in (40) then implies that:

$$h(X_{2i},U_i) = c_0h_0(X_{2i},U_i) + m_u(X_{2i}) \qquad h(X_{2i},U_i) = c_0h_0(X_{2i},U_i) + m_j(X_2^{(-j)},U_i)\ \ \forall\ 1 \leq j \leq d_{x_2}\ a.s. \tag{41}$$

Since all $h \in \mathcal{H}$ are differentiable, m_u and m_j must be differentiable as well. Therefore, exploiting (41) we obtain:

$$P(m_u(X_{2i}) = m_j(X_{2i}^{(-j)},U_i)) = 1 \quad \Rightarrow \quad P(\frac{d}{dx_2^{(j)}}m_u(X_{2i}) = 0) = 1\ . \tag{42}$$

We conclude from (42) that m_u is constant almost surely. Thus, it follows from (39) and (41) that for some $c_1 \in \mathbf{R}$:

$$\begin{aligned} P(\Lambda(X_{1i}'\gamma + h(X_{2i},U_i),\beta) &= \Lambda(X_{1i}'\gamma_0 + h_0(X_{2i},U_i),\beta_0)) \\ &= P(\Lambda(c_0(X_{1i}'\gamma_0 + h_0(X_{2i},U_i)) + c_1,\beta_0) = \Lambda(X_{1i}'\gamma_0 + h_0(X_{2i},U_i))) \\ &= P(c_0\Lambda^{-1}(Y_i,\beta_0) + c_1 = \Lambda^{-1}(Y_i,\beta))\ . \end{aligned} \tag{43}$$

However, by condition (v) the final expression is equal to one only if $(c_0,c_1) = (1,0)$ and $\beta = \beta_0$. ■

Lemma A.1. *Let $\bar{W}(x,t,\alpha) \equiv p^{k'_n}(x,t)(P'P)^{-1}\sum_i p^{k_n}(X_i,T_i)W(X_i,T_i,\alpha)$. If Assumptions 2.1(i)-(iv), 3.1(i)-(iii) and 3.2(i)-(iii) hold, it then follows that (i) $\sup_{\alpha\in\mathcal{A}_n,\alpha_0} \frac{1}{n}\sum_i(\hat{W}(X_i,T_i,\alpha) - W(X_i,T_i,\alpha))^2 = O_p(\frac{k_n}{n} + k_n^{-\frac{2r_p}{d_x+1}})$ and in addition (ii) $\sup_{\alpha\in\mathcal{A}_n} E[(\bar{W}(X_i,T_i,\alpha) - W(X_i,T_i,\alpha))^2] = O_p(k_n^{-\frac{2r_p}{d_x+1}})$.*

PROOF: In this proof we follow Lemmas B.2 and B.3 in Chen and Pouzo (2011). As noted in Newey (1997), under Assumption 3.2(i) we may assume without loss of generality that $E[p^{k_n}(X_i,T_i)p^{k'_n}(X_i,T_i)] = I$. Then notice:

$$\begin{aligned} &\sup_{\alpha\in\mathcal{A}_n,\alpha_0} \frac{1}{n}\sum_{i=1}^n(\hat{W}(X_i,T_i,\alpha) - W(X_i,T_i,\alpha))^2 \\ &\qquad \leq \sup_{\alpha\in\mathcal{A}_n,\alpha_0} \frac{2}{n}\sum_{i=1}^n(\hat{W}(X_i,T_i,\alpha) - \bar{W}(X_i,T_i,\alpha))^2 + \sup_{\alpha\in\mathcal{A}_n,\alpha_0} \frac{2}{n}\sum_{i=1}^n(\bar{W}(X_i,T_i,\alpha) - W(X_i,T_i,\alpha))^2\ . \end{aligned} \tag{44}$$

Let G_n denote the linear span of p^{k_n} and notice $\hat{W}(\cdot,\cdot,\alpha), \bar{W}(\cdot,\cdot,\alpha) \in G_n$ for all $\alpha \in \mathcal{A}$. Also, for any $g = \sum_{j=1}^{k_n} a_j p_j$:

$$\frac{\|g\|_\infty}{\|g\|_{L^2}} = \sup_{(x,t)\in\mathcal{X}\times[0,1]} \frac{|\sum_{j=1}^{k_n} a_j p_j(x,t)|}{\left[\sum_j a_j^2\right]^{\frac{1}{2}}} \leq \sup_{(x,t)\in\mathcal{X}\times[0,1]} \|p^{k_n}(x,t)\| \ , \tag{45}$$

where the equality and inequality follow by $E[p^{k_n}(X_i,T_i)p^{k_n'}(X_i,T_i)] = I$ and Cauchy-Schwarz respectively. Let $A_n = \sup_{G_n} \|g\|_\infty/\|g\|_{L^2}$ and note that (45) and Assumption 3.2(iii) implies $A_n^2 \times k_n/n \to 0$. Therefore, Lemma 2.3(i) in Huang (2003) implies that with probability tending to one we have:

$$\frac{1}{2}E[g^2(X_i,T_i)] \leq \frac{1}{n}\sum_{i=1}^n g^2(X_i,T_i) \leq 2E[g^2(X_i,T_i)] \qquad \forall g \in G_n \ . \tag{46}$$

Let $\epsilon_i(\alpha) \equiv \{(1\{Y_i \leq m(X_i,T_i,\alpha)\} - T_i) - (P(Y_i \leq m(X_i,T_i,\alpha)|X_i,T_i) - T_i)\}$, $\epsilon(\alpha) \equiv (\epsilon_1(\alpha),\ldots,\epsilon_n(\alpha))'$ and note:

$$\begin{aligned}\sup_{\alpha\in\mathcal{A}} E[(\hat{W}(X_i,T_i,\alpha) - \bar{W}(X_i,T_i,\alpha))^2] &= \sup_{\alpha\in\mathcal{A}} \text{trace}\{E[p^{k_n'}(X_i,T_i)(P'P)^{-1}P'\epsilon(\alpha)\epsilon(\alpha)'P(P'P)^{-1}p^{k_n}(X_i,T_i)]\} \\ &= \sup_{\alpha\in\mathcal{A}} \text{trace}\{(P'P)^{-1}P'\epsilon(\alpha)\epsilon(\alpha)'P(P'P)^{-1}\} \ , \end{aligned} \tag{47}$$

where the second equality is implied by $E[p^{k_n}(X_i,T_i)p^{k_n'}(X_i,T_i)] = I$. Taking expectations conditional on $\{X_i,T_i\}_{i=1}^n$, noticing that $E[\epsilon_i^2(\alpha)|\{X_i,T_i\}_{i=1}^n] \leq 1$ and $E[\epsilon_l(\alpha)\epsilon_j(\alpha)|\{X_i,T_i\}_{i=1}^n] = 0$ for all $l \neq j$, we obtain:

$$\sup_{\alpha\in\mathcal{A}} E[\text{trace}\{(P'P)^{-1}P'\epsilon(\alpha)\epsilon(\alpha)'P(P'P)^{-1}\}|\{X_i,T_i\}_{i=1}^n] \leq \text{trace}\{P(P'P)^{-\frac{1}{2}}(P'P)^{-1}(P'P)^{-\frac{1}{2}}P\} \ . \tag{48}$$

However, by Theorem 1 in Newey (1997), the smallest eigenvalue of $P'P/n$ converges in probability to 1. Therefore, it follows by results (47) and (48) together with Markov's inequality that:

$$\sup_{\alpha\in\mathcal{A}} E[(\hat{W}(X_i,T_i,\alpha) - \bar{W}(X_i,T_i,\alpha))^2] = O_p(k_n/n) \ . \tag{49}$$

Next, using Assumptions 3.1(i)-(ii), 3.2(i)-(ii) and Lemma A.1(B) in Ai and Chen (2003) we may conclude that:

$$\sup_{\alpha\in\mathcal{A}_n,\alpha_0} \frac{1}{n}\sum_{i=1}^n (\bar{W}(X_i,T_i,\alpha) - W(X_i,T_i,\alpha))^2 = O_p(k_n^{-\frac{2r_p}{d_x+1}}) \ . \tag{50}$$

Therefore, the first claim of the Lemma follows from (49) and (50), together with the inequalities in (44) and (46).

