SENSITIVITY ANALYSIS IN SEMIPARAMETRIC LIKELIHOOD MODELS

Xiaohong Chen, Elie Tamer, and Alexander Torgovitsky

November 2011

Sensitivity Analysis in Semiparametric Likelihood Models*

Xiaohong Chen† Yale
Elie Tamer‡ Northwestern
Alexander Torgovitsky§ Northwestern

First draft: December 2007; This draft: November 13, 2011

Abstract

We provide methods for inference on a finite dimensional parameter of interest, $\theta \in \Re^{d_\theta}$, in a semiparametric probability model when an infinite dimensional nuisance parameter, g, is present. We depart from the semiparametric literature in that we do not require that the pair (θ, g) is point identified and so we construct confidence regions for θ that are robust to non-point identification. This allows practitioners to examine the sensitivity of their estimates of θ to specification of g in a likelihood setup. To construct these confidence regions for θ, we invert a profiled sieve likelihood ratio (LR) statistic. We derive the asymptotic null distribution of this profiled sieve LR, which is nonstandard when θ is not point identified (but is χ^2 distributed under point identification). We show that a simple weighted bootstrap procedure consistently estimates this complicated distribution's quantiles. Monte Carlo studies of a semiparametric dynamic binary response panel data model indicate that our weighted bootstrap procedures performs adequately in finite samples. We provide three empirical illustrations where we compare our results to the ones obtained using standard (less robust) methods.

Keywords: Sensitivity Analysis, Semiparametric Models, Partial Identification, Irregular Functionals, Sieve Likelihood Ratio, Weighted Bootstrap

*This paper benefits from the collaborations with Z. Liao and D. Pouzo on some related research. We acknowledge useful comments from D. Andrews, X. Cheng, O. Linton, D. Pouzo, A. Santos, X. Shi, and other participants at the 2010 North American Winter Meeting of the Econometric Society in Atlanta and many other conferences and departmental seminars.

†Department of Economics, Yale University, xiaohong.chen@yale.edu, ph: 203 432 5852. Support from The National Science Foundation is gratefully acknowledged.

‡Department of Economics, Northwestern University, tamer@northwestern.edu, ph: 847 491 8218. Support from The National Science Foundation is gratefully acknowledged.

§Department of Economics, Northwestern University, a-torgovitsky@northwestern.edu.

1 Introduction

We consider inference on a finite dimensional parameter of interest θ in semiparametric likelihood models when an infinite dimensional nuisance parameter g is present. Existing semiparametric methods to estimate θ in the presence of g, such as sieve maximum likelihood (ML), penalized ML or locally polynomial likelihood (e.g., Fan and Gijbels (1996)), have become increasingly popular in applied econometrics. However, the existing work so far on the asymptotic properties of these procedures rely on a key assumption that the model is (*globally*) *point identified,* i.e., $P_0 = P(.;\theta;g) = P(.;\theta',g')$ means that $(\theta,g) = (\theta',g')$ where P_0 is the true probability distribution of the data, and $P(.;\theta;g)$ is the model probability distribution indexed by parameters (θ,g). The objective of this paper is to construct confidence regions for θ allowing for violations of this assumption, i.e, for the case when point identification does not hold: $P_0 = P(.;\theta;g) = P(.;\theta',g')$ but $(\theta,g) \neq (\theta',g')$. Although this paper focuses on likelihood based models, our approach can be extended to other contexts such as semi/nonparametric moment conditions based models.[1]

There are at least three reasons that motivate our semiparametric likelihood based approach. The *First motivation* is sensitivity analysis. Empirical economists use parametric likelihood methods routinely to do inference on some finite dimensional parameter of interest (θ) in the presence of nuisance parameters g. Maximum likelihood is attractive, since under point identification (and standard regularity conditions) it is efficient. The usual approach is to use parametric assumptions on g and then to show that θ is point identified, and hence standard likelihood based inference methods apply. However, typically, the assumptions made on g, such as functional forms or distributional assumptions, are not plausible and are usually not derived from an economic model; rather they are used because of some computational advantage or based on familiarity. Naturally then, one is worried whether inferences using these parametric models are *sensitive* to specification of g. The *second motivation* for our approach is that the starting point of almost all standard semiparametric models in which both θ and g are treated as unknown parameters, is the point identification conditions where θ is assumed to be globally point identified. These assumptions are not easy to verify outside of simple models, and when available, these point identification conditions might be difficult to satisfy in standard data sets. In addition, available statistical methods for inference in these semiparametric models are invalid when point identification fails. Finally, in models where the parameter of interest is not point identified, an important issue becomes one of finding the tightest set of observationally equivalent parameters. In

[1]See, e.g., Chen, Pouzo, and Tamer (2011) for inference on nonparametric conditional moment models under partial identification using a sieve quasi likelihood ratio statistic.

some cases, showing that one's approach delivers estimates of this sharp set is not easy. So, a *third motivation* for our approach is that a likelihood based model delivers the sharp set which is by definition the argmax of the likelihood function of the observed data.

Current empirical approaches to address the issue of how robust is one's estimate of θ to potential misspecification of g in a likelihood model is to estimate θ given g and then examine the change in this value of the θ's estimate as one changes g. So, a "robust" model is then one where estimates of the parameter of interest "do not change much" as one changes g. In this paper we formalize this mostly heuristic exercise and provide valid methods for inference on θ allowing for partial identification as g changes in its logical domain. This function g, which is the object of this *sensitivity analysis*, is chosen as the piece in an empirical model that causes the most unease among economists. These typically are functions or latent distributions where economists have least prior information about their shape, are less likely to be learned even with further data collection (as in equilibrium selection functions) and so are a prime candidate for sensitivity analysis.

To build a confidence region for θ allowing for non-point identification, we exploit the equivalence between testing and confidence region, and construct this confidence region by collecting all the parameter values that we fail to reject using a profiled sieve likelihood ratio statistic (LR). So, our construction is based on the distribution of the profiled sieve LR statistic under non-point identification in which the unknown infinite dimensional nuisance functions are approximated by a sequence of finite dimensional sieves. There is a recent literature on the large sample distribution of a parametric LR statistic when some parameters are not point identified under the null; see for example Liu and Shao (2003). Unfortunately, this Liu and Shao approach is no longer applicable in the presence of infinite dimensional nuisance functions. For example, a direct generalization of Liu and Shao's work would require $\sqrt{n}-$consistent estimation of the nuisance parameter, g, which is not possible in general when g is infinite dimensional and might belong to a non-compact function space (see Section 4 for details). The first main contribution of our paper is to show that a profiled sieve LR statistic, under a set of conditions, admits a tight asymptotic distribution when the likelihood could depend on partially identified infinite dimensional nuisance parameters. This asymptotic distribution holds whether or not the parameter of interest is point identified. For point identified models, our profiled sieve LR statistic converges to the usual χ^2 distribution regardless of whether the parameter of interest is regular (i.e., $\sqrt{n}-$estimable) or irregular (i.e., slower than $\sqrt{n}-$estimable). For partially identified models, our statistic has a complicated limiting distribution, which is a natural extension of that of Liu and Shao (2003) to allow for unknown nuisance parameters belonging to infinite dimensional non-compact function spaces.

The asymptotic null distribution of the profiled sieve LR is difficult to simulate in general partially identified problems. Our next contribution is to show that a simple weighted bootstrap procedure consistently estimates quantiles of the asymptotic null distribution of the profiled sieve LR statistic. This bootstrap procedure appears to behave adequately in small sample numerical simulations. In our Monte Carlo, we simulate a dynamic (short) panel data binary choice model where know that the parameters of interest are not point identified due to the initial conditions problem.

We apply our methods to three empirical examples. In the first empirical example, we consider the duration model with unobserved heterogeneity of Heckman and Singer (1984). Economists typically have no information about the form of the unobserved heterogeneity distribution and so we estimate the structural parameters of interest using NLSY data without making any assumptions on this distribution. In this model, we know that the structural parameter of interest can be point identified at infinity as the durations tend to zero. This sufficient point identification condition does not appear to be valid in our data. Our confidence region construction is robust to failure of point identification. In addition, this model presents a case where even if the parameter is point identified, it can be really difficult to estimate (since rates of convergence are slower than root n). Our confidence regions remain valid in this case when the parameter is irregularly point identified. In the second empirical example, we build a confidence regions for parameters in an intergenerational schooling example as in Plug (2004) in which the dependent variable in a linear model is observed in bins, and so our analysis examines the sensitivity of the estimates to functional form assumptions on the distribution of the errors. The last empirical example estimates a version of the Berry (1992) entry model using airline data by allowing for heteroskedasticity of unknown shape. In all examples, we provide marginal and joint confidence regions on parameters and contrast those to ones obtained from parametric likelihood models.

Literature Review: Our paper contributes to two literatures in econometrics, the partial identification literature and the sieve semiparametric inference literature. There are many recent papers in econometrics that deal with the question of inference in partially identified models (without infinite dimensional nuisance parameters). See Imbens and Manski (2004), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2008), Andrews and Soares (2010)[2]. In addition, Redner (1981) establishes consistency of the parametric MLE without assuming that it is uniquely identified, and Liu and Shao (2003) obtain the

[2]See also the papers of Rosen (2008), Bugni (2010), Canay (2010), Chernozhukov, Lee, and Rosen (2009), Andrews and Shi (2010) and other papers referenced therein. For a recent survey on partial identification in econometrics, see Tamer (2010).

limiting distribution of the parametric LR statistic under non-point identification. Andrews and Cheng (2010) provide methods for constructing confidence regions and tests including parametric likelihood models that are robust to lack of or weak identification. The sieve literature for semiparametric models under point identification is exposited in Chen (2007).[3] In particular, Murphy and Van der Vaart (2000) and Shen and Shi (2005) respectively established that the profiled LR and sieve LR statistic converges to χ^2 distribution for point identified and regular parameters. Chen and Liao (2009) established that the sieve LR statistic converges to a χ^2 distribution for the case where the parameters are point identified and irregular. Chen and Pouzo (2009) considered profiled sieve quasi likelihood ratio inference for point identified semi-nonparametric conditional moment models where the problems could be nonlinear and ill-posed. Finally, Santos (2011), in an interesting paper, considers testing in a nonparametric instrumental variables (IV) regression model without requiring identification. His paper also approximates the unknown functions by sieves but builds confidence regions by inverting a Bierens' type test statistic.

On the other hand, sensitivity analysis has a long history in econometrics. Formally, in some setups, our sensitivity approach is mathematically equivalent to partial identification analysis. See for example Manski (1995). In addition, our approach to sensitivity analysis which constructs confidence regions that reflect model uncertainty in addition to accounting for sampling noise, is similar in spirit to the "extreme bounds approach" as advocated by Leamer (1985). See also Leamer (1987).

The paper is organized as follows. The next section provides some motivating examples. These are likelihood based econometric models in which sensitivity analysis is desirable. These models are mostly structural models that are empirically relevant where a likelihood function is used to conduct policy simulations. Section 3 provides general results for consistency and rates of convergence for a sieve MLE where the parameter belongs to a general function space. We provide the main results of the paper in Sections 4 and 5. There, we have conditions under which a profiled sieve LR statistic admits a tight asymptotic distribution. This limit distribution is difficult to characterize in general. In section 5, we provide a bootstrap procedure that is empirically implementable and show that this procedure is consistent. Section 6 examines the numerical properties of our sieve LR procedure in a limited Monte Carlo experiment where we simulate a binary dynamic panel discrete choice model. We apply our work in Section 7 to three empirical examples. Section 8 concludes with questions for future research.

[3]See also this chapter for background material and other important references for the sieve semiparametrics literature.

2 General Setup and Motivating Examples

This section introduces a family of semiparametric probability models and provides a definition of the identified set for the parameters of interest. Then, we give several motivating examples.

Let $\mathcal{P} \equiv \{p(\cdot;\theta,g) : (\theta,g) \in \Theta \times \mathcal{G}\}$ be a family of probability densities with respect to a dominating sigma finite positive measure μ on a measurable space $(\mathcal{Z},\mathcal{B})$. Let the data $\{Z_i = (Y_i, X_i)\}_{i=1}^n$ be a random sample of $Z = (Y,X)$ that has true (but unknown) probability measure $P_0 \equiv P_Z$ on $(\mathcal{Z},\mathcal{B})$, with $p_0 \equiv \frac{dP_Z}{d\mu}$ being its density wrt the dominating measure μ. We assume that the true probability density p_0 is unique. We say the family of probability models $\mathcal{P} = \{p(\cdot;\theta,g) : (\theta,g) \in \Theta \times \mathcal{G}\}$ is correctly specified if $p_0 \in \mathcal{P}$, that is, $p_0 \in \mathcal{P}_0 \equiv \{p(\cdot;\theta,g) = p_0(\cdot) : (\theta,g) \in \Theta \times \mathcal{G}\} = \{p_0\}$. In this paper we use $E_0()$ to denote expectation under true probability density p_0.[4]

The family of probability models $\mathcal{P}$ is semiparametric in that the parameter of interest θ is finite dimensional and the nuisance parameter g is infinite dimensional. In particular, we assume that Θ is a compact subset in $\Re^{d_\theta}$ and $\mathcal{G}$ is a function space. We assume that the semiparametric probability model $\mathcal{P} = \{p(\cdot;\theta,g) : (\theta,g) \in \Theta \times \mathcal{G}\}$ is correctly specified, i.e., there is at least one $(\theta_0, g_0) \in \Theta \times \mathcal{G}$ such that $p(\cdot;\theta_0,g_0) \equiv p_0(\cdot)$. Complications arise because the class of semiparametric models $\mathcal{P}$ could be *partially identified*, i.e., there are (θ_0,g_0) and (θ_1,g_1) in $\Theta \times \mathcal{G}$ such that $p_0 \equiv p(\cdot;\theta_1,g_1) = p(\cdot;\theta_0,g_0)$ but $(\theta_1,g_1) \neq (\theta_0,g_0)$; Given a random sample $\{Z_i = (Y_i,X_i)\}_{i=1}^n$ and the class of models $\mathcal{P}$, we are interested in inference on $\theta \in \Theta$ allowing for partial identification (although we initially focus on θ as the parameter of interest, the formal results allow for the parameter of interest to be a finite dimensional function of (θ,g)).

To conduct inference, we use maximum likelihood which is a natural approach in our setting. The sample log-likelihood objective function is

$$L_n(\theta,g) = \sum_{i=1}^n \log p(Z_i;\theta,g).$$

We define the identified set Θ_I for parameters of interest θ to be

$$\Theta_I \equiv \arg\sup_{\theta\in\Theta} \left(\sup_{g\in\mathcal{G}} E_0[\log p(Z_i;\theta,g)] \right) \tag{2.1}$$

This set can also be defined as

$$\Theta_I = \{\theta \in \Theta : p(\cdot,\theta,g) = p_0(\cdot) \text{ for some } g \in \mathcal{G}\} \tag{2.2}$$

[4]Sometimes we also use $E_Z()$ or $E_{P_Z}()$ or $E_{p_0}()$ for $E_0()$.

The above likelihood procedure would still be reasonable if the model is not correctly specified (i.e., (2.2) is empty) although now the interpretation of (2.1) is the set of pseudo true parameters minimizing the KL distance.

Next, we provide some motivating examples.

1) Unobserved Heterogeneity in a Heckman-Singer Model Accounting for unobserved heterogeneity across individuals or firms has become a quintessential ingredient in modern microeconometric models. An early work on this is that of Heckman and Singer (1984) (HS) in the context of estimating a job duration distribution in the presence of unobserved heterogeneity. There, the density of observed durations $p_0(t)$ is related to an economic job search model via the following integral equation

$$p_0(t) = \int_u f(t|u;\theta)dg(u) \tag{2.3}$$

where $f(\cdot|u;\theta)$ is the density of duration conditional on unobserved heterogeneity u, and $g(.)$ is the distribution of u. Economic theory typically provides suggestions about the functional form of $f(\cdot|u;\theta)$ but on the other hand, it is rarely the case that economists have information about the form of the distribution g of the unobserved heterogeneity. The empirical question of interest is whether information about θ is sensitive to assumptions made about the functional form of $g(.)$.

For a given $f(\cdot|u;\theta)$, HS provided conditions for point identification of θ when $g(.)$ is nonparametric, and these conditions show that this is possible in the limit as durations approach zero.

It is common in empirical papers to derive the likelihood conditional on some unobservable random variable (u here) that stands for unobserved heterogeneity, and then obtain the observed likelihood by integrating out the unobserved heterogeneity[5][6]. It is not easy to derive point identification conditions for θ in these models without making a functional form restriction on the distribution of u. In addition, HS style sufficient conditions for point identification of θ often times rely on this identification in the limit argument which creates practical (and theoretical) difficulties, such as slower than root n rates of convergence. Our approach to inference in this class of models 1) is valid whether or not θ is point identified, and 2) remains valid if point identification of θ is non-regular. In Section 7 below, we esti-

[5]Typically, the unobserved heterogeneity is assumed to have discrete support with finite known support points, and so the problem becomes one of inference in a discrete finite mixture model where the mixing probabilities become parameters. It is hard in these models to establish conditions under which the parameters are point identified.

[6]In recent empirical IO models, unobserved heterogeneity is motivated as a market level variable that is observed by the players, but not the econometrician.

mate a version of this duration model using NLSY data and compare our estimates of θ to ones obtained from standard parametric approaches.