For the second claim of the Lemma, let $\bar{\epsilon}_i(\alpha) \equiv W(X_i,T_i,\alpha) - \pi_n(\alpha)p^{k_n}(X_i,T_i)$ and $\bar{\epsilon}(\alpha) \equiv (\bar{\epsilon}_1(\alpha),\ldots,\bar{\epsilon}_n(\alpha))'$. It then follows from Assumption 3.2(ii) and arguing as in (47) that:

$$\begin{aligned}\sup_{\alpha\in\mathcal{A}_n} E[(\bar{W}(X_i,T_i,\alpha) - W(X_i,T_i,\alpha))^2] &\leq 2E[(p^{k_n'}(X_i,T_i)(P'P)^{-1}P'\bar{\epsilon}(\alpha))^2] + O(k_n^{-\frac{2r_p}{d_x+1}}) \\ &= \sup_{\alpha\in\mathcal{A}_n} 2\text{trace}\{(P'P)^{-1}P'\bar{\epsilon}(\alpha)\bar{\epsilon}(\alpha)'P(P'P)^{-1}\} + O(k_n^{-\frac{2r_p}{d_x+1}}) = \sup_{\alpha\in\mathcal{A}_n} 2\|(P'P)^{-1}P'\bar{\epsilon}(\alpha)\|^2 + O(k_n^{-\frac{2r_p}{d_x+1}}) \ . \end{aligned} \tag{51}$$

Using once again that the smallest eigenvalue of $P'P/n$ converges to 1 by Theorem 1 in Newey (1997), and noticing that the matrix $P(P'P)^{-1}P$ is idempotent, we conclude that with probability tending one:

$$\sup_{\alpha\in\mathcal{A}_n} \|(P'P)^{-1}P'\bar{\epsilon}(\alpha)\|^2 \leq \sup_{\alpha\in\mathcal{A}_n} \bar{\epsilon}(\alpha)'P(P'P)^{-1}P'\bar{\epsilon}(\alpha)/n \leq \sup_{\alpha\in\mathcal{A}_n} \|\bar{\epsilon}(\alpha)\|^2/n = O(k_n^{-\frac{2r_p}{d_x+1}}) \ . \tag{52}$$

Combining (51) and (52) establishes the second claim of the Lemma. ■

Proof of Lemma 3.1: The proof proceeds by verifying the conditions of Corollary 3.1 in Chen and Pouzo (2011). Towards this end, first notice that the mean value theorem, Assumption 3.3(iv) and (4) imply that:

$$Q(\alpha) = E[(P(Y_i \leq m(X_i,T_i,\alpha)|X_i,T_i) - P(Y_i \leq m(X_i,T_i,\alpha_0)|X_i,T_i))^2] \asymp E[(m(X_i,T_i,\alpha) - m(X_i,T_i,\alpha_0))^2] \tag{53}$$

Hence, by Assumption 2.2 α_0 is the unique zero of Q on $\mathcal{A}$, which together with Assumptions 3.1(i) and 3.3(i) verifies Assumption 3.1 in Chen and Pouzo (2011). In turn, their Assumption, 3.2(i), 3.6(i) and 3.7(i) are implied by our

Assumption 3.3(i)-(iii). Moreover, by the Cauchy-Schwarz inequality and (53) we obtain:

$$\begin{aligned}(Q(\alpha)-Q(\tilde{\alpha}))^2 &\leq 4E[(P(Y_i \leq m(X_i,T_i,\alpha)|X_i,T_i) - P(Y_i \leq m(X_i,T_i,\tilde{\alpha})|X_i,T_i))^2] \\ &\lesssim E[(m(X_i,T_i,\alpha)-m(X_i,T_i,\tilde{\alpha}))^2] \\ &\leq E[G^2(X_i,T_i)]\|\alpha-\tilde{\alpha}\|_{c,\kappa}^2 \ , \end{aligned} \tag{54}$$

where the final inequality follows by Assumption 3.3(iv). It follows that Q is continuous under $\|\cdot\|_{c,\kappa}$, which verifies their Assumption 3.2(ii) and 3.7(ii). Their requirement 3.4 and 3.6(ii) are trivially satisfied as we do not use a penalty function, and their Assumptions 3.5(i) and 3.5(iii) are established by our Assumption 3.3(iv) arguing as in (54) and Lemma A.1(i). Finally, notice that $(x,u,\alpha) \mapsto m(x,u,\alpha)$ is continuous with respect to $\max\{\|\cdot\|, \|\cdot\|_{c,\kappa}\}$ due to all $h \in \mathcal{H}$ being continuous. Therefore, by compactness of $\mathcal{X} \times [0,1] \times \mathcal{A}$, it follows that $(x,u,\alpha) \mapsto m(x,u,\alpha)$ is uniformly bounded on $\mathcal{X} \times [0,1] \times \mathcal{A}$, verifying Assumption 3.5(ii) of Chen and Pouzo (2011). Hence, all conditions for Corollary 3.1 in Chen and Pouzo (2011) are met and the Lemma follows. ■

Proof of Proposition 3.1: Our assumptions imply $\sup_{\alpha\in\mathcal{A}} \|W(\cdot,\cdot,\alpha)\|_{\infty,r_p} < \infty$, and hence our choice of sieve ensures Assumption 3.2(ii) and 3.3(iii) are satisfied; see Chen (2007). To verify Assumption 3.3(iv), first notice that:

$$\begin{aligned}&|\Lambda(x_1'\gamma + h(x_2,u),\beta) - \Lambda(x_1'\tilde{\gamma} + \tilde{h}(x_2,u),\tilde{\beta})| \\ &\qquad \leq |\Lambda(x_1'\gamma + h(x_2,u),\beta) - \Lambda(x_1'\gamma + h(x_2,u),\tilde{\beta})| + |\Lambda(x_1'\gamma + h(x_2,u),\tilde{\beta}) - \Lambda(x_1'\tilde{\gamma} + \tilde{h}(x_2,u),\tilde{\beta})| \ , \end{aligned} \tag{55}$$

for any $(x_1,x_2,t) \in \mathcal{X} \times [0,1]$. Thus, applying the mean value theorem it follows that for some β^* between β and $\tilde{\beta}$:

$$|\Lambda(x_1'\gamma + h(x_2,u),\beta) - \Lambda(x_1'\gamma + h(x_2,u),\tilde{\beta})| = |\nabla_\beta \Lambda(x_1'\gamma + h(x_2,u),\beta^*)[\beta-\tilde{\beta}]| \lesssim \|\beta-\tilde{\beta}\| \ , \tag{56}$$

where the second inequality follows by Cauchy Schwarz and the assumption that $\nabla_\beta\Lambda(a,\beta)$ is uniformly bounded. Similarly, applying the mean value theorem we obtain that for some a^* between $x_1'\gamma + h(x_2,u)$ and $x_1'\tilde{\gamma} + \tilde{h}(x_2,u)$:

$$|\Lambda(x_1'\gamma+h(x_2,u),\tilde{\beta})-\Lambda(x_1'\tilde{\gamma}+\tilde{h}(x_2,u),\tilde{\beta})| = |\lambda(a^*,\tilde{\beta})| \times |x_1'(\gamma-\tilde{\gamma})+h(x_2,u)-\tilde{h}(x_{,2},u)| \lesssim \|\gamma-\tilde{\gamma}\|+\|h-\tilde{h}\|_{\infty,\kappa} \tag{57}$$

where the inequality is implied by $\lambda(a,\tilde{\beta})$ bounded, Cauchy Schwarz, $\mathcal{X}$ being compact and $\kappa \geq 0$. Thus, (55)-(57) establish Assumption 3.3(iv) holds and the claim of the Proposition follows by Lemma 3.1. ■

Lemma A.2. *Under Assumptions 2.1(iii), 3.1, 3.3(i), 3.3(iv) and 3.4(i)-(ii), (i)* $E[(\frac{dm(X_i,T_i,\tilde{\alpha})}{d\alpha}[\bar{\alpha}])^2] \leq M\|\bar{\alpha}\|_{c,L^2}^2$ *for some* $M > 0$ *and all* $(\tilde{\alpha},\bar{\alpha}) \in \mathcal{A} \times \bar{\mathcal{A}}$*, (ii)* $\exists c_0, c_1 > 0$ *with* $c_0\|\bar{\alpha}\|_{c,L^2} \leq E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\bar{\alpha}])^2] \leq c_1\|\bar{\alpha}\|_{c,L^2}^2$ $\forall \bar{\alpha} \in \bar{\mathcal{A}}$.

Proof: First observe that since $h \in \mathcal{H}$ are uniformly bounded and $\mathcal{X}, \Theta$ are compact, $\sup_{(h,\theta)\in\mathcal{H}\times\Theta} \|(X_i, h(X_i,T_i),\theta)\|$ is contained in a compact set almost surely. Thus, since derivatives of $m(x,e,\theta)$ in (e,θ) are continuous, they are also almost surely uniformly bounded in $(h(x,u),\theta)$ for $(h,\theta) \in \mathcal{H} \times \Theta$. Hence, for all $(\bar{h},\bar{\theta}) = \bar{\alpha} \in \bar{\mathcal{A}}$, we have:

$$\begin{aligned}\sup_{\tilde{\alpha}\in\mathcal{A}} E[(\frac{dm(X_i,T_i,\tilde{\alpha})}{d\alpha}[\bar{\alpha}])^2] &= \sup_{\tilde{\alpha}\in\mathcal{A}} E[(\frac{dm(X_i,T_i,\tilde{\alpha})}{d\theta}[\bar{\theta}] + \frac{dm(X_i,T_i,\tilde{\alpha})}{dh}[\bar{h}])^2] \\ &\leq \sup_{\tilde{\alpha}\in\mathcal{A}} 2E[(\frac{dm(X_i,T_i,\tilde{\alpha})}{d\theta}[\bar{\theta}])^2] + \sup_{\tilde{\alpha}\in\mathcal{A}} 2E[(\frac{dm(X_i,T_i,\tilde{\alpha})}{dh}[\bar{h}])^2] \lesssim \|\bar{\theta}\|^2 + \|\bar{h}\|_{L^2}^2 \leq \|\bar{\alpha}\|_{c,L^2}^2 \ . \end{aligned} \tag{58}$$

For the second claim of the lemma, notice that (58) implies $E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\bar{\alpha}])^2] \leq c_1\|\bar{\alpha}\|_{c,L^2}^2$ for all $\bar{\alpha} \in \bar{\mathcal{A}}$ for some $c_1 > 0$. Also observe that the following orthogonality condition must hold for all $\bar{h} \in \bar{\mathcal{H}}$ by definition of h^*:

$$E[f_{Y|X}^2(m(X_i,T_i,\alpha_0)|X_i)(\frac{dm(X_i,T_i,\alpha_0)}{d\theta} - \frac{dm(X_i,T_i,\alpha_0)}{dh}[h^*])\frac{dm(X_i,T_i,\alpha_0)}{dh}[\bar{h}]] = 0 \ . \tag{59}$$

Hence, since $f_{Y|X}$ is bounded above by Assumption 3.3(iv), result (59) and the definition of R_{h^*} yield that:

$$\begin{aligned}E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\bar{\alpha}])^2] &\gtrsim E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\bar{\alpha}])^2 f_{Y|X}^2(m(X_i,T_i,\alpha_0)|X_i)] \\ &= \bar{\theta}' E\left[R_{h^*}(X_i,T_i)R_{h^*}'(X_i,T_i)\right]\bar{\theta} + E[(\frac{dm(X_i,T_i,\alpha_0)}{dh}[h^{*\prime}\bar{\theta}+\bar{h}])^2 f_{Y|X}^2(m(X_i,T_i,\alpha_0)|X_i)] \ . \end{aligned} \tag{60}$$

Moreover, $E[R_{h^*}(X_i,T_i)R'_{h^*}(X_i,T_i)]$ positive definite by Assumption 3.4(ii), (60) and Assumption 3.3(iv) imply:

$$E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\bar{\alpha}])^2] \gtrsim \|\bar{\theta}\|^2 + E[(\frac{dm(X_i,T_i,\alpha_0)}{dh}[h^{*\prime}\bar{\theta}+\bar{h}])^2] \ . \tag{61}$$

However, as h_0 is continuous, it follows that $\frac{dm(x,e,\theta_0)}{de}|_{e=h_0(x,u)}$ is continuous as well by Assumption 3.4(i). Therefore, by compactness of $\mathcal{X}\times[0,1]$ and Assumption 2.1(iii), we obtain that for some $\epsilon>0$:

$$\inf_{(x,u)\in\mathcal{X}\times[0,1]} \frac{dm(x,e,\theta_0)}{de}\Big|_{e=h_0(x,u)} > \epsilon \ . \tag{62}$$

Thus, result (62) $E[h_j^*(X_i,T_i)]<\infty$ for all $1\leq j\leq d_\theta$ and $\frac{dm(X_i,T_i,\alpha)}{dh}$ being uniformly bounded almost surely, imply:

$$\begin{aligned} E[\bar{h}^2(X_i,T_i)] &\lesssim E[(\frac{dm(X_i,T_i,\alpha_0)}{dh}[\bar{h}])^2] \\ &\leq 2E[(\frac{dm(X_i,T_i,\alpha_0)}{dh}[h^{*\prime}\bar{\theta}+\bar{h}])^2] + 2E[(\frac{dm(X_i,T_i,\alpha_0)}{dh}[h^{*\prime}\bar{\theta}])^2] \\ &\lesssim E[(\frac{dm(X_i,T_i,\alpha_0)}{dh}[h^{*\prime}\bar{\theta}+\bar{h}])^2] + \|\bar{\theta}\|^2 \ . \end{aligned} \tag{63}$$

In turn, (61) and (63) establish there is a $c_0>0$ such that $c_0\|\bar{\alpha}\|^2_{c,L^2} \leq E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\bar{\alpha}])^2]$, concluding the proof. ■

Lemma A.3. *Under Assumptions 2.1(iii), 3.1(ii)-(iii), 3.3(i), 3.3(iv) and 3.4(i)-(ii), there exists $\mathcal{N}(\alpha_0)$ a $\|\cdot\|_{c,\kappa}$ neighborhood of α_0 and $c_0,c_1>0$ so that $c_0\|\alpha-\alpha_0\|^2_{c,L^2} \leq Q(\alpha) \leq c_1\|\alpha-\alpha_0\|^2_{c,L^2}$, for all $\alpha\in\mathcal{N}(\alpha_0)\cap\mathcal{A}$.*

Proof: Recall that as argued in Lemma A.2 the first two derivatives of $m(x,e,\theta)$ in (e,θ) are uniformly bounded evaluated on $(x,h(x,u),\theta)$ for $(x,u)\in\mathcal{X}\times[0,1]$ and $(h,\theta)\in\mathcal{H}\times\Theta$. Hence, for $\theta^{(i)}$ the i^{th} coordinate of θ, we obtain

$$\Big|\frac{dm(x,t,\alpha)}{d\theta^{(i)}} - \frac{dm(x,t,\alpha_0)}{d\theta^{(i)}}\Big| = \Big|\frac{d^2m(x,t,\bar{\alpha})}{d\theta d\theta^{(i)}}[\theta-\theta_0] + \frac{d^2m(x,t,\bar{\alpha})}{dhd\theta^{(i)}}[h-h_0]\Big| \lesssim \|\theta-\theta_0\| + |h(x,t)-h_0(x,t)| \ , \tag{64}$$

by the mean value theorem for $\bar{\alpha}$ a convex combination of α and α_0, with $\bar{\alpha}\in\mathcal{A}$ by Assumption 3.3(ii). Similarly, the same manipulations as in (64) further establish that for any $\alpha\in\mathcal{A}$ we also have that:

$$\Big|\frac{dm(x,t,\alpha)}{dh} - \frac{dm(x,t,\alpha_0)}{dh}\Big| \lesssim \|\theta-\theta_0\| + |h(x,t)-h_0(x,t)| \ . \tag{65}$$