2) Dynamic Binary Response Panel Data Model Consider the following dynamic panel data binary response model with individual effects

$$y_{it} = 1\left\{x_{it}'\beta + y_{i,t-1}\gamma + u_i + \varepsilon_{it} \geq 0\right\}. \tag{2.4}$$

We observe a sample of N individuals for T time periods starting with $t = 1$, to get a sample of $\{(y_{it}, x_{it})\}_{t=1}^{T}$ where we use the notation $x_i^T = (x_{i1}, ..., x_{iT})$. For each i, u_i is an individual specific random variable that is unobserved. The presence of a lagged dependent variable in the above model requires one to model the distribution of the first period of individual i's economic life, $g(u_i, x_i^T) \equiv P(y_{i1} = 1|u_i, x_i^T)$. Here, let $F_u(u_i)$ be the distribution of u_i which is assumed here, as in Honoré and Tamer (2006), to be independent of x_i^T and let the parameter of interest be $\theta \equiv (\beta, \gamma, \sigma^2)$ and the nuisance function is $(g(.,.), F_u(.))$. Here, we allow for the regressors x to be continuous. Finally, ε_{it} is a standard normal random variable that is i.i.d. across t and i and statistically independent of x_i^T (again, these assumptions are all imposed here for simplicity and illustration).
The conditional probability density of $y_i^T = (y_{i1}, ..., y_{iT})$ given x_i^T can be written as follows

$$\begin{aligned} & p(y_i^T|x_i^T, \theta, g, F_u) \\ = & \int \{g(u, x_i^T)\}^{y_{i1}}\{1 - g(u, x_i^T)\}^{1-y_{i1}} \prod_{t=2}^{T} P(y_{it}|x_i^T, y_{it-1}; \theta, u) dF_u(u) \end{aligned}$$

where

$$P\left(y_{it} = 0|\, x_i^T, y_{it-1}; \theta, u\right) = \Phi(-x_{i,t}'\beta - \gamma y_{i,t-1} - u).$$

Economists rarely have any information about the shape of the initial distribution $g()$ and so inference on θ when using a parametric form for g might be sensitive to these ad-hoc assumptions made on g. On the other hand, Honoré and Tamer (2006) show that without making assumptions on g, the parameter θ is partially identified in a pure random effects version of the above model in which F_u (the distribution of u) is assumed to be known up to finite dimensional parameter. They provide a linear programming method for characterizing the identified set on θ based on a minimum distance approach. Though sharp, the approach is problematic when x contains a continuous regressor. In this paper, we replace g with a sieve function and construct confidence regions for θ after profiling out the function g. The size (and shape) of the confidence regions we obtain partially reflect the information we have about θ. Though the model is meant to study the sensitivity of the parameters

to specification of the initial condition distribution, one can in principle also study the sensitivity to specification of the u or ϵ distribution also.

3) Inference in Discrete Games There has been a lot of work recently on inference in discrete games. In the two player version of a discrete game, interest is focused on the finite dimensional parameter θ in the two equation system

$$\begin{aligned} y_1 &= 1\left[x_1'\beta_1 + \Delta_1 y_2 + \epsilon_1 + \nu_1 \geq 0\right] \\ y_2 &= 1\left[x_2'\beta_2 + \Delta_2 y_1 + \epsilon_2 + \nu_2 \geq 0\right] \end{aligned} \tag{2.5}$$

This is a representation of a bivariate discrete game with 2 decision makers and with a general information structure: the public information that is unobservable to the econometrician is the vector (ϵ_1, ϵ_2) while player 1's private information is captured in ν_1, and similarly for player 2. The vector (ν_1, ν_2) is not observed by the econometrician. We allow the ϵ's to be correlated (and observed by the players), while the ν's are independent. This game was recently studied by Grieco (2011) and is termed a discrete game with incomplete information (as opposed to games with pure incomplete information, this game allows one to have incomplete information *and* unobserved heterogeneity). There can be multiple equilibria in this game. Assuming that the epsilons are normally distributed, and that from the perspective of the econometricians the ν's are also normal, Grieco shows that one can write the observed data distribution as

$$P_y(\mathbf{x}, \theta, g(.)) = \int \sum_{e \in \mathcal{E}(\epsilon, x, \theta)} g^e(\epsilon, x) P_y^e(\epsilon, x; \theta) dF_\epsilon \tag{2.6}$$

where θ is the finite dimensional vector of parameters that include $(\beta_1, \beta_2, \Delta_1, \Delta_2)$ and the parameters of the joint distribution of the ϵ's, $\mathcal{E}(.)$ is the set of equilibria that is known up to θ, $P_y^e(.)$ is the probability of the outcome $y = (y_1, y_2)$ given equilibrium e, and $g(.)$ is the unknown selection function here. A likelihood approach to inference in this game is attractive since it delivers the sharp set, which by definition is the argmax of the likelihood of the observed data. Generally, economists have no information about the functional forms of g and, without assumptions on g, it is hard to obtain sufficient conditions for point identification of θ (see Grieco for some of these conditions). So, our approach in this model is attractive, since we profile out the unknown selection functions $g(.)$ and provide a way to construct confidence regions on θ that hold whether or not θ is identified. Grieco uses this approach to fully estimate a model of entry and exit of grocery stores. More generally, studying inference in statistical structures as in the mixture model in (2.6) above is important in *any model* with multiple equilibria, or multiple potential outcome because the likelihoods of these models involve a selection function and since economists do not typically have any

information about these functions. In Section 7, we estimate a simpler version of the model where we restrict the information structure to a game of complete information and estimate the entry decisions of airlines in various markets.

4) Schooling Model with Discretized Outcomes We estimate an intergenerational schooling example where the aim is to examine the impact of parents' schooling on a child's level of education. So, here, for parents whose child is still in school, this child's final level of education would be missing but known to belong to one of finitely many bins. In particular, consider the problem of inference on a parameter θ in the linear regression model

$$y = x'\theta + \epsilon$$

where for some observations, y (a child's education) is missing. All we know in these cases is that y belongs to one of a finite number of ranges. For example, in a Tobit model, we know that missing y are ones for which y is negative. A common approach to conduct inference here is to assume a parametric functional form for the distribution F_ϵ of ϵ (as in an ordered probit/Tobit model). Estimates in such nonlinear parametric models are known to be sensitive to the specification of such a distribution. Let $d = 1$ signify that y is missing, and $d = 0$ otherwise, and when it is missing, we know that y belongs to one of k intervals $[y_i, y_{i+1}]$ for $i = 1, \ldots, k$ where the y_i's are fixed constants. Assuming that ϵ is independent of x, with unknown distribution $F(.)$ with mean zero, the likelihood of an observation is

$$(F(y_{j+1} - x'\theta) - F(y_j - x'\theta))^{[d=1]} f(y - x'\beta)^{[d=0]}$$

for some $j \in \{1, \ldots, k\}$. Notice here that this is a kind of an "ordered Tobit" model in that when the outcome is not observed, it takes finitely many values, and the likelihood of these values are uniquely determined by the distribution of ϵ. In Section 7, we use the same data as in Plug (2004) to estimate this model and compare our results to other parametric models.

3 Consistency and convergence rate of sieve MLE

In this section we present general consistency and rate of convergence results that allow for partial identification in semiparametric likelihood models. These results extend the existing set consistency results to cover cases where the parameter space is potentially infinite dimensional non-compact. Also, we extend the consistency results for sieve MLE to cover cases where the parameters are potentially partially identified. We first provide various definitions needed for the statement of the results.

Definition 3.1 *Let p_1 and p_2 be two densities with respect to a σ-finite measure μ. Define the following pseudo distances $D(p_1, p_2)$ between p_1 and p_2:*

1. *Squared **Hellinger distance**: $H^2(p_1, p_2) \equiv \int (\sqrt{p_1} - \sqrt{p_2})^2 d\mu = 2\left[1 - \int \sqrt{p_1 p_2} d\mu\right] = E_{p_2}\left[\left(\sqrt{\frac{p_1}{p_2}} - 1\right)^2\right]$.*
2. ***Pearson distance:** $\chi^2(p_1, p_2) \equiv \int \left(\frac{p_1}{p_2} - 1\right)^2 p_2 d\mu = E_{p_2}\left[\left(\frac{p_1}{p_2} - 1\right)^2\right]$ which is also called the χ^2 distance.*
3. *γ-**divergence:** $\rho_\gamma(p_1, p_2) \equiv \int \gamma^{-1}\left[\left(\frac{p_1}{p_2}\right)^\gamma - 1\right] p_1 d\mu$ for $\gamma \in (-1, 0)$ or $(0, 1]$.*
4. ***Kullback-Leibler divergence:** $K(p_1, p_2) \equiv \int p_1 \log(p_1/p_2) d\mu = E_{p_1}[\log(p_1) - \log(p_2)]$ (if $p_1 << p_2$ and $= +\infty$ otherwise).*

Remark 3.1 *From Definition 3.1, it is easy to see that*

$$H^2(p_1, p_2) = \rho_{-1/2}(p_1, p_2), \quad K(p_1, p_2) = \rho_{0+}(p_1, p_2), \quad \chi^2(p_1, p_2) = \rho_1(p_1, p_2),$$

$$H^2(p_1, p_2) \leq K(p_1, p_2) \leq \rho_\gamma(p_1, p_2) \leq \chi^2(p_1, p_2) \quad for \quad \gamma \in (0, 1].$$

Denote $\mathcal{A} \equiv \Theta \times \mathcal{G}$ as the parameter space and $\mathcal{P} \equiv \{p(\cdot; \theta, g) : (\theta, g) \in \mathcal{A}\}$ as the family of probability models. For any $p_1, p_2 \in \mathcal{P}$, let $D(p_1, p_2)$ be any one of the distances in Definition 3.1. Then: $D(p_1, p_2) \geq 0$, and $D(p_1, p_2) = 0$ iff $p_1 = p_2$ $a.s. - \mu$. Under correct specification, we can define the identified set for all the parameters as

$$\begin{aligned} \mathcal{A}_I &\equiv \Theta_I \times \mathcal{G}_I \equiv \{\alpha = (\theta, g) \in \mathcal{A} : p(\cdot; \alpha) = p_0(\cdot)\} \\ &= \{\alpha = (\theta, g) \in \mathcal{A} : D(p_0, p(\cdot; \alpha)) = 0\} \end{aligned}$$

Also, Θ_I defined in (2.1) could be expressed as

$$\Theta_I = \{\theta \in \Theta : D(p_0, p(\cdot; \theta, g)) = 0 \text{ for some } g \in \mathcal{G}\}.$$

Let $\mathcal{A}_{k(n)} \equiv \Theta \times \mathcal{G}_{k(n)}$. Let $\mathcal{P}_{k(n)} \equiv \{p(\cdot; \theta, g) : (\theta, g) \in \mathcal{A}_{k(n)}\}$ be a sequence of sieve spaces that is dense in $\mathcal{P} \equiv \{p(\cdot; \theta, g) : (\theta, g) \in \mathcal{A}\}$ under one of the $D(p_1, p_2)$ distances as given in Definition 3.1. Denote $\widehat{p}(\cdot) \equiv p(\cdot; \widehat{\theta}, \widehat{g})$ as the η_n-sieve MLE which is defined as follows:

$$\frac{1}{n}\sum_{i=1}^n \log p(Z_i; \widehat{\theta}, \widehat{g}) \geq \sup_{(\theta, g) \in \mathcal{A}_{k(n)}} \frac{1}{n}\sum_{i=1}^n \log p(Z_i; \theta, g) - \eta_n \quad \text{with } \eta_n = o_{P_Z}(1); \tag{3.1}$$

or equivalently,

$$\frac{1}{n}\sum_{i=1}^{n}\log\widehat{p}(Z_i) \geq \sup_{p\in\mathcal{P}_{k(n)}}\frac{1}{n}\sum_{i=1}^{n}\log p(Z_i) - \eta_n \quad \text{with } \eta_n = o_{P_Z}(1);$$

It is easy to see that the η_n-sieve MLE $\widehat{p}(\cdot) \equiv p(\cdot;\widehat{\theta},\widehat{g})$ defined in (3.1) is numerically equivalent to the following η_n-sieve profile MLE $\widehat{p}(\cdot) \equiv p(\cdot;\hat{\theta},\tilde{g}^{\hat{\theta}})$, computed in two steps where in the Step 1 we get for each $\theta \in \Theta$,

$$\frac{1}{n}\sum_{i=1}^{n}\log p(Z_i;\theta,\tilde{g}^{\theta}) \geq \sup_{g\in\mathcal{G}_{k(n)}}\frac{1}{n}\sum_{i=1}^{n}\log p(Z_i;\theta,g) - \eta_n \quad \text{with } \eta_n = o_{P_Z}(1).$$

and then in Step 2, compute

$$\hat{\theta} \in \widehat{\Theta} = \arg\max_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\log p(Z_i;\theta,\tilde{g}^{\theta}), \quad \widehat{g} = \tilde{g}^{\hat{\theta}} \in \left\{\tilde{g}^{\hat{\theta}} \in \mathcal{G}_{k(n)} : \hat{\theta} \in \Theta\right\}.$$

In the next assumption, we provide the various conditions that we need for consistency of the sieve MLE estimators.

Assumption 3.1 *Let the followings hold:*

1. ***Parameter space and objective function:*** *(i) $\mathcal{A} = \Theta\times\mathcal{G} \subseteq \mathbf{A} = \Re^{d_\theta}\times\mathbf{G}$, Θ is a compact, nonempty subset of a Euclidean space $(\Re^{d_\theta}, |\cdot|_e)$, and $\mathcal{G}$ is a closed, bounded and nonempty subset of a separable infinite dimensional Banach space $(\mathbf{G}, ||\cdot||_G)$; (ii) $E_0[\log p(Z;\theta,g)]$ is upper semicontinuous on $\mathcal{A}$ under $||\alpha||_A = |\theta|_e + ||g||_G$; (iii) the identified set, $\mathcal{A}_I = \Theta_I \times \mathcal{G}_I = \{\alpha = (\theta,g) \in \mathcal{A} : D(p_0, p(\cdot;\alpha)) = 0\}$ is a nonempty, closed and bounded strict subset of $(\mathcal{A}, ||\cdot||_A)$.*

2. ***Sieve space*** *(i) for each $k \geq 1$, $\mathcal{A}_k = \Theta\times\mathcal{G}_k \subseteq \mathcal{A}$, $\mathcal{G}_k$ is closed under $||\cdot||_G$ with $\dim(\mathcal{G}_k) < \infty$; (ii) $\varnothing \neq \mathcal{G}_k \subseteq \mathcal{G}_{k+1} \subseteq \mathcal{G}$ for all $k \geq 1$, and $\overline{\cup_{k=1}^{\infty}\mathcal{G}_k}$ is dense in $\mathcal{G}$ under $||\cdot||_G$. That is, for any $g \in \mathcal{G}$, there is $\Pi_k g \in \mathcal{G}_k$ such that $||g - \Pi_k g||_G \to 0$ as $k \to \infty$.*

3. ***Penalty function*** *There is a function $Pen : \mathcal{G} \to [0,\infty)$ such that: (i) $Pen(.)$ is a measurable function such that $\sup_{g\in\mathcal{G}_I} Pen(g) < \infty$; (ii) the set $\{g \in \mathcal{G} : Pen(g) \leq M\}$ is compact under $||\cdot||_G$ for all $M \in [0,\infty)$; (iii) $\lambda_n > 0$, and $\lambda_n \sup_{g\in\mathcal{G}_I} |Pen(\Pi_n g) - Pen(g)| = O(\lambda_n) = o(1)$.*

4. ***Uniform convergence on sieve space*** *(i) the data $\{Z_i = (Y_i, X_i)\}_{i=1}^n$ are a random sample of $Z = (Y,X)$ from a unique density p_0; (ii) $E_0\{\sup_{\alpha\in\mathcal{A}_n} |\log p(Z;\alpha)|\}$ is*

bounded; there are a finite $s > 0$ and a random variable $U(Z)$ with $E_0\{U(Z)\} < \infty$ such that

$$\sup_{\alpha,\alpha' \in \mathcal{A}_n : ||\alpha - \alpha'||_A \leq \delta} |\log p(Z;\alpha) - \log p(Z;\alpha')| \leq \delta^s U(Z),$$

and $\log N(\delta^{1/s}, \mathcal{A}_n, ||\cdot||_A) = o(n)$ *for all* $\delta > 0$.

Assumption 3.1.1 is standard and provides conditions on the parameter space. We do require that the identified set is a strict subset of the overall parameter space to avoid cases for which the identified set is the whole parameter space. The assumption that Θ is compact is not needed but we impose it since it is a standard one for semiparametric models. Assumption 3.1.2 concerns conditions on the sieve approximation that are needed. The penalty function is used to regularize the optimization problem as in Chen and Pouzo (2011). Finally, Assumption 3.1.4 implies uniform convergence of the objective function over the sieve space.

The next theorem presents consistency in the one sided Hausdorff metric.

Theorem 3.1 *Let Assumption 3.1 hold. Let $\widehat{\mathcal{A}}_n$ be the collection of $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{g}_n) \in \mathcal{A}_{k(n)} = \Theta \times \mathcal{G}_{k(n)}$ that solves*

$$\frac{1}{n}\sum_{i=1}^{n} \log p(Z_i;\widehat{\alpha}_n) - \lambda_n Pen(\widehat{g}_n) = \sup_{\alpha \in \mathcal{A}_{k(n)}} \left[\frac{1}{n}\sum_{i=1}^{n} \log p(Z_i;\alpha) - \lambda_n Pen(g)\right].$$

Then:

1. $K(p_0, \widehat{p}) = o_{as-Z}(1)$;
2. $d_A(\widehat{\alpha}_n, \mathcal{A}_I) \equiv \inf_{\alpha \in \mathcal{A}_I} ||\widehat{\alpha}_n - \alpha||_A = o_{P_Z}(1)$ *and* $Pen(\widehat{g}_n) = O_{P_Z}(1)$.

In the Appendix, we provide and prove a more general consistency Theorem that holds for semiparametric extremum estimators. Given the consistency result, the next Remark provides information about the relationship between the various distances.