In turn, applying the mean value theorem inside the integral implies the equality in (66) for $\bar{\alpha}_{xt}$ depending on the point (x,t) at which the integrand is evaluated. Further, for $\bar{\alpha}_{xt}=(\bar{\theta}_{xt},\bar{h}_{xt})$ the first inequality in (66) follows by (64) and (65), and the second inequality is established by exploiting that $\bar{\alpha}_{xt}$ is a convex combination of α and α_0:

$$\begin{aligned} E[(m(X_i,T_i,\alpha)-&m(X_i,T_i,\alpha_0) - \frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha-\alpha_0])^2] \\ &= E[(\frac{dm(X_i,T_i,\bar{\alpha}_{xt})}{d\alpha}[\alpha-\alpha_0] - \frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha-\alpha_0])^2] \\ &\lesssim E[(\|\bar{\theta}_{xt}-\theta_0\| + |\bar{h}_{xt}(X_i,T_i)-h_0(X_i,T_i)|)^2(\|\theta-\theta_0\| + |h(X_i,T_i)-h_0(X_i,T_i)|)^2] \\ &\lesssim E[(\|\theta-\theta_0\|^2 + (h(X_i,T_i)-h_0(X_i,T_i))^2)^2] \ , \end{aligned} \tag{66}$$

where the last expectation is finite since $(h,\theta)\equiv\alpha\in\mathcal{A}$ implies $\|h\|_{\infty,\kappa}<\infty$. Hence, for any $\alpha\in\mathcal{A}$ we obtain:

$$E[(m(X_i,T_i,\alpha)-m(X_i,T_i,\alpha_0) - \frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha-\alpha_0])^2] \lesssim \|\alpha-\alpha_0\|^2_{c,L^2}\times\|\alpha-\alpha_0\|^2_{c,\kappa} \ . \tag{67}$$

To conclude notice that Assumption 3.3(iv) and the mean value theorem yields the first equality in (68), while the second equality holds for $\alpha\in\mathcal{A}$ due to result (67), the Cauchy-Schwarz inequality and Lemma A.2.

$$\begin{aligned} Q(\alpha) &\asymp E[(m(X_i,T_i,\alpha)-m(X_i,T_i,\alpha_0))^2] \\ &= E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha-\alpha_0])^2] + O(\|\alpha-\alpha_0\|^2_{c,L^2}\times\|\alpha-\alpha_0\|_{c,\kappa}) + O(\|\alpha-\alpha_0\|^2_{c,L^2}\times\|\alpha-\alpha_0\|^2_{c,\kappa}) \ . \end{aligned} \tag{68}$$

The claim of the Lemma then follows by result (68), Lemma A.2 and setting $\mathcal{N}(\alpha_0)$ so that $\|\alpha-\alpha_0\|_{c,\kappa}$ is sufficiently small for all $\alpha\in\mathcal{N}(\alpha_0)\cap\mathcal{A}$. ■

PROOF OF THEOREM 3.1: The proof proceeds by verifying the conditions for Theorem 4.1 in Chen and Pouzo (2011). Note that their Assumptions 3.1(i)-(ii), 3.2 and 3.4 were already established to hold under our Assumptions 2.1-3.3 in the proof of Lemma 3.1. Furthermore, for $\bar{W}(x,t,\alpha)$ as defined in Lemma A.1 we have:

$$\sup_{\alpha\in\mathcal{A}_n} E[(\hat{W}(X_i,T_i,\alpha)-W(X_i,T_i,\alpha))^2]$$
$$\leq \sup_{\alpha\in\mathcal{A}_n} 2E[(\hat{W}(X_i,T_i,\alpha)-\bar{W}(X_i,T_i,\alpha))^2]+\sup_{\alpha\in\mathcal{A}_n} 2E[(\bar{W}(X_i,T_i,\alpha)-W(X_i,T_i,\alpha))^2]=O_p(\frac{k_n}{n}+k_n^{-\frac{2r_p}{d_x+1}}) \quad (69)$$

due to (49) and Lemma A.1(ii). Result (69) verifies their Assumption 3.8(i) with $\delta_{m,n}^2 = k_n/n + k_n^{-\frac{2r_p}{d_x+1}}$, while their Assumption 3.8(ii) follows from (46). Further, their Assumption 4.1(i) holds by Assumption 3.3(ii), 3.3(iv) and 3.4(i), while their Assumption 4.2(ii) is established by Lemma A.3. We therefore conclude that:

$$\|\hat{\alpha}-\alpha_0\|_{c,L^2} \leq \|\hat{\alpha}-\Pi_n\alpha_0\|_{c,L^2}+\|\Pi_n\alpha_0-\alpha_0\|_{c,L^2}=O_p(\frac{\sqrt{k_n}}{\sqrt{n}}+k_n^{-\frac{r_p}{d_x+1}}+j_n^{-\frac{r_q}{d_x+1}}) \ , \quad (70)$$

where the final result follows by Theorem 4.1 in Chen and Pouzo (2011) and Assumption 3.4(iii). ■

PROOF OF PROPOSITION 3.2: As a result of the proof of Proposition 3.1, it only remains to verify Assumptions 3.4(i)-(iii). Assumption 3.4(i) follows from $(a,\beta)\mapsto\Lambda(a,\beta)$ being twice continuously differentiable in (a,β). Let $\beta\in\mathbf{R}^{d_\beta}$ and $\gamma\in\mathbf{R}^{d_\gamma}$. As discussed in Remark 3.4, Assumption 3.4 holds if for all $0\neq(\pi_\beta,\pi_\gamma)\in\mathbf{R}^{d_\beta}\times\mathbf{R}^{d_\gamma}$:

$$\min_{h\in\bar{\mathcal{H}}} E[(\nabla_{\beta'}\Lambda(X_{1i}'\gamma_0+h_0(X_{2i},T_i),\beta_0)\pi_\beta+\lambda(X_{1i}'\gamma_0+h_0(X_{2i},T_i),\beta_0)(X_{1i}'\pi_\gamma-h(X_{2i},T_i)))^2]>0 \ , \quad (71)$$

where the minimum in (71) is indeed attained by Remark 3.5. Suppose by way of contradiction that (71) does not hold. Then there exist (π_β,π_γ,h) with $(\pi_\beta,\pi_\gamma)\neq 0$ and $h\in\bar{\mathcal{H}}$ such that with probability one:

$$\nabla_{\beta'}\Lambda(X_{1i}'\gamma_0+h_0(X_{2i},T_i),\beta_0)\pi_\beta+\lambda(X_{1i}'\gamma_0+h_0(X_{2i},T_i),\beta_0)X_{1i}'\pi_\gamma=\lambda(X_{1i}'\gamma_0+h_0(X_{2i},T_i),\beta_0)h(X_{2i},T_i) \ . \quad (72)$$

We first establish that (72) implies $\pi_\beta=0$. Suppose instead that $\pi_\beta\neq 0$, and define the function $a\mapsto G(a|\pi_\beta)$ by:

$$G(a|\pi_\beta)\equiv\frac{\nabla_{\beta'}\Lambda(a,\beta_0)\pi_\beta}{\lambda(a,\beta_0)} \ . \quad (73)$$

Next, notice that since $\lambda(X_{1i}'\gamma_0+h_0(X_{2i},T_i),\beta_0)>0$ with probability one by assumption, result (72) implies that:

$$G(X_{1i}'\gamma_0+h_0(X_{2i},T_i)|\pi_\beta)=h(X_{2i},T_i)-X_{1i}'\pi_\gamma \ , \quad (74)$$

almost surely. Furthermore differentiating both sides of (74) with respect to x_1 yields also with probability one:

$$G'(X_{1i}'\gamma_0+h_0(X_{2i},T_i)|\pi_\beta)\gamma_0=\pi_\gamma \ , \quad (75)$$

where $G'(a|\pi_\beta)=\frac{d}{da}G(a|\pi_\beta)$. Hence, since by assumption $\gamma_0\neq 0$ we conclude from (75) that $a\mapsto G(a|\pi_\beta)$ is linear for all a in the support of $X_{1i}'\gamma_0+h_0(X_{2i},T_i)$. Therefore, by definition of $G(a|\pi_\beta)$ and $\tilde{Y}_i$, we must have:

$$\nabla_{\beta'}\Lambda(\tilde{Y}_i,\beta_0)\pi_\beta=\lambda(\tilde{Y}_i,\beta_0)(c_0+c_1\tilde{Y}_i) \quad (76)$$

with probability one for some $c_0,c_1\in\mathbf{R}$. Result (76), however, contradicts that $E[Z_iZ_i']$ has full rank. We therefore conclude that $\pi_\beta=0$. By (72), if $\pi_\beta=0$, then $\lambda(X_{1i}'\gamma_0+h_0(X_{2i},T_i),\beta_0)>0$ almost surely implies:

$$X_{1i}'\pi_\gamma=h(X_{2i},T_i) \quad (77)$$

with probability one. However, if $\pi_\gamma\neq 0$, then (77) violates the assumption that (X_{1i},X_{2i}) is continuously distributed. Thus, we conclude that if (π_β,π_γ) satisfy (72) then $(\pi_\beta,\pi_\gamma)=0$, verifying Assumption 3.4(ii). Finally, Assumption 3.4(iii) holds with $r_q/(d_x+1)=\kappa/(d_{x_2}+1)$, which together with Theorem 3.1 establishes the Proposition. ■

Lemma A.4. *Under Assumptions 2.1(i), 2.1(iii)-(iv), 3.1(ii)-(iii), 3.3(i) and 3.4(i), it follows that:*

$$E[(1\{Y_i\leq m(X_i,T_i,\alpha_1)\}-1\{Y_i\leq m(X_i,T_i,\alpha_2)\})^2|X_i,T_i]\lesssim|m(X_i,T_i,\alpha_1)-m(X_i,T_i,\alpha_2)| \ .$$

PROOF: For all y such that $m(x,0,\alpha_0) \leq y \leq m(x,1,\alpha_0)$ let $v(y,x,\alpha_0)$ denote the inverse in u of $m(x,u,\alpha_0)$, which exists by the strict monotonicity from Assumptions 2.1(iii)-(iv). Also define an extension to $\mathbf{R}$ by:

$$v^e(y,x,\alpha_0) = \begin{cases} 0 & \text{if } y \leq m(x,0,\alpha_0) \\ v(y,x,\alpha_0) & \text{if } m(x,0,\alpha_0) \leq y \leq m(x,1,\alpha_0) \\ 1 & \text{if } m(x,1,\alpha_0) \leq y \end{cases} \tag{78}$$

Expanding the square, Assumption 2.1(i), definition (78) and using $U_i \sim U[0,1]$, we then obtain that:

$$E[(1\{Y_i \leq m(X_i,T_i,\alpha_1)\} - 1\{Y_i \leq m(X_i,T_i,\alpha_2)\})^2 | X_i, T_i] = |v^e(m(X_i,T_i,\alpha_1),X_i,\alpha_0) - v^e(m(X_i,T_i,\alpha_2),X_i,\alpha_0)| \tag{79}$$

Recall that if a function f is monotonic and differentiable, then f^{-1} is also monotonic and differentiable with $\frac{df^{-1}(f(x))}{dx} = (\frac{df(x)}{dx})^{-1}$. The mean value theorem therefore implies that if $y, y' \in [m(x,0,\alpha_0), m(x,1,\alpha_1)]$, then:

$$|v(y,x,\alpha_0) - v(y',x,\alpha_0)| = |\frac{dv(\bar{y},x,\alpha_0)}{dy}| \times |y - y'| \leq \sup_{(x,u)\in\mathcal{X}\times[0,1]} [\frac{dm(x,u,\alpha_0)}{du}]^{-1} \times |y - y'| \ , \tag{80}$$

where $\bar{y}$ is a convex combination of y and y'. However, from continuity and compactness we also have:

$$\inf_{(x,u)\in\mathcal{X}\times[0,1]} \frac{dm(x,u,\alpha_0)}{du} > 0 \ , \tag{81}$$

since the infimum must be attained. Therefore, employing results (78), (80) and (81) we obtain that:

$$\sup_{(y,y')\in\mathbf{R}^2} |v^e(y,x,\alpha_0) - v^e(y',x,\alpha_0)| \leq \sup_{(y,y')\in[m(x,0,\alpha_0),m(x,1,\alpha_0)]^2} |v(y,x,\alpha_0) - v(y',x,\alpha_0)| \lesssim |y - y'| \ . \tag{82}$$

The conclusion of the Lemma then follows from (79) and (82). ■

Lemma A.5. *Let $g_\lambda(x,t,\alpha_0) = f_{Y|X}(m(x,t,\alpha_0)|x)\frac{dm(x,t,\alpha_0)}{d\alpha}[v^\lambda]$ and define $\mathcal{F}_\lambda \equiv \{g_\lambda(x,t,\alpha_0)P(Y \leq m(x,t,\alpha)|x,t) : \alpha \in \mathcal{A}\}$. If Assumptions 3.1, 3.3(i)-(ii), 3.4(i), 3.4(iv) and 4.2(ii) hold, then $\mathcal{F}_\lambda$ is Donsker for every $0 \neq \lambda \in \mathbf{R}^{d_\theta}$.*

PROOF: Recall that as argued in (58), $\frac{dm(x,t,\alpha)}{dh}$ and $\|\frac{dm(x,t,\alpha)}{d\theta}\|$ are uniformly bounded in $(x,t,\alpha) \in \mathcal{X}\times[0,1]\times\mathcal{A}$ by Assumptions 3.1(iii), 3.3(i) and 3.4(i). Similarly note that $f_{Y|X}$ and $g_\lambda(\cdot,\cdot,\alpha_0)$ are bounded by Assumption 3.4(iv) and 4.2(ii). Therefore, for any $\alpha_1, \alpha_2 \in \mathcal{A}$, we obtain by successive applications of the mean value theorem that:

$$\begin{aligned} |g_\lambda(x,t,\alpha_0)(P(Y \leq m(x,t,\alpha_1)|x,t) - P(Y \leq m(x,t,\alpha_2)|x,t))| &\lesssim |m(x,t,\alpha_1) - m(x,t,\alpha_2)| \\ &= |\frac{dm(x,t,\bar{\alpha})}{d\theta}[\theta_1 - \theta_2] + \frac{dm(x,t,\bar{\alpha})}{dh}[h_1 - h_2]| \lesssim \|\alpha_1 - \alpha_2\|_{c,0} \ , \end{aligned} \tag{83}$$

where $\bar{\alpha}$ is a convex combination of α_1 and α_2. Hence, $\mathcal{F}_\lambda$ is Lipschitz in $\alpha \in \mathcal{A}$ under $\|\cdot\|_{c,0}$, and since in addition $\mathcal{F}_\lambda$ is uniformly bounded we obtain by Theorem 2.7.11 in van der Vaart and Wellner (1996) that:

$$\begin{aligned} \int_0^\infty \sqrt{\log N_{[\,]}(\epsilon,\mathcal{F}_\lambda,\|\cdot\|_\infty)}d\epsilon &\lesssim \int_0^\infty \sqrt{\log N_{[\,]}(\epsilon,\mathcal{A},\|\cdot\|_{c,0})}d\epsilon \\ &= \int_0^D \sqrt{\log N_{[\,]}(\epsilon,\Theta,\|\cdot\|) + \log N_{[\,]}(\epsilon,\mathcal{H},\|\cdot\|_\infty)}d\epsilon \lesssim \int_0^D \left(\frac{1}{\epsilon}\right)^{\frac{d_x+1}{2\kappa}} d\epsilon < \infty \ , \end{aligned} \tag{84}$$

where the equality holds for D the diameter of $\mathcal{A}$ under $\|\cdot\|_{c,0}$, and the second and third inequalities are implied by Assumption 3.3(i), Theorem 2.7.1 in van der Vaart and Wellner (1996) and $\kappa > \frac{d_x+1}{2}$. The claim of the Lemma then follows by (84) and Theorem 2.5.6 in van der Vaart and Wellner (1996). ■