Remark 3.2 *By our consistency result above, we know that $\widehat{p}$ is "close" to p_0, the true density, as sample size increases. For densities that are close, we provide below relations among the various distances from Remark 3.1.*
1) *By Remark 3.1 we have $[H(p,p_0)]^2 \leq K(p_0,p) \leq \chi^2(p_1,p_2)$ for all $p \in \mathcal{P}$. When $H(p,p_0)$ is small, using the Taylor expansion of $\log(1+x) = x - 0.5x^2(1+o(1))$ for small x, we have the following:*

$$K(p_0,p) \equiv -E_0[\log p - \log p_0] = 2[H(p,p_0)]^2(1+o(1)) = \frac{1}{2}\chi^2(p,p_0)(1+o(1)).$$

2) *Theorem 5 of Wong and Shen (1995) states that for all* $\epsilon^2 \leq 0.5(1-e^{-1})^2$ *and for some* $\gamma \in (0,1]$*, we have*

$$K(p_0,p) \leq \left[const. + \frac{8}{\gamma}\max\left(1, \log\left(\frac{m_\gamma}{\epsilon}\right)\right)\right] \times \epsilon^2$$

for all $p \in \mathcal{P}_\epsilon$*, where*

$$\mathcal{P}_\epsilon \equiv \left\{p \in \mathcal{P} : H(p,p_0) \leq \epsilon,\ m_\gamma^2 \equiv E_{p_0}\left[\left(\frac{p_0}{p}\right)^\gamma \times 1\{\left(\frac{p_0}{p}\right)^\gamma \geq e\}\right] < \infty\right\}.$$

3) *Lemma 3.1 of Liu and Shao (2003) shows that for some small* $\epsilon > 0$*, if the class of functions*

$$\mathcal{D}\chi \equiv \left\{s = \frac{\frac{p}{p_0}-1}{\chi(p,p_0)} : H(p,p_0) \leq \epsilon, p \in \mathcal{P}\backslash\{p_0\}\right\},$$

has a square integrable envelop function, i.e., $E_0\left[\left(\sup_{s\in\mathcal{D}\chi}|s|\right)^2\right] < \infty$*, then:*

$$\lim_{\chi(p,p_0)\to 0}\frac{4H^2(p,p_0)}{\chi^2(p,p_0)} = \lim_{H(p,p_0)\to 0}\frac{4H^2(p,p_0)}{\chi^2(p,p_0)} = 1.$$

Next, we derive rates of convergence results under partial identification. We can directly apply Wong and Shen (1995) or Birgé and Massart (1998) to obtain convergence rate of $H(\widehat{p},p_0)$ under mild conditions. In the following, we denote

$$e_n(\gamma) = \inf_{p\in\mathcal{P}_{k(n)}}\rho_\gamma(p_0,p) \quad for \quad \gamma \in [0+,1]; \tag{3.2}$$

$$\epsilon_n = \inf\left[\epsilon > 0 : \int_{2^{-8}\epsilon^2}^{\sqrt{2}\epsilon}\sqrt{\log N_{[]}(u,\{p\in\mathcal{P}_{k(n)} : H(p,p_0)\leq 2\epsilon\}, H())}du \leq const.\sqrt{n}\epsilon^2\right], \tag{3.3}$$

where $e_n(\gamma)$ is the bias (or the sieve approximation error) under the distance ρ_γ for $\gamma \in [0+,1]$ (see Definition 3.1), and ϵ_n is the measure of sieve model complexity in terms of Hellinger distance with bracketing. Next, we state results on rates of convergence in terms of Hellinger (or Pearson) distance. These results are direct application of Wong and Shen (1995) and hence we omit its proof.

Theorem 3.2 *Let* $\delta_n = \max\left[\epsilon_n, \sqrt{e_n(\gamma)}\right]$ *for* $\gamma \in [0+,1]$*. Under all the conditions of Theorem 3.1 with* $\eta_n = O(\lambda_n) = O([\delta_n]^2)$*, we have: (1)* $H(p(\cdot,\widehat{\alpha}),p(\cdot,\alpha_0)) = O_{P_Z}(\delta_n)$ *for all* $\widehat{\alpha} \in \widehat{\mathcal{A}}_n$ *and all* $\alpha_0 \in \mathcal{A}_I$*; (2) If* $\delta_n = \max\left[\epsilon_n, \sqrt{e_n(1)}\right]$*, then* $\chi(p(\cdot,\widehat{\alpha}),p(\cdot,\alpha_0)) = O_{P_Z}(\delta_n)$ *for all* $\widehat{\alpha} \in \widehat{\mathcal{A}}_n$ *and all* $\alpha_0 \in \mathcal{A}_I$*.*

We provide next a key point regarding characterization of the identified set. An important insight in what follows is that although the likelihood can be maximized on a set, and consistency is based on convergence of some set distances based on the norm $||\cdot||_A$ we define above, *all* the elements of the identified set are "equivalent" in that they induce the same density (P_0), and so the identified set in some way is a *singleton*. We elaborate on this point next. Let

$$D(\alpha_1, \alpha_2) = 2H(p(\cdot, \alpha_1), p(\cdot, \alpha_2)) \text{ or } = \chi(p(\cdot, \alpha_1), p(\cdot, \alpha_2))$$

denote either the rescaled Hellinger distance or Pearson distance induced metric on the parameter space $\mathcal{A}$. Then $D(\alpha, \alpha_0) = D(\alpha, \alpha_0')$ for all $\alpha_0, \alpha_0' \in \mathcal{A}_I$ and all $\alpha \in \mathcal{A}$. So, $\mathcal{A}_I$ *is a singleton (or unique* $\{\alpha_0\}$ *in terms of equivalent class) under the distance* $D(\cdot, \cdot)$ *although* $\mathcal{A}_I$ *is not a singleton under* $||\cdot||_A$. To stress this difference, sometimes we use notations $\{\alpha_0^D\} = (\mathcal{A}_I, D(\cdot, \cdot))$ and $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$. Define

$$\alpha_{0n}^D \equiv \arg\min_{\alpha \in \mathcal{A}_n} D(\alpha, \alpha_0^D) = \arg\min_{\alpha \in \mathcal{A}_n} D(\alpha, \alpha_0). \tag{3.4}$$

Then $D(\alpha_{0n}^D, \alpha_0) \leq const.\sqrt{e_n(\gamma)} = O(\delta_n)$ for all $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$. In the rest of the paper we shall focus on the case of $D(\alpha, \alpha_0) = \chi(p(\cdot, \alpha), p(\cdot, \alpha_0))$ and $\delta_n = \max\left[\epsilon_n, \sqrt{e_n(1)}\right]$, but given Remark 3.2, all our asymptotic results remain valid with $D(\alpha, \alpha_0) = 2H(p(\cdot, \alpha), p(\cdot, \alpha_0))$.

Remark 3.3 *(1) For any* $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$ *let* $\mathcal{B}(\alpha_0) \equiv \{\alpha \in \mathcal{A} : D(\alpha, \alpha_0) \leq \delta_n \log\log n\}$ *and* $\mathcal{B}_n(\alpha_0) \equiv \{\alpha \in \mathcal{A}_{k(n)} : D(\alpha, \alpha_0) \leq \delta_n \log\log n\} = \mathcal{B}(\alpha_0) \cap \mathcal{A}_{k(n)}$. *Then it is clear that* $\mathcal{B}(\alpha_0) = \mathcal{B}(\alpha_0^D)$ *and* $\mathcal{B}_n(\alpha_0) = \mathcal{B}_n(\alpha_0^D)$ *for all* $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$. *By Theorem 3.2 we have:* $\alpha_{0n}^D \in \mathcal{B}_n(\alpha_0)$ *and* $\widehat{\alpha}_n \in \widehat{\mathcal{A}}_n \subset \mathcal{B}_n(\alpha_0)$ *with probability approaching one (wpa1) uniformly in* $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$.

(2) By Remark 3.2, there is a positive sequence $\zeta_n = o(1)$ *such that*

$$\sup_{\alpha \in \mathcal{B}(\alpha_0): D(\alpha, \alpha_0) \neq 0} \frac{K(p(\cdot, \alpha_0), p(\cdot, \alpha))}{\frac{1}{2}[D(\alpha, \alpha_0)]^2} = 1 + o(\zeta_n).$$

The next Section provides the main results in the paper on the asymptotic distribution of the LR statistic.

4 Sieve Likelihood Ratio Statistic

We derive the asymptotic distribution of the profiled sieve log-likelihood ratio statistic

$$LR(\theta_0) \equiv 2\left[\sup_{\theta \in \Theta, g \in \mathcal{G}_{k(n)}} \sum_{i=1}^n \log p(Z_i; \theta, g) - \sup_{g \in \mathcal{G}_{k(n)}} \sum_{i=1}^n \log p(Z_i; \theta_0, g)\right]. \tag{4.1}$$

Our inference is criterion based, and so to build confidence regions that reflect the sensitivity of the model with respect to specification of g, we use the LR test statistic to build confidence regions for parameters of interest. This requires first that we show that the LR statistic above admits a nondegenerate asymptotic distribution. Providing conditions under which this holds constitute one of the main theoretical results of the paper.

Typically, for a regular parametric likelihood model $\{p(\cdot,\theta) : \theta \in \Theta\}$ without unknown functions g, deriving the asymptotic distribution of a parametric LR statistic, $LR(\theta_0) = 2[\sup_{\theta\in\Theta} L_n(\theta) - L_n(\theta_0)]$, under the null of $\theta = \theta_0 \in int(\Theta)$ uses a quadratic approximation to the sample log-likelihood $L_n(\theta) = \sum_{i=1}^{n} \log p(Z_i;\theta)$ in a Euclidean $n^{-1/2}$ neighborhood of the true parameter θ_0:

$$L_n(\theta) - L_n(\theta_0) = (\theta-\theta_0)' \sum_{i=1}^{n} \overline{s}(Z_i) - \frac{n}{2}(\theta-\theta_0)' E_{\theta_0}\left[\overline{s}(Z)\overline{s}(Z)'\right](\theta-\theta_0)(1+o_{P_Z}(1)),$$

where $\overline{s}(z) = \frac{d}{d\theta}\log p\left(z;\theta\right)|_{\theta=\theta_0}$ is the score function and $\mathcal{I}_{\theta_0} \equiv E_{\theta_0}\left[\overline{s}(Z)\overline{s}(Z)'\right]$ is the Fisher information matrix. Suppose that $\mathcal{I}_{\theta_0}$ is non-singular, then one immediately obtains that $|\widehat{\theta}-\theta_0|_e = O_{P_Z}\left(n^{-1/2}\right)$, $LR(\theta_0) = 2[L_n(\widehat{\theta}) - L_n(\theta_0)] = O_{P_Z}\left(1\right)$ and that $LR(\theta_0)$ is asymptotically Chi-Square distributed under the null. See, e.g. Chernoff (1954). Without point identification, this quadratic approximation in a $n^{-1/2}$ Euclidean neighborhood of θ_0 is not natural since the ML estimator $\widehat{\theta}$ may not converge to any fixed point θ_0 in the identified set Θ_I and the Fisher information $\mathcal{I}_{\theta_0}$ could be singular for some θ_0 under the null. These problems arise in finite mixture models, Markov switching models, and some other parametric models. Recently Liu and Shao (2003) use a novel approach to deriving the asymptotic null distribution of the parametric LR statistic under partial identification. Whereas the parameter θ is not unique under the null, the true parametric probability density is unique (the density of the data). So Liu and Shao (2003) obtain a quadratic expansion to $L_n(\theta)$ in a Hellinger (or Pearson) $n^{-1/2}$ neighborhood of the true density $p_0(\cdot) = p(\cdot,\theta_0)$:

$$\begin{aligned} L_n(\theta) - L_n(\theta_0) &= 2H(\theta,\theta_0)\sum_{i=1}^{n}\left[s_H(Z_i;\theta) - E_0(s_H(Z;\theta))\right] - 2n[H(\theta,\theta_0)]^2(1+o_{P_Z}(1)) \\ &= \chi(\theta,\theta_0)\sum_{i=1}^{n} s_\chi(Z_i;\theta) - \frac{n}{2}[\chi(\theta,\theta_0)]^2(1+o_{P_Z}(1)), \end{aligned}$$

where $s_H(z;\theta) \equiv [\sqrt{p(z,\theta)/p_0(z)}-1]/H(\theta,\theta_0)$ (or $s_\chi(z;\theta) \equiv [p(z,\theta)/p_0(z)-1]/\chi(\theta,\theta_0)$) is a so-called generalized score function. Let $D(\cdot,\cdot)$ denote Hellinger or Pearson distance. Under a key assumption that the class of generalized score functions $\{s_D(\cdot;\theta) : \theta\in\Theta, 0 < D(\theta,\theta_0) \le \delta\}$ is Donsker in $L^2(P_Z)$ for a small $\delta > 0$, Liu and Shao (2003) establish that $D(\widehat{\theta},\theta_0) = O_{P_Z}\left(n^{-1/2}\right)$ and $LR(\theta_0) = 2[L_n(\widehat{\theta}) - L_n(\theta_0)] = O_{P_Z}\left(1\right)$, which is then used to obtain the

asymptotic null distribution of the parametric LR statistic $LR(\theta_0)$. Without point identification of the parameter θ_0, the asymptotic null distribution of this parametric LR statistic is very complicated; the existing literature has focused on characterizing this complicated asymptotic null distribution for simple and specific parametric likelihood models; see, e.g., Liu and Shao (2003).

One might wish to directly generalize the approach of Liu and Shao (2003) to the semi-parametric likelihood model $\mathcal{P} = \{p(\cdot;\alpha) : \alpha = (\theta, g) \in \Theta \times \mathcal{G}\}$ with infinite dimensional nuisance functions g, without assuming point identification of θ and g. By Theorem 3.2, any sieve MLE $\widehat{\alpha}_n \in \arg\max_{\alpha \in \mathcal{A}_{k(n)}} \sum_{i=1}^n \log p(Z_i;\alpha)$ has the Hellinger (or Pearson) distance convergence rate $D(\widehat{\alpha}_n, \alpha_0) = O(\delta_n)$. Under some regularity conditions, one can show that the sample log-likelihood $L_n(\alpha) \equiv \sum_{i=1}^n \log p(Z_i;\alpha)$ admits the following expansion in a Hellinger (or Pearson) δ_n-neighborhood of the true density $p_0(\cdot) = p(\cdot,\alpha_0) = p(\cdot,\theta_0,g_0)$:

$$\begin{aligned}\frac{L_n(\alpha) - L_n(\alpha_0)}{n} &= \frac{2H(\alpha,\alpha_0)}{n}\sum_{i=1}^n [s_H(Z_i;\alpha) - E_0(s_H(Z;\alpha))] - 2[H(\alpha,\alpha_0)]^2(1 + o_{P_Z}(1)) \\ &= \frac{\chi(\alpha,\alpha_0)}{n}\sum_{i=1}^n s_\chi(Z_i;\alpha) - \frac{1}{2}[\chi(\alpha,\alpha_0)]^2(1 + o_{P_Z}(1)),\end{aligned}$$

where

$$s_H(z;\alpha) \equiv \frac{\sqrt{p(z,\alpha)/p_0(z)} - 1}{H(\alpha,\alpha_0)}, \quad s_\chi(z;\alpha) \equiv \frac{\{p(z,\alpha)/p_0(z)\} - 1}{\chi(\alpha,\alpha_0)}$$

is the generalized score function. Under the naive assumption that the class of generalized score functions

$$\mathcal{S}_{k(n)} \equiv \left\{s_D(\cdot;\alpha) : \alpha \in \mathcal{A}_{k(n)}, 0 < D(\alpha,\alpha_0) \leq \delta\right\} \quad \text{is } L^2(P_Z) - \text{Donsker}$$

for a small $\delta > 0$, one would immediately obtain that $D(\widehat{\alpha},\alpha_0) = O_{P_Z}\left(n^{-1/2}\right)$ and $2[L_n(\widehat{\alpha}) - L_n(\alpha_0)] = O_{P_Z}(1)$. Unfortunately, the best convergence rate δ_n of any estimator (including the sieve MLE $\widehat{\alpha}$) for α_0 in Hellinger (or Pearson) distance is slower than $n^{-1/2}$ when $g \in \mathcal{G}$ is infinite dimensional.[7] This indicates that when $g \in \mathcal{G}$ is infinite dimensional the class of generalized score functions $\mathcal{S}_{k(n)}$ fails to be $L^2(P_Z)-$Donsker in general. The above expansion actually implies that $2[\sup_{\alpha \in \mathcal{A}_{k(n)}} L_n(\alpha) - L_n(\alpha_0)]$ diverges to infinity whenever $D(\widehat{\alpha},\alpha_0)/n^{-1/2} \to \infty$.

Let $\phi(\alpha) \equiv (\phi_1(\alpha), ..., \phi_{d_\phi}(\alpha))' : \mathcal{A} \to \Re^{d_\phi}$ be a $d_\phi-$ vector valued known functional for a fixed and finite d_ϕ. In this paper, we show that, even if $2[\sup_{\alpha \in \mathcal{A}_{k(n)}} L_n(\alpha) - L_n(\alpha_0)]$

[7] In fact, even if the class of density functions $\mathcal{P} = \{p(\cdot;\theta,g) : (\theta,g) \in \Theta \times \mathcal{G}\}$ is analytic, the best convergence rate in Hellinger distance is still slower than $n^{-1/2}$.

and $2[\sup_{\alpha\in\mathcal{A}_{k(n)}:\phi(\alpha)=r_0} L_n(\alpha) - L_n(\alpha_0)]$ may diverge, the following sieve log-likelihood ratio statistic:

$$LR(r_0) \equiv 2\left[\sup_{\alpha\in\mathcal{A}_{k(n)}} L_n(\alpha) - \sup_{\alpha\in\mathcal{A}_{k(n)}:\phi(\alpha)=r_0} L_n(\alpha)\right] \tag{4.2}$$

has a tight limiting distribution under the null hypothesis $H_0 : \phi(\alpha_0) = r_0 \in \Re^{d_\phi}$ for $\alpha_0 \in \mathcal{A}_I$. This type of null accommodates testing for example "marginal effects" in Example 2 and 4 in Section 2 for example where the parameter of interest can be a (finite dimensional) function of both g and θ. Of course, this result immediately implies that the profiled sieve log-likelihood ratio statistic $LR(\theta_0)$ defined in (4.1) has a tight limiting distribution under the null hypothesis $H_0 : \phi(\alpha_0) = \theta_0 \in \Theta_I$. Note here that the constrained likelihood in (4.2) above is at a *finite dimensional constraint* as opposed to evaluating the constrained likelihood at some α_0 where $\alpha_0 = (\theta_0, g_0)$ can be infinite dimensional as in what a naive extension of Liu and Shao (2003) would do. The above sieve log-likelihood ratio statistic (4.2) could still diverge to infinity when the constraint is infinite dimensional and we do not consider it here.

To provide conditions that are needed for our results, we introduce some definitions and notations. Define the unconstrained approximate sieve MLE $\widehat{\alpha}_n \in \widehat{\mathcal{A}}_n$ as

$$\widehat{\alpha}_n \in \widehat{\mathcal{A}}_n \equiv \arg\max_{\alpha\in\mathcal{A}_n}\left\{\sum_{i=1}^n \log p(Z_i;\alpha) - o_{P_Z}(1)\right\}.$$

Define the constrained approximate sieve MLE $\widetilde{\alpha}_n \in \widetilde{\mathcal{A}}_n$ as

$$\widetilde{\alpha}_n \in \widetilde{\mathcal{A}}_n \equiv \arg\max_{\{\alpha\in\mathcal{A}_n:\ \phi(\alpha)=r_0\}}\left\{\sum_{i=1}^n \log p(Z_i;\alpha) - o_{P_Z}(1)\right\}.$$

Let $\mathcal{A}_I^r \equiv \mathcal{A}_I \cap \{\alpha \in \mathcal{A} : \phi(\alpha) = r_0\} \neq \varnothing$. By Theorem 3.1 we have

$$\inf_{\alpha_0\in\mathcal{A}_I} ||\widetilde{\alpha}_n - \alpha_0||_A \leq \inf_{\alpha_0\in\mathcal{A}_I^r} ||\widetilde{\alpha}_n - \alpha_0||_A = d_A(\widetilde{\alpha}_n, \mathcal{A}_I^r) = o_{P_Z}(1),$$

and by Theorem 3.2 we have

$$D(\widetilde{\alpha}_n, \alpha_0) = O_{P_Z}(\delta_n) \text{ for all } \alpha_0 \in \mathcal{A}_I^r;$$

thus $\widetilde{\alpha}_n \in \mathcal{B}_n(\alpha_0) \cap \{\alpha \in \mathcal{A} : \phi(\alpha) = r_0\}$ with probability approaching one (wpa1) uniformly in $\alpha_0 \in \mathcal{A}_I^r$, where $\mathcal{B}_n(\alpha_0)$ is defined in Remark 3.3.

Let $\langle\cdot,\cdot\rangle$ denote the distance $\chi(\cdot,\cdot)$ or $||\cdot||_{L^2(P_Z)}$ induced inner product. For $\alpha_{0n}^D = \alpha_{0n}^\chi$ defined in (3.4) with distance $D = \chi$, let

$$\mathcal{V}_n = cl\left\{v(z) = \frac{p(z,\alpha) - p(z,\alpha_{0n}^D)}{p_0(z)} : \alpha \in \mathcal{B}_n(\alpha_0), E_0[v(Z)] = 0, E_0\left[(v(Z))^2\right] < \infty\right\}.$$

Then $\mathcal{V}_n \subset L_0^2(P_Z) \equiv \left\{v(z) : E_0[v(Z)] = 0, E_0\left[(v(Z))^2\right] < \infty\right\}$ which is a properly defined Hilbert space. For any candidate $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and any $\lambda \in U^{d_\phi} \equiv \{\lambda \in \Re^{d_\phi} : |\lambda|_e = 1\}$, if

$$\sup_{v=\frac{p(\cdot,\alpha)-p(\cdot,\alpha_{0n}^D)}{p_0(\cdot)} \in \mathcal{V}_n : \langle v,v\rangle \neq 0} \frac{\left|\lambda' \frac{\partial \phi(\alpha_0)}{\partial \alpha}[\alpha - \alpha_{0n}^D]\right|^2}{E_0\left[\left(\frac{p(Z,\alpha)-p(Z,\alpha_{0n}^D)}{p_0(Z)}\right)^2\right]} < \infty,$$

then there is some $v_n^*(\alpha_0, \lambda) \in \mathcal{V}_n$ such that

$$0 < ||v_n^*(\cdot, \alpha_0, \lambda)||^2 = \sup_{v=\frac{p(\cdot,\alpha)-p(\cdot,\alpha_{0n}^D)}{p_0(\cdot)} \in \mathcal{V}_n : \langle v,v\rangle \neq 0} \frac{\left|\lambda' \frac{\partial \phi(\alpha_0)}{\partial \alpha}[\alpha - \alpha_{0n}^D]\right|^2}{E_0\left[\left(\frac{p(Z,\alpha)-p(Z,\alpha_{0n}^D)}{p_0(Z)}\right)^2\right]} < \infty$$

and

$$\begin{aligned} \lambda' \frac{\partial \phi(\alpha_0)}{\partial \alpha}[\alpha - \alpha_{0n}^D] &= E_0\left[\left(\frac{p(Z,\alpha) - p(Z,\alpha_{0n}^D)}{p_0(Z)}\right) v_n^*(Z, \alpha_0, \lambda)\right] \\ &= E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)} - 1\right) v_n^*(Z, \alpha_0, \lambda)\right] \end{aligned}$$

A crucial point of a departure here from the semiparametric sieve literature is that for any pair $\alpha_0, \alpha_0' \in (\mathcal{A}_I^r, ||\cdot||_A)$, we have $D(\alpha_0, \alpha_0') = 0$, but if $||\alpha_0 - \alpha_0'||_A \neq 0$ then we could have that $\lambda' \frac{\partial \phi(\alpha_0)}{\partial \alpha}[v] \neq \lambda' \frac{\partial \phi(\alpha_0')}{\partial \alpha}[v]$ for some λ. Therefore for any candidate $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ *we use a different representer* $\langle v, v_n^*(\alpha_0, \lambda)\rangle$ for $\lambda' \frac{\partial \phi(\alpha_0)}{\partial \alpha}[v]$. If the model were point identified, then the representer would be unique and correspond to the true and point identified parameter.

Assumption 4.1 *Let it be known that uniformly in* $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$*, the following hold: (i)* $\frac{\partial \phi_j(\alpha_0)}{\partial \alpha}$ *is linear in* $\alpha - \alpha_0$ *for all* $j = 1, ..., d_\phi$*, and is linearly independent across* j*; (ii) uniformly in* $\lambda \in U^{d_\phi} = \{\lambda \in \Re^{d_\phi} : |\lambda|_e = 1\}$*,*

$$\frac{\left|\lambda' \frac{\partial \phi(\alpha_0)}{\partial \alpha}[\alpha_{0n}^D - \alpha_0]\right|}{||v_n^*(\alpha_0, \lambda)||} = o_{P_Z}(n^{-\frac{1}{2}}), \tag{4.3}$$

and

$$\frac{\left|\lambda'\{\phi(\alpha) - \phi(\alpha_0)\} - \lambda' \frac{\partial \phi(\alpha_0)}{\partial \alpha}[\alpha - \alpha_0]\right|}{||v_n^*(\alpha_0, \lambda)||} = o_{P_Z}(n^{-\frac{1}{2}}) \tag{4.4}$$

uniformly in $\alpha \in \mathcal{B}_n(\alpha_0)$*.*

The above Assumptions is simple to verify. For example, in cases where the null is of the form $H_0 : \phi(\alpha) = \theta_0$, it is trivially satisfied since this restriction is linear.

Next, denote

$$u_n^*(z,\alpha_0,\lambda) \equiv \frac{v_n^*(z,\alpha_0,\lambda)}{||v_n^*(\alpha_0,\lambda)||} = \frac{v_n^*(z,\alpha_0,\lambda)}{\sqrt{Var_0[v_n^*(Z,\alpha_0,\lambda)]}}.$$

Let also $\ell(Z,\alpha) \equiv \log p(Z,\alpha)$, $\chi(\alpha,\alpha_0) \equiv \chi(p(\cdot,\alpha),p_0)$ and for any $\alpha \in \mathcal{B}_n(\alpha_0)$, let

$$R(Z;\alpha,\alpha_0) \equiv \ell(Z,\alpha) - \ell(Z,\alpha_0) - \chi(\alpha,\alpha_0)s_\chi(Z;\alpha).$$

We now consider perturbation in probability density sieve space: for all $\alpha \in \mathcal{B}_n(\alpha_0)$ and $t \in \mathcal{T}_n \equiv \{t \in [-1,1] : |t| \leq const. \times n^{-1/2}\}$, we let

$$p(z,\alpha(t)) \equiv p(z,\alpha) + t \times u_n^*(z,\alpha_0,\lambda) \times p_0(z).$$

We denote an empirical process indexed by function f as $\mu_n(f) = \frac{1}{n}\sum_{i=1}^{n}[f(Z_i) - E_0(f(Z_i)]$. This then leads us to the next assumption that we require on the remainder.

Assumption 4.2 *Uniformly in $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A), \lambda \in U^{d_\phi}$, the following stochastic equicontinuity holds:*

$$\sup_{\alpha\in\mathcal{B}_n(\alpha_0),t\in\mathcal{T}_n} \mu_n\{R(Z;\alpha,\alpha_0) - R(Z;\alpha(t),\alpha_0)\} = o_{P_Z}(n^{-1}).$$

Define the set of efficient scores in the sieve space as

$$\mathcal{D}_{k(n)}^{eff} \equiv \left\{d(\cdot) = u_n^*(\cdot,\alpha_0,\lambda) : \alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A), \lambda \in U^{d_\phi}\right\} \tag{4.5}$$

This set of efficient scores, though well defined, has no explicit closed-form solution for partially identified semiparametric models. In the following remark, we provide a link between this set of efficient scores, and the set of efficient scores one would get in a point identified model. This would give an intuition of what is involved.

Finally, let $\mathcal{D}^{eff}$ denote the set of all limit points in $L^2(p_0\mu)$ of sequences of functions in $\mathcal{D}_{k(n)}^{eff}$, as $k(n) \to \infty$. The last assumption which is the most substantial, requires that this set of efficient scores be Donsker.

Assumption 4.3 *Let the following hold: (i) $\mu_n\{u_n^*(z,\alpha_0,\lambda)\} = O_{P_Z}(n^{-1/2})$ uniformly in $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A), \lambda \in U^{d_\phi}$; (ii) $\mathcal{D}^{eff}$ is Donsker in $L^2(p_0\mu)$ and has a $p_0\mu$-square interable envelope function F.*

Under assumption 4.3(ii), $\mathcal{D}^{eff}$ is a compact subset of the unit sphere of $L^2(p_0\mu)$. Note that the set of efficient scores in (4.5) is defined as α_0 ranges over the *constrained null*. So, this Donsker assumption is more reasonable as the set of efficient scores under the null is not as large.

Remark 4.1 *Although the set $\mathcal{D}_{k(n)}^{eff}$ and its $L^2(p_0\mu)-$limit set $\mathcal{D}^{eff}$ are well-defined, they have no closed-form expressions for partially identified semiparametric models in general. If $(\mathcal{A}_I^r, ||\cdot||_A) = \{\alpha_0\}$ (a singleton), then we could use Fisher norm $||v|| = \sqrt{E_0\left(\{\frac{d\ell(Z;\alpha_0)}{d\alpha}[v]\}^2\right)}$ instead of the (rescaled Hellinger or Pearson) distance $D(,)$ to compute a Riesz representer on the sieve space, which leads to an alternative yet equivalent expression for $\mathcal{D}_{k(n)}^{eff}$. In particular,*

$$||v_n^*(\alpha_0,\lambda)||^2 = \sup_{\alpha\in\mathcal{A}_n:||\alpha-\alpha_{0n}^D||\neq 0} \frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha-\alpha_{0n}^D]\right|^2}{E_0\left(\frac{d\ell(Z;\alpha_0)}{d\alpha}[\alpha-\alpha_{0n}^D]\right)^2},$$

and

$$\mathcal{D}_{k(n)}^{eff} = \left\{ d(\cdot) = \frac{\frac{d\ell(\cdot;\alpha_0)}{d\alpha}[v_n^*(\alpha_0,\lambda)]}{||v_n^*(\alpha_0,\lambda)||} : \lambda\in U^{d_\phi} \right\}.$$

For example, when $\phi(\alpha)=\theta$ and $\phi(\alpha_0)=\theta_0=r_0$ for $\{\theta_0\}=\Theta_I\subset int\,(\Theta)$, we have,

$$\begin{aligned} ||v_n^*(\alpha_0,\lambda)||^2 &= \sup_{\alpha\in\mathcal{A}_n:||\alpha-\alpha_{0n}^D||\neq 0} \frac{|\lambda'(\theta-\theta_0)|^2}{E_0\left(\frac{d\ell(Z;\alpha_0)}{d\theta'}(\theta-\theta_0)+\frac{d\ell(Z;\alpha_0)}{dg}[g-g_{0n}^D]\right)^2} \\ &= \lambda'\left(\mathcal{I}_{k(n)}\right)^{-1}\lambda, \end{aligned}$$

where

$$\mathcal{I}_{k(n)} \equiv E_0\left(\left[\frac{d\ell(Z;\alpha_0)}{d\theta'}-\frac{d\ell(Z;\alpha_0)}{dg}[w_n^*(\alpha_0)]\right]'\left[\frac{d\ell(Z;\alpha_0)}{d\theta'}-\frac{d\ell(Z;\alpha_0)}{dg}[w_n^*(\alpha_0)]\right]\right),$$

$w_n^*(\alpha_0)=\left(w_{n1}^*(\alpha_0),...,w_{nd_\theta}^*(\alpha_0)\right)$*, and for each $j=1,...,d_\theta$, $w_{nj}^*(\alpha_0)$ solves*

$$\inf_{w_j\in\mathcal{G}_{k(n)}} E_0\left[\left(\frac{d\ell(Z;\alpha_0)}{d\theta_j}-\frac{d\ell(Z;\alpha_0)}{dg}[w_j]\right)^2\right].$$

Then

$$\begin{aligned} \frac{d\ell(z;\alpha_0)}{d\alpha}[v_n^*(\alpha_0,\lambda)] &= \frac{d\ell(z;\alpha_0)}{d\theta'}v_{n,\theta}^*(\alpha_0,\lambda)+\frac{d\ell(z;\alpha_0)}{dg}[v_{n,g}^*(\alpha_0,\lambda)] \\ &= \left[\frac{d\ell(z;\alpha_0)}{d\theta'}-\frac{d\ell(z;\alpha_0)}{dg}[w_n^*(\alpha_0)]\right]v_{n,\theta}^*(\alpha_0,\lambda) \\ &= \left[\frac{d\ell(z;\alpha_0)}{d\theta'}-\frac{d\ell(z;\alpha_0)}{dg}[w_n^*(\alpha_0)]\right]\left(\mathcal{I}_{k(n)}\right)^{-1}\lambda, \end{aligned}$$

and

$$\mathcal{D}_{k(n)}^{eff} = \left\{ d(\cdot) = \frac{\left[\frac{d\ell(\cdot;\alpha_0)}{d\theta'}-\frac{d\ell(\cdot;\alpha_0)}{dg}[w_n^*(\alpha_0)]\right]\left(\mathcal{I}_{k(n)}\right)^{-1}\lambda}{\sqrt{\lambda'\left(\mathcal{I}_{k(n)}\right)^{-1}\lambda}} : \lambda\in U^{d_\phi} \right\}.$$

The following Theorem is the main result in this section.

Theorem 4.1 *Suppose that all the assumptions of Theorem 3.2 hold with* $\eta_n = O(\lambda_n) = o(n^{-1})$. *In addition, let assumptions 4.1 and 4.3 hold. For any* $r_0 \in \Re^{d_\phi}$, *let* $\varnothing \neq \mathcal{A}_I^r \equiv \{\alpha \in \mathcal{A}_I : \phi(\alpha) = r_0\} \subseteq \mathcal{A}$. *Then:*

(1) If assumption 4.3(i) holds, then: under the null hypothesis of $\alpha_0 \in \mathcal{A}_I^r$,

$$\begin{aligned} LR(r_0) &\equiv 2\left[\sup_{\alpha\in\mathcal{A}_{k(n)}}\sum_{i=1}^{n}\log p(Z_i;\alpha) - \sup_{\alpha\in\mathcal{A}_{k(n)}:\phi(\alpha)=r_0}\sum_{i=1}^{n}\log p(Z_i;\alpha)\right] \\ &= \sup_{d\in\mathcal{D}_{k(n)}^{eff}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}d(Z_i)\right)^2 + o_{P_Z}(1). \end{aligned}$$

(2) If assumption 4.3(ii) holds, then: under the null hypothesis of $\alpha_0 \in \mathcal{A}_I^r$,

$$LR(r_0) \Rightarrow \sup_{d\in\mathcal{D}^{eff}}(W(d))^2 \text{ in distribution,}$$

where $\{W(d) : d \in \mathcal{D}^{eff}\}$ *is a tight centered Gaussian process with covariance function* $\Gamma(d_1, d_2) = E_0[d_1 d_2]$ *defined on* $\mathcal{D}^{eff} \times \mathcal{D}^{eff}$.