PROOF OF COROLLARY 4.1: To establish the corollary notice that the definition of v^λ and result (59) imply:

$$\begin{aligned} F_\lambda(\alpha - \alpha_0) &= (\theta - \theta_0)'E[R_{h^*}(X_i,T_i)R'_{h^*}(X_i,T_i)](E[R_{h^*}(X_i,T_i)R'_{h^*}(X_i,T_i)])^{-1}\lambda \\ &= E[f^2_{Y|X}(m(X_i,T_i,\alpha_0)|X_i)(\frac{dm(X_i,T_i,\alpha_0)}{d\theta}[\theta - \theta_0] - \frac{dm(X_i,T_i,\alpha_0)}{dh}[h^{*\prime}(\theta - \theta_0)])(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[v^\lambda])] \\ &= E[f^2_{Y|X}(m(X_i,T_i,\alpha_0)|X_i)(\frac{dm(X_i,T_i,\alpha_0)}{d\theta}[\theta - \theta_0] + \frac{dm(X_i,T_i,\alpha_0)}{dh}[h - h_0])(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[v^\lambda])] \ , \end{aligned}$$

which establishes the claim of the corollary. ■

PROOF OF THEOREM 4.1: The proof proceeds by appealing to Theorem 3.1 in Chen and Pouzo (2009). We note, however, that several of their assumptions are employed to obtain a rate of convergence for the penalized minimum distance estimator. In lieu of Theorem 3.1, we therefore do not need to verify all their assumptions, but only those pertinent to asymptotic normality, which are 2.5-2.10 and 3.1-3.6.

First notice that their Assumptions 2.5(i)-(iii) are automatically satisfied with $\hat{\Sigma}(x,t) = \Sigma(x,t) = 1$, and $\delta_{\Sigma,n} = 0$. Their Assumptions 2.6(i)-(ii) hold with $\delta_{m,n}^2 = k_n/n + k_n^{-\frac{2r_p}{d_x+1}}$ by Lemma A.1(i)-(ii) and result (46). In turn, their Assumption 2.7(i)-(iii) is verified by our Assumptions 3.1(ii)-(iii), 3.2(i) and 3.2(iii). Their Assumption 2.8(i) holds with $K = 1$, while Assumption 2.8(ii) holds with $b_{m,J_n} = k_n^{-\frac{r_p}{d_x+1}}$ by our Assumption 3.2(ii). Additionally, their Assumption 2.9(i) holds by our Assumptions 3.3(ii) and 3.4(i), their Assumption 2.9(ii) holds with $\|\cdot\| = \|\cdot\|_{c,L^2}$ by Lemma A.3, and their Assumption 2.9(iii) is satisfied for $\|\cdot\|_s = \|\cdot\|_{c,0}$ due to the densities of X_i and T_i being bounded by Assumption 3.1(ii)-(iii). Moreover, their Assumption 2.10(i) holds since $\Sigma(X_i,T_i) = 1$ a.s. and $\|\frac{dm(x,t,\alpha_0)}{d\alpha}\|$ is bounded uniformly in $(x,t) \in \mathcal{X} \times [0,1]$ by Assumption 3.1(iii), 3.4(i) and h_0 being continuous. In turn, their Assumption 2.10(ii) holds by our Assumption 3.4(ii).

As argued in (58), $\frac{dm(x,t,\alpha)}{dh}$ and $\|\frac{dm(x,t,\alpha)}{d\theta}\|$ are uniformly bounded in $(x,t,\alpha) \in \mathcal{X} \times [0,1] \times \mathcal{A}$ by Assumptions 3.1(iii), 3.3(i) and 3.4(i). Therefore, by Lemma A.4 and the mean value theorem we obtain:

$$E[(1\{Y_i \leq m(X_i,T_i,\alpha_1)\} - 1\{Y_i \leq m(X_i,T_i,\alpha_2)\})^2|X_i,T_i] \lesssim |m(X_i,T_i,\alpha_1) - m(X_i,T_i,\alpha_2)| \lesssim \|\alpha_1 - \alpha_2\|_{c,0} \ , \quad (85)$$

which verifies Assumption 3.1(i) of Chen and Pouzo (2009) with $r = 2$. In turn, their requirement 3.1(ii) is satisfied with $C(Y_i,X_i,T_i) = 1$ almost surely. Next, observe that as noted in Newey (1997), under Assumption 4.2(iii) we may assume without loss of generality that $E[q^{j_n}(X_i,T_i)q^{j_n'}(X_i,T_i)] = I$, and hence note that as in (45), $\|h\|_\infty/\|h\|_{L^2} \leq \xi_{q,n}$ for all $h \in \bar{\mathcal{H}}$. Therefore, letting $\Pi_n\alpha = (\theta, \Pi_n h)$ for any $\alpha = (\theta,h)$ we obtain from Assumption 3.4(iii) that:

$$\begin{aligned}\|\hat{\alpha} - \alpha_0\|_{c,0} &\leq \|\hat{\alpha} - \Pi_n\alpha_0\|_{c,0} + O(j_n^{-\frac{r_q}{d_x+1}}) \\ &\leq \|\hat{\theta} - \theta_0\| + \{\sup_{h \in \bar{\mathcal{H}}_n} \frac{\|h\|_\infty}{\|h\|_{L^2}}\} \times \|\hat{h} - \Pi_n h_0\|_{L^2} + O(j_n^{-\frac{r_q}{d_x+1}}) \lesssim \xi_{q,n} \times \|\hat{\alpha} - \alpha_0\|_{c,L^2} + O(j_n^{-\frac{r_q}{d_x+1}}) \ . \quad (86)\end{aligned}$$

Therefore, Theorem 3.1 and Assumption 4.2(iv) verifies Assumption 3.1(iii) in Chen and Pouzo (2009) with $\delta_n = \sqrt{k_n}/\sqrt{n} + k_n^{-\frac{r_p}{d_x+1}} + j_n^{-\frac{r_q}{d_x+1}}$ and $\delta_{s,n} = \xi_{q,n} \times \delta_n$. Additionally, notice that then by Assumptions 4.1(ii) and 4.2(iv), $\delta_n = o(n^{-\frac{1}{4}})$ and $\|\Pi_n v^\lambda - v^\lambda\|_{c,L^2} \times \delta_n = o(n^{-\frac{1}{2}})$ verifying their conditions 3.2(iii) and 3.3(i). In turn, their Assumptions 3.2(i) and 3.2(ii) hold by our Assumption 4.2(i) with $\Sigma_0(X_i,T_i) = T_i(1-T_i)$, while their conditions 3.3(ii)-(iii) hold trivially since $\delta_{\Sigma_n} = \lambda_n = 0$. Also, note their Assumption 3.5 is verified by Lemma A.5.