Remark 4.2 *When* $\phi(\alpha) = \theta$ *and* $\phi(\alpha_0) = \theta_0 = r_0$, *Theorem 4.1 immediately yields a limiting distribution for the sieve profile log-LR statistic* $LR(\theta_0)$ *defined in (4.1): Under the null hypothesis of* $\theta_0 \in \Theta_I \cap int(\Theta)$ *and some regularity conditions, we have:*

$$LR(\theta_0) = \sup_{d\in\mathcal{D}_{k(n)}^{eff}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}d(Z_i)\right)^2 + o_{P_Z}(1) \Rightarrow \sup_{d\in\mathcal{D}^{eff}}(W(d))^2 \text{ in distribution.}$$

This result extends the profile likelihood ratio statistic result of Murphy and Van der Vaart (2000) to partially identified semiparametric models. It also extends Theorem 3.1 of Liu and Shao (2003) to allow for unknown nuisance functions belonging to non-compact parameter spaces.

Remark 4.3 *If the restriction* $\phi(\alpha) = r_0$ *point identifies the parameter, i.e., if* $\{\alpha_0\} = (\mathcal{A}_I^r, ||\cdot||_A)$ *(a singleton), and assumption 4.3 is automatically satisfied and* $\sup_{d\in\mathcal{D}^{eff}}(W(d))^2$ *will reduce to the usual Chi-squared distribution with degree of freedom* d_ϕ. *Previously, under point identification (i.e.,* $\{\alpha_0\} = (\mathcal{A}_I, ||\cdot||_A)$*), the chi-squared distribution result has been derived by Shen and Shi (2005) for the case of regular* $\phi()$ *(i.e., root-n estimable) and by Chen and Liao (2009) for the case of irregular* $\phi()$ *(i.e., slower than root-n estimable).*

When $(\mathcal{A}_I^r, ||\cdot||_A)$ is not a singleton, the limiting distribution in Theorem 4.1 no longer has a simple closed-form expression. The next section provides a computationally attractive approach to inference based on a weighted bootstrap approximation to this complicated asymptotic null distribution under partial identification.

5 Weighted bootstrap

Heuristically, the approach we take to consistently estimate the asymptotic distribution of the sieve LR statistic is as follows. We generate n-size samples of positive "weights" from a known distribution with mean 1, and for each of these samples, we compute *the weighted likelihood*. We compute the value of LR statistic for each of these weighted likelihoods, and we show that the empirical distribution of that sample of the weighted likelihood ratio values consistently estimation the distribution of the (unweighted) LR statistic. This consistency holds whether or not the parameter is on the boundary, the problem is ill-posed, or rates are non standard. We first provide assumptions on the weights.

Assumption 5.1 *(i)* $\{\omega_i\}_{i=1}^n$ *is a positive, i.i.d. sequence drawn from the distribution of a positive random variable* ω *with* $E[\omega] = 1$, $Var[\omega] = \sigma_\omega^2 \in [0, \infty)$ *and* $\|\omega\|_{2,1} \equiv \int_0^\infty \sqrt{\Pr(|\omega| > t)}dt < \infty$*; (ii)* $\{\omega_i\}_{i=1}^n$ *is independent of* $\{Z_i\}_{i=1}^n$.

We assume that the bootstrap weights $\{\omega_i\}_{i=1}^n$ defined on $(\mathcal{W}, \Omega, P_W)$. For the joint randomness involved, the product probability space is defined as

$$(\mathcal{Z}^\infty, \mathcal{A}^\infty, P_Z) \times (\mathcal{W}, \Omega, P_W) = (\mathcal{Z}^\infty \times \mathcal{W}, \mathcal{A}^\infty \times \Omega, P_{ZW}).$$

Since the bootstrap weights $\{\omega_i\}_{i=1}^n$ is independent of the data $\{Z_i\}_{i=1}^n$, we have $P_{ZW} = P_Z \times P_W$.

We only need assumption 5.1 in bootstrap consistency theorem 5.1.

Theorem 5.1 *Suppose that all the assumptions of Theorem 4.1 hold and assumption 5.1 holds. Let* $\widehat{p} \equiv p(\cdot; \widehat{\alpha}) = \arg\max_{\alpha \in \mathcal{A}_{k(n)}} \sum_{i=1}^n \log p(Z_i; \alpha)$ *be the sieve MLE and* $\phi(\widehat{\alpha}) = \widehat{r}$*. Then: conditional on the data* $\{Z_i\}_{i=1}^n$ *satisfying the null hypothesis of* $\alpha_0 \in \mathcal{A}_I^r$*,*

$$\begin{aligned} LR^\omega(\widehat{r}) &\equiv 2\left[\sup_{\alpha \in \mathcal{A}_{k(n)}} \sum_{i=1}^n \omega_i \log p(Z_i; \alpha) - \sup_{\alpha \in \mathcal{A}_{k(n)}: \phi(\alpha) = \widehat{r}} \sum_{i=1}^n \omega_i \log p(Z_i; \alpha)\right] \\ &= \sup_{d \in \mathcal{D}_{k(n)}^{eff}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (\omega_i - 1) d(Z_i)\right)^2 + o_{P_{ZW}}(1) \\ &\Rightarrow \sigma_\omega^2 \times \sup_{d \in \mathcal{D}^{eff}} (W(d))^2 \text{ in distribution,} \end{aligned}$$

where $\{W(d) : d \in \mathcal{D}^{eff}\}$ *is a tight centered Gaussian process with covariance function* $\Gamma(d_1, d_2) = E_0[d_1 d_2]$.

The following result directly follows from Theorem 5.1.

Corollary 5.1 *Suppose that all the assumptions of Theorem 4.1 hold and assumption 5.1 holds. Let* $\widehat{p} \equiv p(\cdot;\widehat{\theta},\widehat{g}) = \arg\max_{\theta\in\Theta, g\in\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \log p(Z_i;\theta,g)$ *be the sieve MLE. Then: conditional on the data* $\{Z_i\}_{i=1}^n$ *and* $\theta_0 \in \Theta_I \cap int\,(\Theta)$,

$$\begin{aligned} LR^{\omega}(\widehat{\theta}) &\equiv 2\left[\sup_{\theta\in\Theta, g\in\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \omega_i \log p(Z_i;\theta,g) - \sup_{g\in\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \omega_i \log p(Z_i;\widehat{\theta},g)\right] \\ &\Rightarrow \sigma_\omega^2 \times \sup_{d\in\mathcal{D}^{eff}} (W(d))^2 \ \textit{in distribution.} \end{aligned}$$

We could apply Theorem 5.1 to construct confidence sets for $\theta_0 \in \Theta_I$. Recall that the sieve profile log-likelihood ratio statistic for θ is

$$LR(\theta) \equiv 2\left[\sup_{\theta'\in\Theta, g\in\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \log p(Z_i;\theta',g) - \sup_{g\in\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \log p(Z_i;\theta,g)\right].$$

Our confidence set $\mathcal{C}_n$ is constructed by inverting the log likelihood ratio statistic:

$$\mathcal{C}_n = \{\theta \in \Theta : \quad LR(\theta) \leq \widehat{c}_n(\theta, 1-\tau)\}$$

where $\widehat{c}_n(\theta, 1-\tau)$ is the $(1-\tau)$ quantile using the weighted bootstrap with a weight such that $\sigma_\omega^2 = 1$:

$$\widehat{c}_n(\theta, 1-\tau) = \inf\left\{x : \frac{1}{B_n}\sum_{j=1}^{B_n} I\{LR_j^{\omega}(\widehat{\theta}) \leq x\} \geq 1-\tau\right\},$$

where B_n is the number of bootstrap replications, $LR_j^{\omega}(\widehat{\theta})$ is the j-th bootstrapped version of the weighted likelihood ratio statistic $LR^{\omega}(\widehat{\theta})$ defined in Theorem 5.1.

6 Monte Carlo

To examine the finite sample behavior of our inferential procedures, we conduct a series of Monte Carlo experiments. We consider two different versions of the binary dynamic discrete choice model discussed in Example 2 in Section 2 above. The general model we consider is

$$y_{it} = 1\{x_{it}'\beta + y_{i,t-1}\gamma + u_i + \varepsilon_{it} \geq 0\} \tag{6.1}$$

where we have $u_i \sim N(0, \sigma_0^2)$, with $\sigma_0 = 1$ and $\gamma_0 = .8$,. The random variable ϵ_{it} is standard normal independent of the regressors at all time periods. Also, we set the initial condition distribution $g(u_i, x_i^T) \equiv P(y_{i1} = 1|u_i, x_i^T) \equiv \frac{1}{2}$.

To build the sieve MLE, the sieve space $\mathcal{G}_k$ was taken to be the space of all Bernstein polynomials of degree $k = 4$.[8] We added penalties on the L_2-norms of g and its derivative

and took $\lambda = .1$. Numerical integration of the integral in the sample log likelihood $L_n(\theta, g) = \sum_{i=1}^{n} \log p(y_i^T | x_i^T, \theta, g, F_u)$ where

$$\begin{aligned} & p(y_i^T | x_i^T, \theta, g, F_u) \\ = & \int \{g(u, x_i^T)\}^{y_{i1}} \{1 - g(u, x_i^T)\}^{1-y_{i1}} \prod_{t=2}^{T} P(y_{it} | x_i^T, y_{it-1}; \theta, u) dF_u(u) \end{aligned}$$

where

$$P\left(y_{it} = 0 \middle| x_i^T, y_{it-1}; \theta, u\right) = \Phi(-x_{i,t}'\beta - \gamma y_{i,t-1} - u).$$

was performed using Halton sequences of length 40. Bootstrap weights ω_i were generated as independent draws from an exp(1) distribution, however we obtained very similar results using other distributions, including simple two-point distributions such as $P[\omega_i = 1 - a] = 1/2 = P[\omega_i = 1 + a]$ for various choices of $a \in (0, 1)$. All simulations were performed with $B_n = 500$ bootstrap replications and 500 Monte Carlo repetitions.[9]

We first report results for the case where $\beta = 0$. So, the observed data are a vector of binary choices of size T.

Table 1 shows the actual sizes for individual and joint confidence regions of (γ, σ), defined as the proportion of Monte Carlo repetitions in which the confidence region contained γ_0, σ_0 or (γ_0, σ_0) (Marginal Confidence Regions in the Table). The results show that the actual sizes are quite accurate even in small samples. This accuracy is not affected by whether the identified set is large ($T = 3$) or small ($T = 4$). For the sake of comparison, k and λ were kept fixed across different values of sample size n, contrary to the prescription of our theory. This is manifest in the bias that starts to appear for $n = 800, 1600$ and is completely consistent with our results. The filled contour plot, Figure 1, shows the coverage function for the joint confidence region of (γ, σ) overlaid on the identified set, which is delimited by the solid white line. In this diagram, the color at any (γ, σ) on the graph corresponds, via the key on the right, to the actual size of the confidence interval at that point. Additionally, for

[8] See Chen (2007) for the definitions of all sieve spaces used in this paper. The j^{th} Bernstein polynomial of degree k is given by

$$B_j^k(x) = \binom{k}{j} x^j (1 - x)^{k-j}$$

and is defined for $x \in [0, 1]$ and $j = 0, 1, \ldots, k$. (To build a basis for a function defined outside of $[0, 1]$, one can scale the evaluation points to lie in $[0, 1]$.) A helpful property of the Bernstein polynomials is that if $g = \sum_{j=0}^{k} \phi_j B_j^k$, then $g \in [0, 1]$ if and only if $\phi \in [0, 1]^{k+1}$. This property means that simple lower and upper bounds in maximizing over ϕ will guarantee that g is a proper probability over its entire domain.

[9] The program was written in the AMPL modeling language and optimizations were solved using KNITRO. Each Monte Carlo replication takes approximately 1, 3 or 12 minutes on a 2.4Ghz Intel Core 2 Quad for $T = 2, 3$ or 4, respectively.

Table 1: MC Results for Dynamic Discrete Choice with No Regressors

γ		T=3		
level	n=1600	n=800	n=400	n=200
0.500	0.604	0.532	0.500	0.532
0.900	0.954	0.928	0.882	0.898
0.950	0.986	0.966	0.942	0.948
0.990	0.994	0.990	0.992	0.986

γ		T = 4		
level	n=1600	n=800	n=400	n=200
0.500	0.560	0.548	0.468	0.536
0.900	0.940	0.926	0.906	0.878
0.950	0.956	0.960	0.950	0.942
0.990	0.996	0.994	0.990	0.980

σ		T=3		
level	n=1600	n=800	n=400	n=200
0.500	0.534	0.530	0.500	0.504
0.900	0.904	0.904	0.908	0.900
0.950	0.950	0.958	0.944	0.938
0.990	0.988	0.990	0.988	0.986

σ		T = 4		
level	n=1600	n=800	n=400	n=200
0.500	0.504	0.500	0.480	0.518
0.900	0.882	0.908	0.868	0.902
0.950	0.938	0.964	0.936	0.946
0.990	0.980	0.994	0.978	0.990

Joint Confidence Regions

		T=3		
level	n=1600	n=800	n=400	n=200
0.500	0.592	0.544	0.494	0.502
0.900	0.960	0.928	0.894	0.898
0.950	0.972	0.954	0.956	0.942
0.990	0.994	0.982	0.992	0.990

		T = 4		
level	n=1600	n=800	n=400	n=200
0.500	0.550	0.526	0.494	0.524
0.900	0.908	0.932	0.900	0.886
0.950	0.960	0.964	0.946	0.940
0.990	0.990	0.996	0.976	0.984

Table 2: MC Results for Dynamic Discrete Choice with Regressors: $n = 200, T = 3$

	Marginal CI		Joint CI
level	β	γ	(β, γ)
0.500	.442	.464	.462
0.900	.866	.87	.87
0.950	.934	.93	.942
0.990	.984	.982	.986

any given value of n, the results were fairly insensitive to different values of λ and k. Figure 1 shows the coverage functions (one less the power function) obtained from our Monte Carlo experiment with $T = 3$ and $n = 200, 400, 800$ and 1600.

We also investigated a similar dynamic binary response model with a single continuous, time-invariant, covariate $x_i \sim N(0, 1)$, independently of u_i and ϵ_{it} with coefficient $\beta_0 = 1$ in (6.1) above. This model is especially interesting because of the presence of a continuous regressor. Due to computational concerns, in this case we assumed that the distribution of u_i was known to be discrete with equal probability on support points $.2, .4, .6, .8$.[10] The actual sizes when $n = 200$ and $T = 3$ are shown in Table 2 and the one and two dimensional coverage functions are shown in Figure 3.

As we can see from above for the design we have, the Monte Carlo experiments show adequate small sample performance and more importantly a reasonable computational burden.

7 Empirical Applications

We applied our methods to three interesting economic applications. In the first,we study a duration model with unobserved heterogeneity as in Heckman and Singer. There, it is not clear whether the structural parameters are point identified, and even if they are, the identification is at infinity and the parameter is estimated slower than root-n rate. We provide confidence region using our approach and contrast ours to the ones commonly implemented in empirical work assuming the estimated parameter is root-n asymptotic normal. In the second example, we examine a model of intergenerational schooling where a child's schooling level is explained by the parents'. The regression suffers from a censoring problem since we do not observe the full schooling of some of the children due to the fact that those children

[10]These issues arise from the computational difficulties involved with numerical integration. As our empirical examples show, our method is applicable to much more computationally complex models than we investigated in our Monte Carlo studies. However, in Monte Carlo simulations, where the model must be estimated several hundred times, simpler models are more manageable.

Figure 1: Confidence Regions for γ, σ

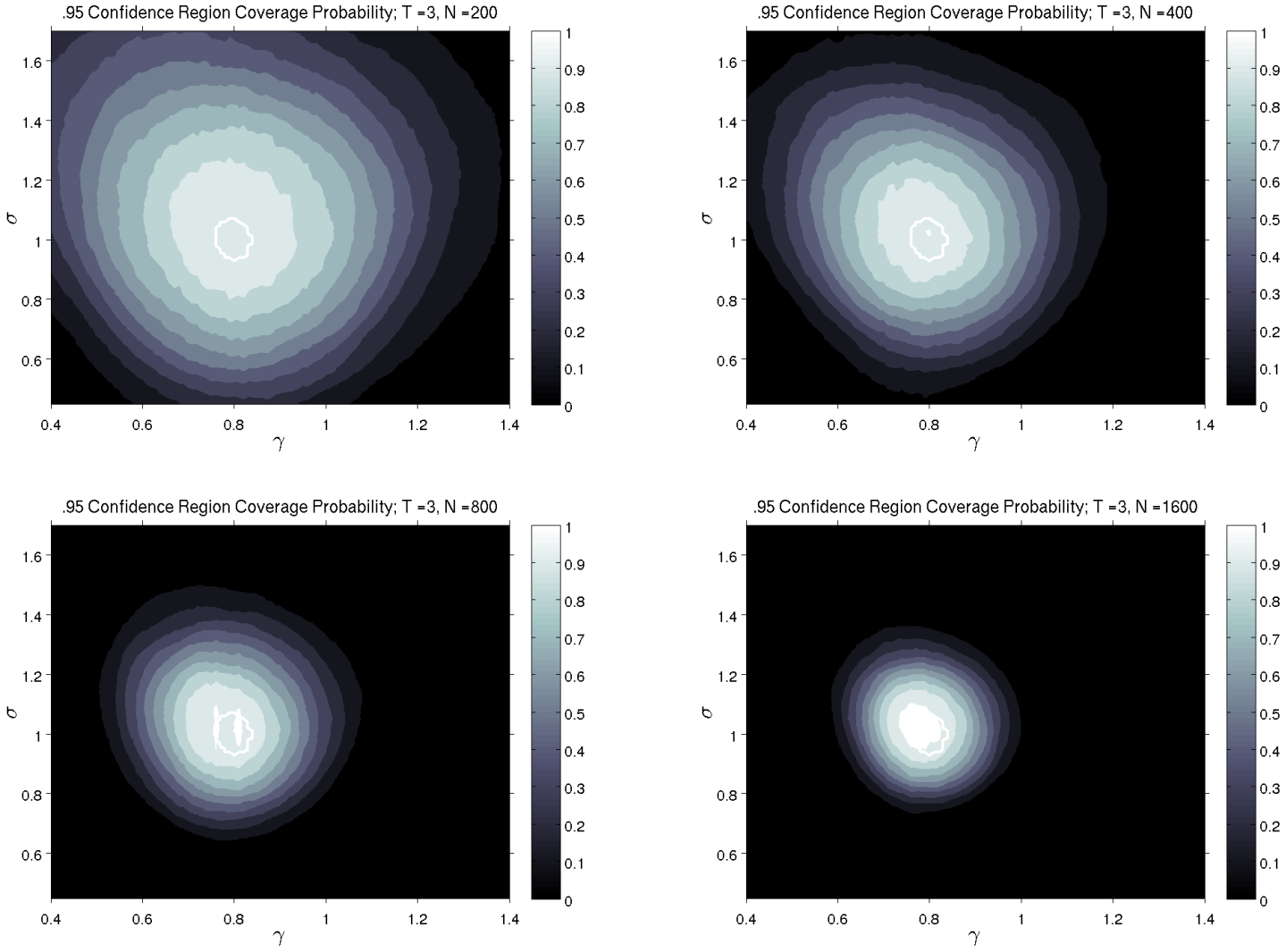


Figure 2: Coverage Functions for γ

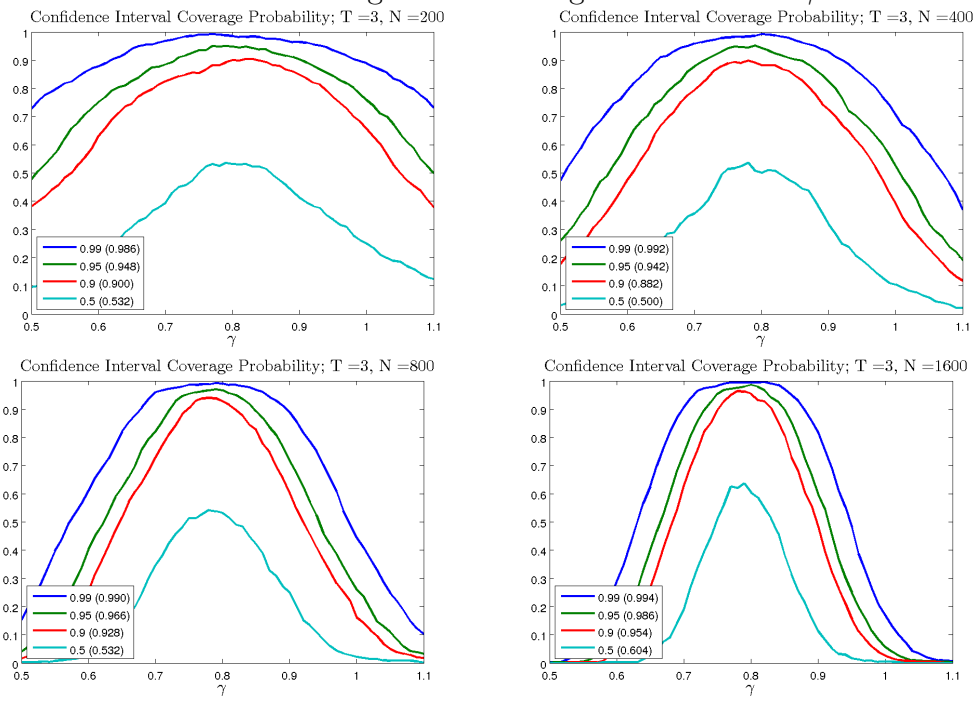
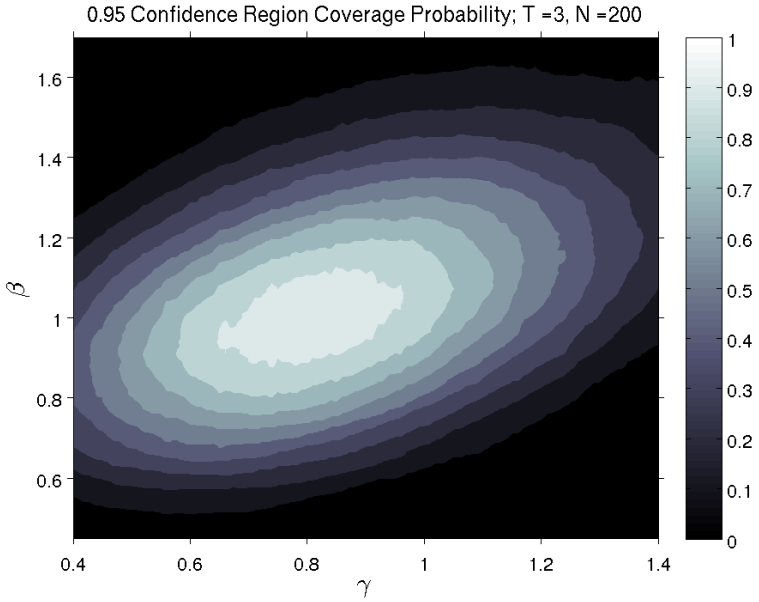


Figure 3: Confidence Regions for γ, β for Model (6.1)



were still enrolled in school at the time of survey. This is an example of linear regression with interval data on the outcome.

7.1 Length of Unemployment Spells

We estimated a duration model of the type described in Example 1 of Section 2 with a Weibull hazard function,

$$\log \lambda(t|x,u) = \alpha' x + \beta \log t + u,$$

where t is time until becoming employed, X are covariates, u is an unobservable, assumed to be independent of X, and $\beta > -1$ is a shape parameter for the Weibull distribution. The observable density of spell lengths in is then

$$p(t|x;\theta,g) = \int_u t^\beta \exp\left(\alpha'^{\beta+1}\left(\frac{e^{\alpha' x+u}}{\beta+1}\right)\right) dg(u),$$

where $\theta = (\alpha, \beta)$ and $g()$ is a parameter corresponding to the distribution of unobserved heterogeneity. Theoretically, it is rare that one has any information about the heterogeneity distribution, and so the empirical literature has used various distributions $g(.)$ based on computational ease and familiarity. Heckman and Singer (1984) in important work showed that under certain support conditions, the parameter of interest θ is point identified in the limit without making assumptions on g, and provided a non-parametric likelihood type estimator for it. This identification of θ in this model (with unknown g) is fragile because point identification is reached in the limit as durations approach zero. This is an example of a model that is "identified at infinity" (See Ishwaran (1996)). More generally, it is a semiparametric mixture model with a structural parameter and in these models, it is difficult to provide sufficient point identification conditions. Regardless, our inference procedure is consistent whether or not θ is point identified, or whether or not in case of point identification, it can be estimated at a regular $\sqrt{n}$ rate of convergence.

Our sample is drawn from the 1979 youth cohort National Longitudinal Survey of Labor Market Experience. A detailed description of this data set can be found for example in Keane and Wolpin (1997). We follow 1,119 young men who were 16 years old in 1977 and appear in the survey every year between 1980 and 1986, inclusive, restricting attention only to those who were unemployed with 12 years of schooling or less in 1980 and who did not re-enter school during the sample. For covariates we use an indicator for black, years of schooling and an indicator for having completed high school (12 years of schooling). Approximately half of the sample remained unemployed for the entire time period and are thus considered to be censored in our construction of the likelihood. In the left hand side of Table **??**, we provide first confidence regions for the parameters constructed marginally using

our bootstrap procedure. This top half uses the semiparametric Heckman/Singer like model with a nonparametric $g(.)$ while the bottom half uses uses a standard normal distribution for the heterogeneity distribution and is there for comparison. We see that the estimates are close but with important differences. Compare for example the effect of education. Both the coefficients on education and high school seem to be higher in magnitude in the normal model, while Figure 4 shows the joint confidence regions for pairs of parameters. One can see that upper and lower endpoints of these confidence regions change and are different than the marginal cases as to be expected. The shape of these confidence regions reflect first the shape of the identified set. Also, these confidence regions reflect a slight departure from asymptotic normality (in case we think the model parameters are point identified). These confidence regions again are robust to both departures from normality and regular rates, and to failure of point identification. In the figure in Table 3 below, we plot the "estimated" density of $g(.)$ using the estimated sieve coefficients. It shows departures from normality[11] As we can see in this example, using a normal density for the unobserved heterogeneity provides a slightly biased estimates in this sample.

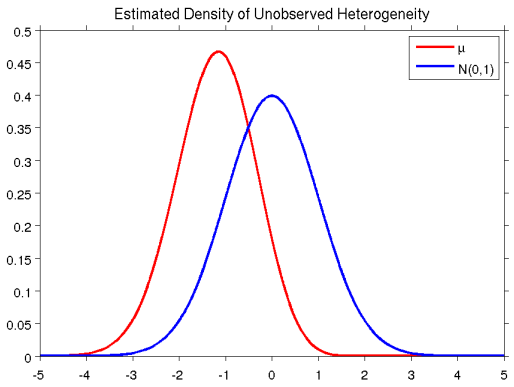
Table 3: Sensitivity with respect to heterogeneity distribution

Nonparametric g

Parameter	Marginal Confidence Region
β	$[-.0488, .0798]$
Black	$[-.83, -.429]$
Education	$[-.0813, -.0369]$
High School	$[.868, 1.352]$

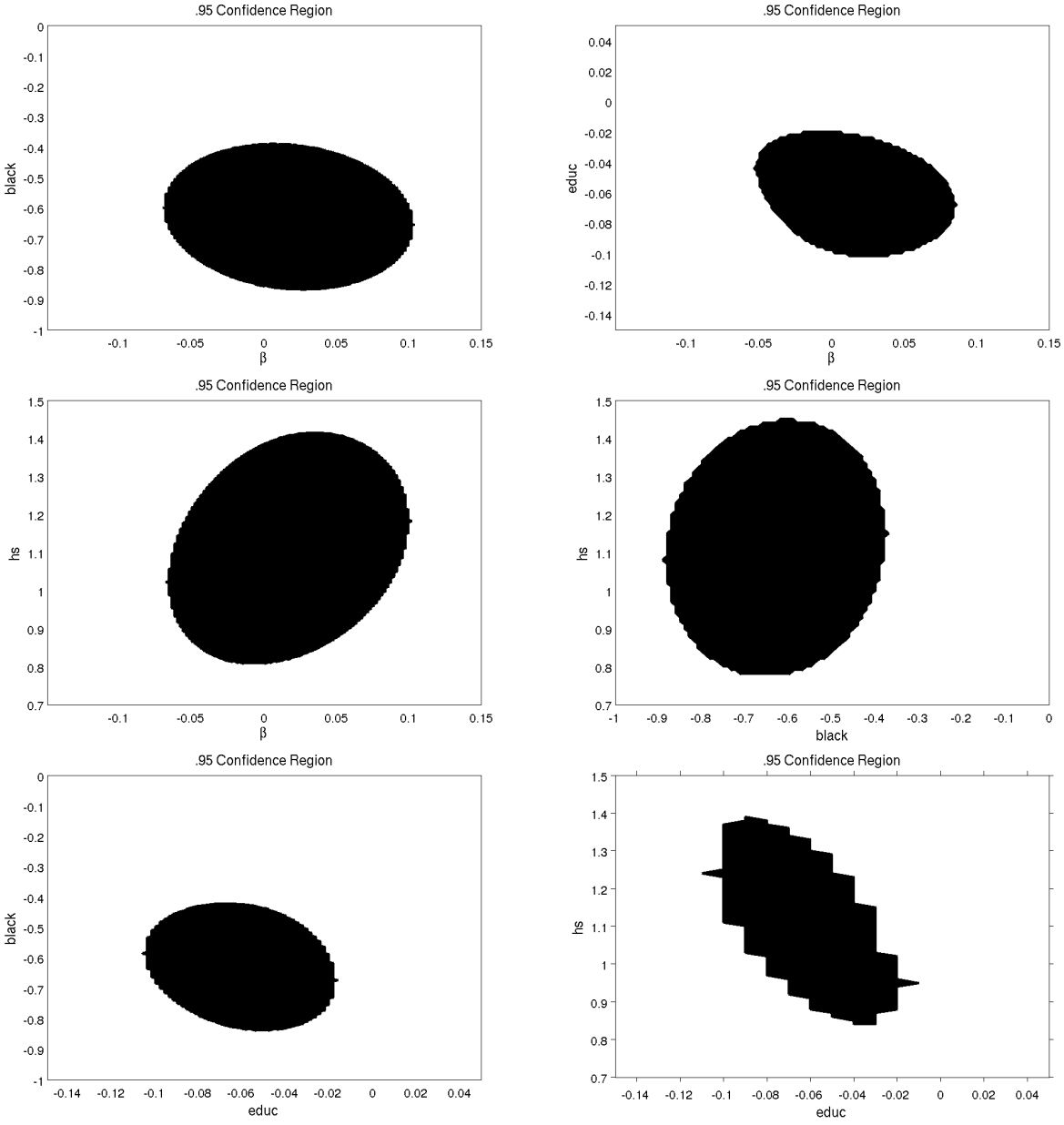
Standard Normal g

Parameter	Marginal Confidence Region
β	$[-.0035, .135]$
Black	$[-.88, -.465]$
Education	$[-.20, -.167]$
High School	$[1.205, 1.704]$



[11]It is not clear whether the density of $g(.)$ is identified. Moreover, this paper does not contain a theory for inference on infinite dimensional parameters that are not point identified. So, a comparison of the estimate of density of $g()$ constructed using estimated sieve coefficients is an approximation at best.

Figure 4: Joint Confidence Regions



7.2 Intergenerational Schooling Mobility with Discretized Samples

Censoring occurs frequently in applied work and often causes practitioners to adopt *ad hoc* functional forms especially in cases when the dependent variable is interval measured. One such situation arises when trying to determine the impact of a parent's schooling level on that of their children. But, if at the time of the survey, some of the children are still in school, schooling for those children will belong to a predefined interval (is censored). Given that schooling is discrete, then, for the censored observations, we observe that these belong to a bin, as in *higher than highschool*, or *more than a college*, etc. So, the dependent variable when it is censored can take finitely many values. So, this is an example where the likelihood for the econometrics model depends on the specification of the error distribution. An interesting data set that we use is the most recent wave of the Wisconsin Longitudinal Study (WLS) which contains data on the completed levels of schooling for adults who graduated from high school in 1957 and their children for the years 1975, 1992 and 2004. In the 1992 data, children's schooling level is censored for some observations, whereas in the 2004 data it is not since the survey updated the censored observations. De Haan and Plug (2011) and Plug (2004) exploit this unique feature of the WLS to evaluate the effectiveness of various approaches for censored data by comparing these approaches applied to the 1992 data to the estimates from the uncensored 2004 data. Our motivation is similar to that of De Haan and Plug (2011) in that we want to evaluate the effectiveness of our semiparametric approach in solving this censoring problem by comparing its performance to the results obtained using the uncensored data.

Consider a simple linear regression of the form

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

where X_i is parent's completed level of schooling and Y_i is their child's level of schooling which is either censored or uncensored depending on the sample. We also controlled for age and gender. Here, since when Y is unobserved, we know that it must belong to a finite number of ranges, so the likelihood of the observed data can be written easily as a function of the distribution of ϵ. The Tobit model assumes that $X_i \perp \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$. The model we estimate is a semiparametric generalization of this,

$$Y_i = \alpha + \beta X_i + \sigma_c \sigma(X_i)\epsilon_i = \alpha + \beta X_i + u_i,$$

where $\epsilon_i | X_i$ is distributed according to some unknown G that has mean zero, variance 1 and density g. Here σ denotes an unknown heteroskedasticity function and σ_c is a common scalar

Table 4: Sensitivity in Censored Model of Schooling

Nonparametric $G(.)$ and $\sigma(.)$

Parameter	Marginal Confidence Region
β	$[.292, .453]$
Age	$[-.1975, -.058]$
Gender	$[-.12, .185]$
Intercept	$[11.82, 14.71]$

Uncensored

Parameter	Marginal Confidence Region
β	$[.324, .372]$
Age	$[-.099, -.067]$
Gender	$[-.24, .226]$
Intercept	$[11.35, 12.56]$

Comparing CI's for β under Various Assumptions

Model	β
Tobit	$[.366, .421]$
Normal and Heteroskedastic	$[.368, .426]$
NonNormal and Homoskedastic	$[.267, .382]$

scale parameter (which is unnecessary but we use it for ease of computations). It follows that $u_i|X_i \sim \sigma_c\sigma(X_i)\epsilon_i$, so that $E(u_i|X_i) = 0$ and $Var(u_i|x_i) = \sigma_c^2\sigma(X_i)^2$. Letting $d_i = 1$ if the observation is censored and 0 otherwise and writing $\theta = (\alpha, \beta, \sigma_c)$, the sample log-likelihood for this model is

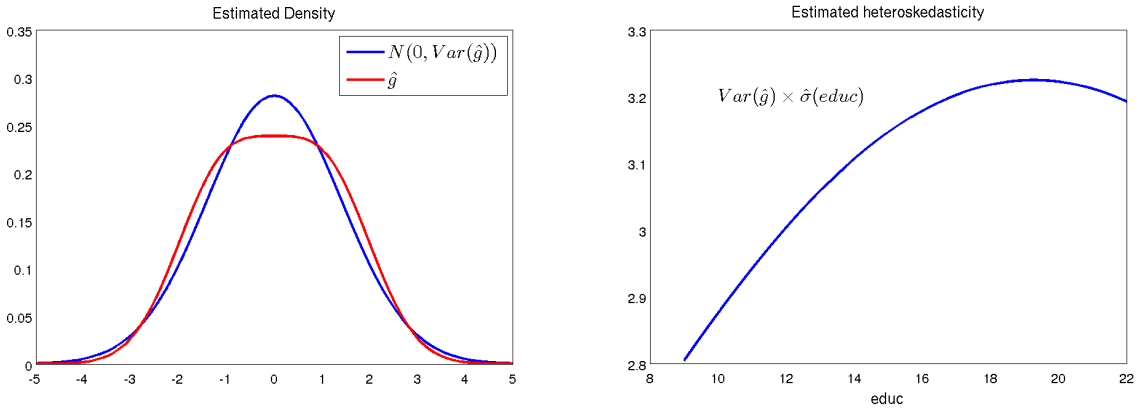
$$L_n(\theta, \sigma, g) = \sum_{i=1}^{n}(1 - d_i)\log\left(\frac{1}{\sigma_c\sigma(X_i)}g\left(\frac{Y_i - \alpha - \beta X_i}{\sigma_c\sigma(X_i)}\right)\right) + \sum_{i=1}^{n} d_i \log\left(1 - G\left(\frac{Y_i - \alpha - \beta X_i}{\sigma_c\sigma(X_i)}\right)\right).$$

We approximated σ by a second degree polynomial spline with a knot at $X_i = 12$ years (completed high school) and g by a fifth degree Hermite polynomial constrained to be a proper density with zero mean and variance one. Penalties were added on the L_2-norms of σ, σ' and g with $\lambda_\sigma = 1$ and $\lambda_g = .01$. We took X_i to be father's schooling and added a small amount of mean zero measurement error to Y_i to make it continuous.[12] We also added controls for age and gender. The program was written in AMPL, used SNOPT as the solver and took about six hours to run.