For any $r : \mathcal{X} \times [0,1] \to \mathbf{R}$, let $\bar{E}[r(X_i,T_i)|X_i,T_i] \equiv p^{k_n'}(X_i,T_i)(P'P)^{-1}P'R$, with $R \equiv (r(X_1,T_1),\ldots,r(X_n,T_n))$. Defining $g_\lambda(x,t,\alpha_0) = f_{Y|X}(m(x,t,\alpha_0)|x)\frac{dm(x,t,\alpha_0)}{d\alpha}[v^\lambda]$, we then obtain that:

$$E[(g_\lambda(X_i,T_i,\alpha) - \bar{E}[g_\lambda(X_i,T_i,\alpha_0)|X_i,T_i])^2] = O_p(k_n^{-\frac{2r_p}{d_x+1}}) \ , \quad (87)$$

by Assumptions 3.1(i)-(iii), 3.2(i), 4.1(ii) and Lemma A.1(B) in Ai and Chen (2003). Hence, Assumptions 3.4(i)-(ii) in Chen and Pouzo (2009) is satisfied due to (87), $\Sigma(X_i,T_i) = 1$, $k_n^{-\frac{2r_p}{d_x+1}} = o(n^{-\frac{1}{2}})$ by Assumption 4.2(iv) and $\delta_n^2 = o(n^{-\frac{1}{2}})$. To verify their Assumption 3.6(i), notice that since $h \in \mathcal{H}$ are uniformly bounded and $\mathcal{X}, \Theta$ are compact, $\sup_{(h,\theta)\in\mathcal{H}\times\Theta} \|(X_i,h(X_i,T_i),\theta)\|$ is contained in a compact set almost surely. Therefore, since the first two derivatives of $m(x,e,\theta)$ in (e,θ) are continuous, they are uniformly bounded in $(x,t,\alpha) \in \mathcal{X} \times [0,1] \times \mathcal{A}$. Similarly, since Assumption 4.1(ii) and 4.2(ii) imply $\|v^\lambda\|_{c,0} < \infty$ and $\|v_n^\lambda\|_{c,0}$ is uniformly bounded, we conclude:

$$\sup_{\alpha\in\mathcal{A}} |\frac{d^2}{d\alpha d\alpha} P(Y_i \leq m(X_i,T_i,\alpha)|X_i,T_i)[v_n^\lambda, v_n^\lambda]| \leq M \ , \quad (88)$$

almost surely for some constant M by Assumptions 3.1(iii) and 4.1(i). Result (88) verifies their condition 3.6(i).

Next, let $\mathcal{N}_{0n} \equiv \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_0\|_{c,0} \leq \delta_{s,n}\}$ and employ $\|v_n^\lambda\|_{c,0}$ uniformly bounded, $f_{Y|X}$ and $f'_{Y|X}$ bounded by Assumptions 3.1(iii) and 4.1(i), the mean value theorem and results (64), (65) and (83) to obtain:

$$\begin{aligned} E[\sup_{\alpha\in\mathcal{N}_{0n}} (f_{Y|X}(m(X_i,T_i,\alpha)|X_i)\frac{dm(X_i,T_i,\alpha)}{d\alpha}[v_n^\lambda] - f_{Y|X}(m(X_i,T_i,\alpha_0)|X_i)\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[v_n^\lambda])^2] \\ \lesssim E[\sup_{\alpha\in\mathcal{N}_{0n}} (m(X_i,T_i,\alpha) - m(X_i,T_i,\alpha_0))^2] + E[\sup_{\alpha\in\mathcal{N}_{0n}} (\frac{dm(X_i,T_i,\alpha)}{d\alpha}[v_n^\lambda] - \frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[v_n^\lambda])^2] \\ \lesssim \sup_{\alpha\in\mathcal{N}_{0n}} \|\alpha - \alpha_0\|_{c,0}^2 \ . \end{aligned} \tag{89}$$

Hence, since $\delta_{s,n} = o(n^{-\frac{1}{4}})$ by result (86) and Assumption 4.2(iv), (89) verifies Assumption 3.6(ii) in Chen and Pouzo (2009). Similarly, let $\mathcal{N}_0 \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_{c,0} \leq \delta_{s,n}\}$ and use the Cauchy-Schwarz inequality, the mean value theorem, Assumption 4.1(i), together with result (53), Lemmas A.3 and A.2(ii) and $\|\cdot\|_{c,L^2} \lesssim \|\cdot\|_{c,0}$ to obtain:

$$\begin{aligned} \sup_{(\alpha_1,\alpha_2)\in\mathcal{N}_0\times\mathcal{N}_{0n}} E[|(f_{Y|X}(m(X_i,T_i,\alpha_0)|X_i) - f_{Y|X}(m(X_i,T_i,\alpha_1)|X_i))\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha_2 - \alpha_0]|] \\ \lesssim \sup_{\alpha_1\in\mathcal{N}_0} (E[(m(X_i,T_i,\alpha_0) - m(X_i,T_i,\alpha_1))^2])^{\frac{1}{2}} \times \sup_{\alpha_2\in\mathcal{N}_{0n}} (E[(\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha_2 - \alpha_0])^2])^{\frac{1}{2}} = O(\delta_{s,n}^2) \ . \end{aligned} \tag{90}$$

Since $f_{Y|X}$ is bounded by Assumption 3.3(iv), (64), (65) and the triangle and Cauchy-Schwarz inequalities yield:

$$\sup_{(\alpha_1,\alpha_2)\in\mathcal{N}_0\times\mathcal{N}_{0n}} E[|f_{Y|X}(m(X_i,T_i,\alpha_1)|X_i)(\frac{dm(X_i,T_i,\alpha_1)}{d\alpha}[\alpha_2 - \alpha_0] - \frac{dm(X_i,T_i,\alpha_2)}{d\alpha}[\alpha_2 - \alpha_0])|] = O(\delta_{s,n}^2) \ . \tag{91}$$

Therefore, since $g_\lambda(x,t,\alpha_0)$ is bounded in $(x,t) \in \mathcal{X}\times[0,1]$, we conclude from results (90) and (91) that:

$$\begin{aligned} \sup_{(\alpha_1,\alpha_2)\in\mathcal{N}_0\times\mathcal{N}_{0n}} |E[g_\lambda(X_i,T_i,\alpha_0)(f_{Y|X}(m(X_i,T_i,\alpha_1)|X_i)\frac{dm(X_i,T_i,\alpha_1)}{d\alpha}[\alpha_2 - \alpha_0] \\ - f_{Y|X}(m(X_i,T_i,\alpha_0)|X_i)\frac{dm(X_i,T_i,\alpha_0)}{d\alpha}[\alpha_2 - \alpha_0])]| = O(\delta_{s,n}^2) \ . \end{aligned} \tag{92}$$

Hence, $\Sigma(X_i,T_i) = 1$ almost surely, $\delta_{s,n}^2 = o(n^{-\frac{1}{2}})$ by Assumption 4.2(iv) and (92) verify Assumption 3.6(iii) of Chen and Pouzo (2009). The claim of the Theorem then follows by Theorem 3.1 in Chen and Pouzo (2009). ■

Proof of Lemma 4.1: Let P_θ denote a probability under the model $Y_i = m(X_i, h_0(X_i,U_i),\theta)$ and similarly, $f_{YX}(y,x|\theta)$ denote a joint pdf under θ. First observe that due to strict monotonicity and $U_i \perp X_i$ we obtain:

$$\begin{aligned} P_\theta(Y_i \leq y|X_i) = P_\theta(m(X_i,h_0(X_i,U_i),\theta) \leq y|X_i) = P_\theta(U_i \leq h_0^{-1}(X_i,m^{-1}(X_i,y,\theta))|X_i) \\ = 1\{h_0^{-1}(X_i,m^{-1}(X_i,y,\theta)) \geq 0\} \times \min\{1, h_0^{-1}(X_i,m^{-1}(X_i,y,\theta))\} \ , \end{aligned} \tag{93}$$

where $h_0^{-1}(x,\cdot)$ and $m^{-1}(x,\cdot,\theta)$ are the inverses of $h_0(x,\cdot)$ and $m(x,\cdot,\theta)$ respectively. Hence, by result (93):

$$f_{YX}(y,x|\theta) = 1\{m(x,h_0(x,0),\theta) \leq y \leq m(x,h_0(x,1),\theta)\} \times [\frac{d}{dy}h_0^{-1}(x,m^{-1}(x,y,\theta))] \times f_X(x) \ . \tag{94}$$

Recall that for any function g such that $\frac{d}{dx}g(x) \neq 0$, we have $\frac{dg^{-1}(g(x))}{dx} = (\frac{d}{dx}g(x))^{-1}$. Since by continuity and compactness both $\frac{d}{du}h_0(x,u)$ and $\frac{d}{de}m(x,e,\theta)\Big|_{e=h_0(x,u)}$ are uniformly bounded in (x,u,θ), it follows that:

$$\frac{d}{dy}h_0^{-1}(x,m^{-1}(x,y,\theta)) = \frac{d}{du}h_0^{-1}(x,u)\Big|_{u=m^{-1}(x,y,\theta)} \times \frac{d}{dy}m^{-1}(x,y,\theta) \geq \epsilon \ , \tag{95}$$

for some $\epsilon > 0$. Next, select $\lambda \in \mathbf{R}^{d_\theta}$ so that (31) holds, and assume without loss of generality that in fact:

$$P(\frac{dm(X_i,h_0(X_i,1),\theta_0)}{d\theta}[\lambda] \neq 0) > 0 \ . \tag{96}$$