[12]Note that the fully parametric Tobit model applied to unadjusted schooling data is vulnerable to the same objection.

The results are presented in Table 4. The estimates from the uncensored data should be viewed as the benchmark to which the other models are compared. The uncensored model is a linear regression and so is consistent under arbitrary heteroskedasticity, while our model restricts the model to multiplicative heteroskedasticity. The results in Table 4 for our estimator reports "marginal" confidence regions for every parameter, i.e., say for β, a marginal CI is defined as the CI constructed while we profile out the rest of the parameters (along with the unknown function). As we can see, the results are close across regressors. The Tobit model without heteroskedasticity shows some bias. While this is corrected somewhat by the Tobit model with unknown heteroskedasticity, some bias remains. When we relax the normality assumption and continue to allow for heterogeneity, the resulting confidence interval for schooling contains that from the uncensored model. These results show that while specific functional forms on unobservables introduce misspecification bias, our semiparametric estimator does not, albeit at the cost of wider confidence intervals. In addition, and for the semiparametric results, we obtained confidence regions as we varied the parameters for the Sieve and the penalizations and found that varying these tuning parameters did not have a noticeable impact[13]. The conclusion here is that in this sample, the estimates of β are not very sensitive to specifying a distribution for the errors. Finally, in Figure 5, we plot the estimated distribution $\hat{G}$ and also the heteroskedasticity function $\hat{\sigma}(.)$ using the estimated sieve parameters.

Figure 5: Estimated density and heteroskedasticity function in Example 4.2



7.3 Entry in the Airline Industry

Berry (1992) examined several models of entry in the airline industry with a general profit

[13]The results for these runs can be obtained from the authors upon request.

function of the form

$$\pi_{ik}(N) = c + X_i\beta - \delta \log N + Z_{ik}\alpha + \rho u_{i0} + \sqrt{1-\rho^2}u_{ik}, \tag{7.1}$$

where $\pi_{ik}(N)$ is the profit of the k^{th} firm in the i^{th} market when there are N other firms present, X_i are observed market-specific covariates, Z_{ik} are observed firm-market covariates, $u_{i0} \sim N(0,1)$ is market-specific unobserved heterogeneity and u_{ik} is unobserved firm-specific heterogeneity. Berry (1992) proved that under this profit function all pure strategy Nash equilibria involve a unique number of firms N_i^*, which is assumed to be what is observed. We consider a special case of (7.1), also considered by (Berry 1992), where there is no unobserved firm heterogeneity, i.e. $\rho = 1$, and hence the model reduces to an ordered probit. While Berry (1992) assumed homoskedasticity across markets, we will allow for heteroskedasticity to enter in an unknown way so that our profit equation becomes

$$\pi_{ik}(N) = c + X_i\beta - \delta \log N + Z_{ik}\alpha + \sigma(X_i)u_{i0}, \tag{7.2}$$

where σ is an unknown function that is positive and bounded away from zero. Although with enough variation in the regressor the above model might be point identified, it is not clear that with the data we have, the model parameters are point identified when we allow for heteroskedasticity of unknown forms.

To start, let $\phi_{ik} = X_i\beta + Z_{ik}\alpha$ and order the firms from most to least profitable so that $\phi_{i1} > \phi_{i2} > \cdots > \phi_{iK_i}$. The probability that N or more firms enter in market i is the probability that the N_i^{th} most profitable firm entered, or

$$Prob[N_i^* \geq N] = Prob\left[\pi_{iN_i}(N) > 0\right] = \Phi\left[\frac{\delta \log N - c - \phi_{iN_i}}{\sigma(X_i)}\right].$$

It is then straightforward to construct the corresponding likelihood function of $\theta = (c, \alpha, \beta, \delta)$ and σ by differencing these probabilities.

The data come from the Origin and Destination Survey of Air Passenger Traffic for the first and third quarters of 2001. Markets $i = 1, \ldots, N = 1028$ are defined as city-city pairs of routes and k indexes firm identities.[14] Like Berry (1992), we model the entry decision as a 6 month (two quarter) process and we define the number of potential entrants for market i in the third quarter, P_i, as all firms who served a route originating or ending in one or both of the cities corresponding to pair i in the first quarter and we use distance between markets (X_i) and market presence (Z_{ik}) as covariates.

Let $\mathcal{G}_k$ be the collection of all second degree polynomial splines with a knot at the median of X_i. In practice we take $g_k \in \mathcal{G}_k$ and let $\sigma_k = \exp(g_k)$, which ensures that

[14]These firm identities are American Airlines, Continental, Delta, Northwest, USAir, medium airline and low-cost airline.

$\sigma_k > 0$. We also place an L_2 penalty on the norm of g_k and its derivative and set $\lambda = .01$, although the results were fairly insensitive to this choice. The 500 bootstrap draws are taken from an exp(1) distribution. The program was written in AMPL and the likelihood was optimized with SNOPT. Estimation takes about 2 hours with the same processor used for the simulations. The marginal confidence regions are presented in Table 5 and compared alongside the parametric case where $\sigma(X_i) \equiv \rho$. Also, joint confidence regions are presented in Figure 6.

There, we also present results as a function of the various tuning parameters that we chose. Overall, these suggest that the homoskedastic model considerably understates the impact on profits of the number of firms in the market relative to distance, but is approximately accurate on the relative import of number of firms to market presence (compare the CI for δ between our model and the parametric homoskedastic MLE). This is important since the parameter δ measures the relative impact of competition (having an extra entrant). Overall, though, and looking across the estimates, in these data, the impact of the various tuning parameters seems minimal[15].

8 Conclusion

Empirical economic models are built upon a set of assumptions that define the model. Some of these assumptions are motivated by economic theory such as optimizing behavior but other assumptions are made solely for the purpose of "closing the model", or to obtain a complete econometric structure and are motivated by simplicity, familiarity and ease of computation. On the one hand, these assumptions allow economists to use standard methods for inference that rely on simple computational procedures to obtain estimates of the key parameters. On the other hand, these estimates suffer from the serious criticism that they are sensitive to the ad-hoc assumptions made. A response to this criticism is to weaken these extraneous assumptions. But often times this weakening leads to partial identification of the parameter of interest and (sufficient) conditions for point identification, when available, rely on support conditions that are hard to satisfy in typical data.

The loss of point identification can have serious consequences on the way one conducts inference since standard asymptotic distribution theory results derived under point identification are no longer valid. We fill an important gap here by examining the question of inference in likelihood models in the presence of unknown nuisance functions, allowing for the parameter of interest to be partially identified or irregular even if point identified. A

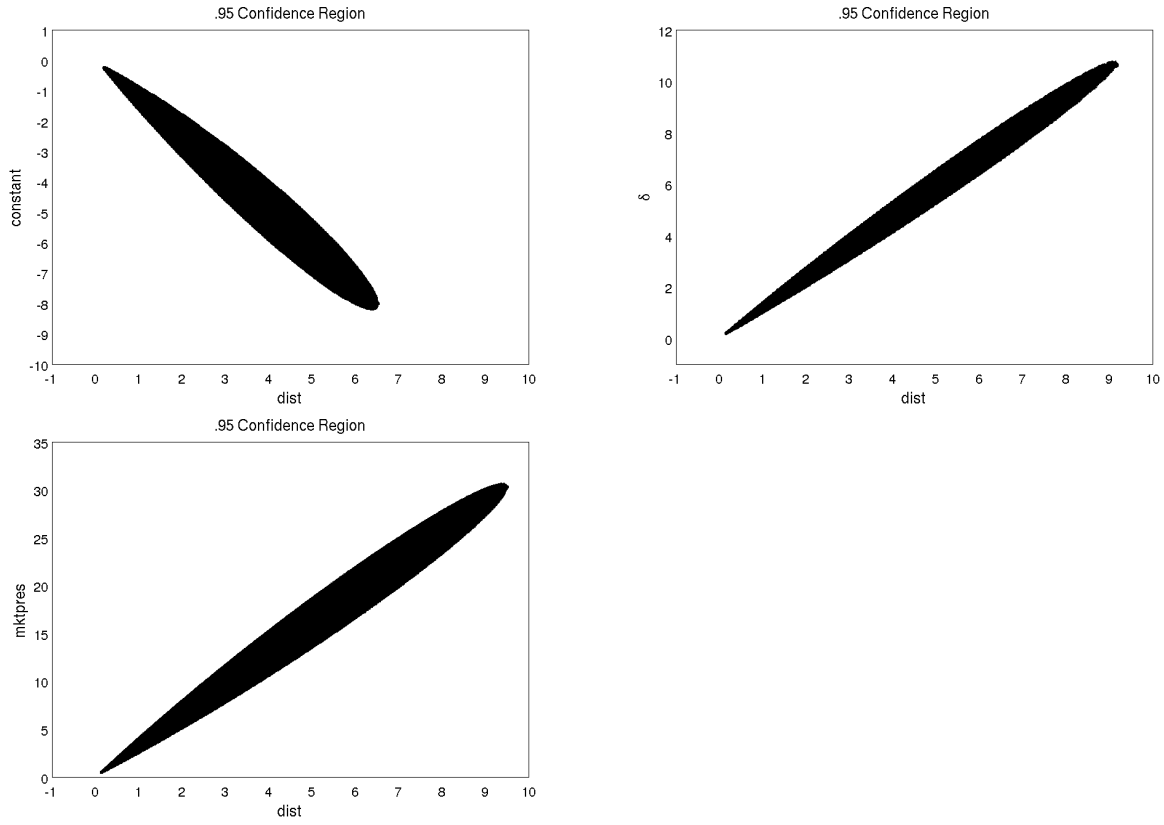
[15]We ran more specifications with more tuning parameters. We are only reporting a smaller, but representative set.

Table 5: Sensitivity in Entry Model to Fixed Cost Distribution

	Semiparametric Results (marginal CI)[a].							Homoskedastic
	2, 1, 10^{-4}	2, 1, 10^{-3}	2, 1, 10^{-2}	2, 1, 10^{-4}	2,2, 10^{-4}	2,3, 10^{-4}	3,3, 10^{-2}	MLE
Int	$[-1.707, -.792]$	$[-1.691, -.805]$	$[-1.621, -.857]$	$[-1.59, .89]$	$[-1.78, -.78]$	$[-1.87, -.77]$	$[-1.95, -.73]$	$[-1.62, -.98]$
δ	$[1.005, 1.511]$	$[1.007, 1.515]$	$[1.013, 1.492]$	$[.98, 1.311]$	$[.97, 1.538]$	$[.94, 1.52]$	$[.97, 1.56]$	$[.92, 1.12]$
MktPres	$[2.291, 3.920]$	$[2.23, 3.902]$	$[2.38, 3.78]$	$[2.54, 3.76]$	$[2.23, 4.1]$	$[2.22, 4.38]$	$[2.16, 4.15]$	$[2.70, 3.75]$
Dist	$[.811, 1.163]$	$[.815, 1.153]$	$[.82, 1.11]$	$[.848, 1.096]$	$[.78, 1.24]$	$[.78, 1.31]$	$[.71, 1.31]$	$[.835, 1.108]$

[a]The parameters in the third row correspond to various combinations of tuning parameters: (2,1, 10^{-4}) signifies a second degree polynomial, with 1 knot, and the penalty parameter is set to 10^{-4}

Figure 6: Confidence Regions for the Berry Entry Model



semiparametric likelihood approach is attractive since it automatically leads to efficiency and chi-square inference when the parameter of interest happens to be point identified. Our weighted bootstrap procedure is very easy to implement and performs well in finite sample Monte Carlo studies.

Although our semiparametric approach is more robust, our model might still be misspecified. In this case, the identified set can be interpreted as the *pseudo true identified set* and represents the set of parameters that minimize the KL distance between the data distribution and the model implied distribution. More generally, the issue of misspecification in partially identified models is delicate, especially in terms of interpreting the identified set, and so we leave that for future research. Finally, our theoretical results in this paper, specifically Theorem 4.1, hold at a fixed distribution P_0, i.e. pointwise. It is not clear that this limit distribution holds uniformly over all P_0 in some space of implied distributions. This might be relevant in some cases, such as models where $\phi(.)$ or θ_0 lie on the boundary of the parameter space. Though our asymptotic distribution still holds pointwise, it might suffer from uniformity issues. This is difficult even in parametric likelihood models such as Liu and Shao's (See Andrews and Cheng (2010)). We view this as an important problem that we leave for future work.

References

Andrews, D., and X. Cheng (2010): "Estimation and Inference with Weak, Semi-Strong, and Strong Identification," Cowles Foundation, Yale University.

Andrews, D., and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78(1), 119–157.

Berry, S. (1992): "Estimation of a model of entry in the airline industry," *Econometrica*, 60(4), 889–917.

Birgé, L., and P. Massart (1998): "Minimum contrast estimators on sieves: exponential bounds and rates of convergence," *Bernoulli*, pp. 329–375.

Bugni, F. (2010): "Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set," *Econometrica*, 78(2), 735–753.

Canay, I. (2010): "EL inference for partially identified models: large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156(2), 408–425.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chen, X., and Z. Liao (2009): "On Limiting Distributions of Sieve M-Estimators of Irregular Functionals," Working Paper.

Chen, X., and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152(1), 46–60.

——— (2011): "Estimation of nonparametric conditional moment models with possibly nonsmooth moments," *forthcoming, Econometrica*.

Chen, X., D. Pouzo, and E. Tamer (2011): "QLR Inference on Partially Identified Nonparametric Conditional Moment Models," Working Paper.

Chernoff, H. (1954): "On the distribution of the likelihood ratio," *The Annals of Mathematical Statistics*, pp. 573–578.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models1," *Econometrica*, 75(5), 1243–1284.

Chernozhukov, V., S. Lee, and A. Rosen (2009): "Interesection Bounds: Estimation and Inference," working paper, MIT.

De Haan, M., and E. Plug (2011): "Estimating intergenerational schooling mobility on censored samples: consequences and remedies," *Journal of Applied Econometrics*.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications*, vol. 66. Chapman & Hall/CRC.

Grieco, P. (2011): "Discrete Games with Flexible Information Structures: An Application to Local Grocery Markets," Penn State Working Paper.

Heckman, J., and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*.

Honoré, B., and E. Tamer (2006): "Bounds on parameters in panel dynamic discrete choice models," *Econometrica*, 74(3), 611–629.

Imbens, G., and C. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72(6), 1845–1857.

Ishwaran, H. (1996): "Uniform rates of estimation in the semiparametric Weibull mixture model," *The Annals of Statistics*, pp. 1572–1585.

Keane, M., and K. Wolpin (1997): "The career decisions of young men," *Journal of Political Economy*, 105(3), 473–522.

Leamer, E. (1985): "Sensitivity analyses would help," *The American Economic Review*, 75(3), 308–313.

Leamer, E. (1987): "Error in Variables in Linear Systems," *Econometrica*, 55(4), 893–809.

Liu, X., and Y. Shao (2003): "Asymptotics for likelihood ratio tests under loss of identifiability," *Annals of Statistics*, pp. 807–832.

Manski, C. (1995): *Identification Problems in the Social Sciences*. Harvard University Press.

Murphy, S., and A. Van der Vaart (2000): "On profile likelihood," *Journal of the American Statistical Association*, pp. 449–465.

Plug, E. (2004): "Estimating the effect of mother's schooling on children's schooling using a sample of adoptees," *The American Economic Review*, 94(1), 358–368.

Redner, R. (1981): "Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions," *The Annals of Statistics*, pp. 225–228.

Romano, J., and A. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.

Rosen, A. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146(1), 107–117.

Santos, A. (2011): "Inference in nonparametric instrumental variables with partial identification," forthcoming, Econometrica.

Shen, X., and J. Shi (2005): "Sieve Likelihood ratio inference on general parameter space," *Science in China Series A: Mathematics*, 48(1), 67–78.

Tamer, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2(1), 167–195.

Wong, W., and X. Shen (1995): "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs," *The Annals of Statistics*, pp. 339–362.

Zeidler, E. (1985): *Nonlinear Functional Analysis and its Applications III: Variational Methods and Optimization*. Springe-Verlag.

A Consistency of penalized sieve extremum estimation under partial identification

In this section, we provide a set consistency result for general extreme estimation problems where the parameter is defined in a general function space and where point identification is relaxed. The result stated in Theorem 3.1 is clearly a subset of the result below and hence we skip its proof.