Next observe that since $\theta \mapsto m(x,h_0(x,1),\theta)$ is continuously differentiable, we have for $|\eta|$ sufficiently small:

$$m(x,h_0(x,1),\theta_0 + \lambda\eta) = m(x,h_0(x,1),\theta_0) + \frac{dm(x,h_0(x,1),\bar\theta(x))}{d\theta}[\lambda\eta] \ , \tag{97}$$

for some $\bar{\theta}(x) \in \Theta$ a convex combination of θ_0 and $\theta_0 + \lambda\eta$ by Assumptions 3.3(ii) and 4.2(i). Define the sets:

$$A_+(\eta) \equiv \{x \in \mathcal{X} : \frac{dm(x, h_0(x,1), \bar{\theta}(x))}{d\theta}[\lambda\eta] > 0\} \qquad A_-(\eta) \equiv \{x \in \mathcal{X} : \frac{dm(x, h_0(x,1), \bar{\theta}(x))}{d\theta}[\lambda\eta] < 0\} \ . \tag{98}$$

Since X_i is continuously distributed, continuity of $\theta \mapsto \frac{dm(x,h_0(x,1),\theta)}{d\theta}$ and (96) imply there exists a set A with positive Lebesgue measure, such that for $\eta > 0$ sufficiently small either $A \subset A_+(\eta)$ or $A \subset A_-(\eta)$. Without loss of generality we assume $A \subset A_+(\eta)$. Since f_X is bounded away from zero, $A \subset \mathbf{R}^{d_x}$ and results (94) and (95) in turn imply:

$$\begin{aligned}
&\int_{\mathbf{R}^{d_x}} \int_{\mathbf{R}} (\sqrt{f_{YX}(y,x|\theta_0)} - \sqrt{f_{YX}(y,x|\theta_0+\lambda\eta)})^2 dydx \\
&\quad \gtrsim \int_A \int_{\mathbf{R}} 1\{m(x,h_0(x,1),\theta_0) + \frac{dm(x,h_0(x,1),\bar{\theta}(x))}{d\theta}[\lambda\eta] \geq y \geq m(x,h_0(x,1),\theta_0)\} dydx \qquad (99) \\
&\quad = \eta \int_A \frac{dm(x,h_0(x,1),\theta_0)}{d\theta}[\lambda] + o(\eta) \ , \qquad (100)
\end{aligned}$$

where the final equality follows from continuity of $(x,u,\theta), \mapsto \frac{dm(x,h_0(x,u),\theta)}{d\theta}$ and boundedness of $A \subset \mathcal{X}$. Hence,

$$\lim_{\eta \downarrow 0} \eta^{-2} \int_{\mathbf{R}^{d_x+1}} \left(\sqrt{f_{YX}(y,x|\theta_0)} - \sqrt{f_{YX}(y,x|\theta_0+\lambda\eta)}\right)^2 = \infty \ , \tag{101}$$

which in turn implies the model is not quadratic mean differentiable at θ_0. ■

Proof of Proposition 4.1: First note Assumption 4.1(i)-(ii) are satisfied with $r_q = \kappa$ since our assumptions guarantee $\|f_{Y|X}\|_{\infty,r_p} < \infty$, $\|v_h^\lambda\|_{\infty,\kappa} < \infty$ and $\|f_{Y|X}(m(\cdot,\cdot,\alpha_0)|\cdot)\frac{dm(\cdot,\cdot,\alpha_0)}{d\alpha}[v^\lambda]\|_{\infty,r_p} < \infty$ for all $\lambda \in \mathbf{R}^{d_\theta}$. In turn, Assumption 4.2(iv) hold for the $k_n \asymp n^\zeta$ and the claim follows by Theorem 4.1. For the second claim note:

$$\frac{d}{d\gamma}\Lambda(X_{1i}'\gamma_0 + h_0(X_{2i},U_i),\beta_0) = \lambda(X_{1i}'\gamma_0 + h_0(X_{2i},U_i),\beta_0) \times X_{1i} \ , \tag{102}$$

which establishes (31) holds due to X_{1i} being continuously distributed. The result then follows by Lemma 4.1. ■

Proof of Theorem 4.2: Follows immediately from Theorem 4.1 in Chen and Pouzo (2009) and the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ being continuous. ■

References

Adams, R. A. and Fournier, J. J. (2003). *Sobolev Spaces*. Academic Press, Elsevier, The Netherlands.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** 1795–1844.

Ai, C. and Chen, X. (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, **141** 5–43.

Altonji, J. G. and Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, **73** 1053–1102.

Bajari, P. and Benkard, C. L. (2005). Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of Political Economy*, **113** 1239–1276.

Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap*. Springer- Verlag, New York.

Benkard, L. and Berry, S. (2006). On the nonparametric identification of nonlinear simultaneous equations models: Comment on brown (1983) and roehrig (1988). *Econometrica*, **74** 1429–1440.

Berry, S. T., Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, **63** 841–890.

Bickel, P. J., Klassen, C. A., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

Blundell, R. and Powell, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics*, vol. II. Cambridge University Press, 312–357.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, **26** 211–252.

Brown, B. W. (1983). The identification problem in systems nonlinear in the variables. *Econometrica*, **51** 175–196.

Brown, D. J. and Wegkamp, M. H. (2002). Weighted minimum mean-square distance from independence estimation. *Econometrica*, **70** 2035–2051.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, **60** 567–596.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics 6B* (J. J. Heckman and E. E. Leamer, eds.). North Holland, Elsevier.

Chen, X., Chernozhukov, V., Lee, S. and Newey, W. K. (2011). Local identification of nonparametric and semiparametric models. Working paper, Yale University.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, **152** 46–60.

Chen, X. and Pouzo, D. (2011). Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica* forthcoming.

Chernozhukov, V., Fernandez-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, **78** 1093–1125.

Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, **73** 245–261.

Chernozhukov, V., Imbens, G. W. and Newey, W. K. (2007). Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, **139** 4–14.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, **71** 1405–1441.

Heckman, J. J., Matzkin, R. L. and Nesheim, L. (2010). Nonparametric identification and estimation of nonadditive hedonic models. *Econometrica*, **78** 1569–1591.

Hoderlein, S. and Mammen, E. (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, **75** 1513–1518.

Hoderlein, S. and White, H. (2009). Nonparametric identification in nonseparable panel data models with generalized fixed effects. Working paper, University of California San Diego.

Huang, J. Z. (1998). Projection estimation in multiple regression with applications to functional anova models. *Annals of Statistics*, **26** 242–272.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, **31** 1600–1635.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equation models without additivity. *Econometrica*, **77** 1481–1512.

Komunjer, I. and Santos, A. (2010). Semiparametric estimation of nonseparable models: A mdi approach. *Econometrics Journal*, **13** S28–S55.

Linton, O., Sperlich, S. and VanKeilegom, I. (2008). Estimation of a semiparametric transformation model. *The Annals of Statistics*, **36** 686–718.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. Wiley, New York.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, **96** 190–217.

Manski, C. F. (1983). Closest empirical distribution estimation. *Econometrica*, **51** 305–319.

Matzkin, R. L. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica*, **71** 1339–1375.

Matzkin, R. L. (2008). Identification in nonparametric simultaneous equations. *Econometrica*, **76** 945–978.

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

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

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56** 931–954.

Roehrig, C. (1988). Conditions for identification in nonparametric and parametric models. *Econometrica*, **56** 433–447.

Torgovitsky, A. (2011). Identification of nonseparable models with general instruments. Working paper, Yale University.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer, New York.