Assumption A.1 *Let the followings hold:*

1. ***Parameter space and objective function:*** *(i)* $\mathcal{A} = \Theta \times \mathcal{G} \subseteq \mathbf{A} = \Re^{d_\theta} \times \mathbf{G}$, Θ *is a compact, nonempty subset of a Euclidean space* $(\Re^{d_\theta}, |\cdot|_e)$, *and* $\mathcal{G}$ *is a closed, bounded and nonempty subset of a separable infinite dimensional Banach space* $(\mathbf{G}, ||\cdot||_G)$*; (ii)* $Q : \mathcal{A} \to [0, \infty)$ *is lower semicontinuous on* $\mathcal{A}$ *under* $||\alpha||_A = |\theta|_e + ||g||_G$*; (iii) the identified set,* $\mathcal{A}_I = \Theta_I \times \mathcal{G}_I = \{\alpha \in \mathcal{A} : Q(\alpha) = 0\}$*, is a nonempty, closed, bounded strict subset of* $\mathcal{A}$ *under* $||\cdot||_A$.

2. ***Sieve space*** *(i) for each* $k \geq 1$, $\mathcal{A}_k = \Theta \times \mathcal{G}_k \subseteq \mathcal{A}$, $\mathcal{G}_k$ *is closed under* $||\cdot||_G$ *with* $\dim(\mathcal{G}_k) < \infty$*; (ii)* $\varnothing \neq \mathcal{G}_k \subseteq \mathcal{G}_{k+1} \subseteq \mathcal{G}$ *for all* $k \geq 1$, *and* $\overline{\cup_{k=1}^{\infty} \mathcal{G}_k}$ *is dense in* $\mathcal{G}$ *under* $||\cdot||_G$. *That is, for any* $g \in \mathcal{G}$, *there is* $\Pi_k g \in \mathcal{G}_k$ *such that* $||g - \Pi_k g||_G \to 0$ *as* $k \to \infty$.

3. ***Penalty function*** *There is a function* $Pen : \mathcal{G} \to [0, \infty)$ *such that: (i)* $Pen(.)$ *is a measurable function such that* $\sup_{g \in \mathcal{G}_I} Pen(g) < \infty$*; (ii) the set* $\{g \in \mathcal{G} : Pen(g) \leq M\}$ *is compact under* $||\cdot||_G$ *for all* $M \in [0, \infty)$*; (iii)* $\lambda_n > 0$, *and* $\lambda_n \sup_{g \in \mathcal{G}_I} |Pen(\Pi_n g) - Pen(g)| = O(\lambda_n) = o(1)$.

4. ***Uniform convergence on sieve space***
$\max\left\{\sup_{\alpha \in \mathcal{A}_{k(n)}} |Q_n(\alpha) - Q(\alpha)|, \sup_{(\theta,g) \in \mathcal{A}_I} Q(\theta, \Pi_n g)\right\} = O_{p_0}(\lambda_n) = o_{P_0}(1)$.

Theorem A.1 *Let* $\widehat{\mathcal{A}}_n$ *be the collection of* $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{g}_n) \in \mathcal{A}_{k(n)} = \Theta \times \mathcal{G}_{k(n)}$ *that solves*

$$Q_n(\widehat{\alpha}_n) + \lambda_n Pen(\widehat{g}_n) = \inf_{\alpha \in \mathcal{A}_{k(n)}} \left[Q_n(\alpha) + \lambda_n Pen(g)\right].$$

Let Assumption A.1 hold. Then:

$$d_A(\widehat{\alpha}_n, \mathcal{A}_I) \equiv \inf_{\alpha \in \mathcal{A}_I} ||\widehat{\alpha}_n - \alpha||_A = o_{p_0}(1),$$

and $Pen(\widehat{g}_n) = O_{p_0}(1)$.

PROOF OF THEOREM A.1: Under assumption A.1, $\widehat{\mathcal{A}}_{k(n)}$ is non-empty, compact under $||\cdot||_A$ for any given data. Take any $\widehat{\alpha}_n \in \widehat{\mathcal{A}}_{k(n)}$, it is well-defined and measurable. In the following we denote $\hat{c}_n^Q \equiv \sup_{\alpha\in\mathcal{A}_{k(n)}} |Q_n(\alpha) - Q(\alpha)|$ and $\Pi_n\alpha \equiv (\theta, \Pi_n g)$. For any $\varepsilon > 0$,

$$
\begin{aligned}
& \Pr\left(d_A(\widehat{\alpha}_n, \mathcal{A}_I) > \varepsilon\right) \\
\leq\ & \Pr\left(\inf_{\alpha\in\mathcal{A}_{k(n)}:d_A(\alpha,\mathcal{A}_I)\geq\varepsilon} [Q_n(\alpha) + \lambda_n Pen(g)] \leq \sup_{\alpha\in\mathcal{A}_I} [Q_n(\Pi_n\alpha) + \lambda_n Pen(\Pi_n g)]\right) \\
\leq\ & \Pr\left(\begin{array}{c} \inf_{\alpha\in\mathcal{A}_{k(n)}:d_A(\alpha,\mathcal{A}_I)\geq\varepsilon} \{Q(\alpha) + \lambda_n Pen(g) - |Q_n(\alpha) - Q(\alpha)|\} \\ \leq \sup_{\alpha\in\mathcal{A}_I} [Q(\Pi_n\alpha) + \lambda_n Pen(\Pi_n g) + |Q_n(\Pi_n\alpha) - Q(\Pi_n\alpha)|] \end{array}\right) \\
\leq\ & \Pr\left(\inf_{\alpha\in\mathcal{A}_{k(n)}:d_A(\alpha,\mathcal{A}_I)\geq\varepsilon} \{Q(\alpha) + \lambda_n Pen(g)\} \leq \sup_{\alpha\in\mathcal{A}_I} [Q(\Pi_n\alpha) + \lambda_n Pen(\Pi_n g)] + 2\hat{c}_n^Q\right) \\
\leq\ & \Pr\left(\begin{array}{c} \inf_{\alpha\in\mathcal{A}_{k(n)}:d_A(\alpha,\mathcal{A}_I)\geq\varepsilon} \{Q(\alpha) + \lambda_n Pen(g)\} \\ \leq \sup_{\alpha\in\mathcal{A}_I}\{Q(\Pi_n\alpha)\} + \lambda_n \sup_{g\in\mathcal{G}_I} Pen(g) + 2\hat{c}_n^Q \end{array}\right) \\
\leq\ & \Pr\left(\inf_{\alpha\in\mathcal{A}_{k(n)}:d_A(\alpha,\mathcal{A}_I)\geq\varepsilon} \{Q(\alpha) + \lambda_n Pen(g)\} \leq \Delta_n\right) \quad \text{where } \Delta_n = O(\lambda_n).
\end{aligned}
$$

We divide $\mathcal{A}_{k(n)}(\varepsilon) \equiv \{\alpha \in \mathcal{A}_{k(n)} : d_A(\alpha, \mathcal{A}_I) \geq \varepsilon\}$ into two disjoint sets: $\mathcal{A}^+_{k(n)}(\varepsilon) \equiv \{\alpha \in \mathcal{A}_{k(n)}(\varepsilon) : Pen(g) \leq 2\lambda_n^{-1}\Delta_n + M\}$ for any $M > 0$, and $\mathcal{A}^-_{k(n)}(\varepsilon) \equiv \mathcal{A}_{k(n)}(\varepsilon) \setminus \mathcal{A}^+_{k(n)}(\varepsilon)$. Note that

$$\inf_{\alpha\in\mathcal{A}^-_{k(n)}(\varepsilon)} \{Q(\alpha) + \lambda_n Pen(g)\} \geq 2\Delta_n + \lambda_n M > \Delta_n$$

Thus $\Pr\left(\inf_{\alpha\in\mathcal{A}^-_{k(n)}(\varepsilon)} \{Q(\alpha) + \lambda_n Pen(g)\} \leq \Delta_n\right) = 0$; hence

$$
\begin{aligned}
\Pr\left(d_A(\widehat{\alpha}_n, \mathcal{A}_I) > \varepsilon\right) &\leq \Pr\left(\inf_{\alpha\in\mathcal{A}^+_{k(n)}(\varepsilon)} \{Q(\alpha) + \lambda_n Pen(g)\} \leq \Delta_n\right) \\
&\leq \Pr\left(\inf_{\alpha\in\mathcal{A}^+(\varepsilon)} \{Q(\alpha) + \lambda_n Pen(g)\} \leq \Delta_n\right),
\end{aligned}
$$

where $\mathcal{A}^+(\varepsilon) \equiv \left\{\alpha \in \mathcal{A} : d_A(\alpha, \mathcal{A}_I) \geq \varepsilon, Pen(g) \leq 2\lambda_n^{-1}\Delta_n + M\right\}$.
Given that assumption A.1.3(ii), the fact that $\{\alpha \in \mathcal{A} : d_A(\alpha, \mathcal{A}_I) \geq \varepsilon\}$ is closed, and that $\Delta_n = O(\lambda_n)$, we have that the set $\mathcal{A}^+(\varepsilon)$ is compact under $||\cdot||_A$. Moreover, $Q(\alpha)$ is lower semicontinuous on $\mathcal{A}$ under $||\cdot||_A$ (assumption A.1.1(ii)). Theorem 38.B of Zeidler (1985) now implies that the minimization problem,

$$\inf_{\alpha\in\mathcal{A}^+_{k(n)}(\varepsilon)} \{Q(\alpha) + \lambda_n Pen(g)\}$$

has a solution, α_n, which belongs to the set $\mathcal{A}^+(\varepsilon)$. Therefore, the sequence $\{\alpha_n\}$ must have a further subsequence, denoted as $\{\alpha_{n_k}\}$, that converges to a limit α_∞ in $||\cdot||_A$ and $\alpha_\infty \in \{\alpha \in \mathcal{A} : d_A(\alpha, \mathcal{A}_I) \geq \varepsilon, Pen(g) \leq \overline{M}\}$ for some $\overline{M} \in [0, +\infty)$. By assumption

A.1.1(ii)(iii) and $Pen(g) \geq 0$, we have:

$$0 \leq Q(\alpha_\infty) \leq \liminf_n \{Q(\alpha_n)\} \leq \liminf_n \Delta_n = 0.$$

This implies that $\alpha_\infty \in \mathcal{A}_I$, which contradicts $\alpha_\infty \in \{\alpha \in \mathcal{A} : d_A(\alpha, \mathcal{A}_I) \geq \varepsilon, Pen(g) \leq \overline{M}\}$. Thus $d_A(\widehat{\alpha}_n, \mathcal{A}_I) = o_P(1)$. Next, by definition, there is a $\alpha^* \in \mathcal{A}_I$ such that

$$\begin{aligned} 0 &\leq \lambda_n Pen(\widehat{g}_n) \leq Q_n(\Pi_n \alpha^*) + \lambda_n Pen(\Pi_n g^*) \\ &\leq Q_n(\Pi_n \alpha^*) - Q(\Pi_n \alpha^*) + \lambda_n [Pen(\Pi_n g^*) - Pen(g^*)] + Q(\Pi_n \alpha^*) + \lambda_n Pen(g^*) \\ &\leq \sup_{\alpha \in \mathcal{A}_{k(n)}} |Q_n(\alpha) - Q(\alpha)| + \lambda_n \sup_{g \in \mathcal{G}_I} |Pen(\Pi_n g) - Pen(g)| + \sup_{\alpha \in \mathcal{A}_I} Q(\Pi_n \alpha) + \lambda_n \sup_{g \in \mathcal{G}_I} Pen(g) = O(1) \end{aligned}$$

thus $Pen(\widehat{g}_n) = O_{p_0}(1)$. *Q.E.D.*

B Proof of Theorem 4.1

Denote $\ell(Z, \alpha) \equiv \log p(Z, \alpha)$, $\chi(\alpha, \alpha_0) \equiv \chi(p(\cdot, \alpha), p_0)$ and

$$s(z; \alpha) = s_\chi(z; \alpha) \equiv \frac{\frac{p(z,\alpha)}{p(z,\alpha_0)} - 1}{\chi(\alpha, \alpha_0)};\ E_0\left[s(Z; \alpha)\right] = 0;\ E_0\left[(s(Z; \alpha))^2\right] = 1.$$

For all $\alpha \in \mathcal{B}(\alpha_0)$, using the fact $\log(1 + u) = u - 0.5u^2[1 - \text{rem}(u)]$ with $\text{rem}(u) \to 0$ as $u \to 0$, we have, with $u = \chi(\alpha, \alpha_0)s(Z; \alpha)$,

$$\begin{aligned} &\ell(Z, \alpha) - \ell(Z, \alpha_0) = \log\left(1 + \left[\frac{p(Z, \alpha)}{p(Z, \alpha_0)} - 1\right]\right) = \log\left[1 + \chi(\alpha, \alpha_0)s(Z; \alpha)\right] = \log(1 + u) \\ = \ &\chi(\alpha, \alpha_0)s(Z; \alpha) - \frac{1}{2}[\chi(\alpha, \alpha_0)s(Z; \alpha)]^2 \left\{1 - \text{rem}(\chi(\alpha, \alpha_0)s(Z; \alpha))\right\} \end{aligned}$$

Then, we have

$$\begin{aligned} &E_0[\ell(Z_i, \alpha) - \ell(Z_i, \alpha_0)] \\ = \ &-K(\alpha_0, \alpha) = -\frac{1}{2}[\chi(\alpha, \alpha_0)]^2 + \frac{1}{2}[\chi(\alpha, \alpha_0)]^2 E_0[s(Z; \alpha)^2 \text{rem}(\chi(\alpha, \alpha_0)s(Z; \alpha))]. \end{aligned}$$

By Remark 3.3, $K(\alpha_0, \alpha) = \frac{1}{2}[\chi(\alpha, \alpha_0)]^2(1 + o(1))$ uniformly in α when $\chi(\alpha, \alpha_0)$ is small, we have:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \alpha) - \ell(Z_i, \alpha_0)] \\ = \ &-K(\alpha_0, \alpha) + \chi(\alpha, \alpha_0)\mu_n\left\{s(Z; \alpha)\right\} - \frac{1}{2}[\chi(\alpha, \alpha_0)]^2 \mu_n\left\{s(Z; \alpha)^2[1 - \text{rem}(\chi(\alpha, \alpha_0)s(Z; \alpha))]\right\} \\ = \ &-\frac{1}{2}[\chi(\alpha, \alpha_0)]^2(1 + o(1)) + \chi(\alpha, \alpha_0)\mu_n\left\{s(Z; \alpha)\right\} + \mu_n\left\{R(Z; \alpha, \alpha_0)\right\} \end{aligned}$$

where

$$R(Z;\alpha,\alpha_0) = -\frac{1}{2}[\chi(\alpha,\alpha_0)s(Z;\alpha)]^2\left\{1-\text{rem}(\chi(\alpha,\alpha_0)s(Z;\alpha))\right\}.$$

Recall that

$$u_n^*(z,\alpha_0,\lambda) \equiv \frac{v_n^*(z,\alpha_0,\lambda)}{||v_n^*(\alpha_0,\lambda)||} = \frac{v_n^*(z,\alpha_0,\lambda)}{\sqrt{Var_0[v_n^*(Z,\alpha_0,\lambda)]}}.$$

Then

$$E_0\left[u_n^*(z,\alpha_0,\lambda)\right] = 0;\ E_0\left[(u_n^*(z,\alpha_0,\lambda))^2\right] = 1.$$

We now consider perturbation in probability density sieve space: for all $\alpha \in \mathcal{B}_n(\alpha_0)$ and $t_n \in \mathcal{T}_n \equiv \{t \in [-1,1] : |t| \leq const. \times n^{-1/2}\}$,

$$p(z,\alpha(t_n)) = p(z,\alpha) + t_n u_n^*(z,\alpha_0,\lambda)p_0(z).$$

Thus

$$\begin{aligned}
& [\chi(\alpha(t_n),\alpha_0)]^2 - [\chi(\alpha,\alpha_0)]^2 \\
= \ & E_0\left[\left(\frac{p(Z,\alpha(t_n))}{p_0(Z)}-1\right)^2\right] - E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)}-1\right)^2\right] \\
= \ & E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)}-1+t_n u_n^*(Z,\alpha_0,\lambda)\right)^2\right] - E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)}-1\right)^2\right] \\
= \ & 2t_n E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)}-1\right)u_n^*(Z,\alpha_0,\lambda)\right] + t_n^2.
\end{aligned}$$

$$\begin{aligned}
& \mu_n\left\{\chi(\alpha(t_n),\alpha_0)s(Z;\alpha(t_n))\right\} - \mu_n\left\{\chi(\alpha,\alpha_0)s(Z;\alpha)\right\} \\
= \ & \mu_n\left\{\frac{p(Z,\alpha(t_n))}{p(Z,\alpha_0)} - \frac{p(Z,\alpha)}{p(Z,\alpha_0)}\right\} \\
= \ & t_n \times \mu_n\left\{u_n^*(Z,\alpha_0,\lambda)\right\}.
\end{aligned}$$

Therefore, uniformly over $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$, $\lambda \in U^{d_\phi}$, $\alpha \in \mathcal{B}_n(\alpha_0)$ and $t_n \in \mathcal{T}_n$, we have:

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n [\ell(Z_i,\alpha(t_n)) - \ell(Z_i,\alpha)] \\
= \ & \frac{1}{n}\sum_{i=1}^n [\ell(Z_i,\alpha(t_n)) - \ell(Z_i,\alpha_0)] - \frac{1}{n}\sum_{i=1}^n [\ell(Z_i,\alpha) - \ell(Z_i,\alpha_0)] \\
= \ & -\frac{[\chi(\alpha(t_n),\alpha_0)]^2 - [\chi(\alpha,\alpha_0)]^2}{2}(1+o(1)) + \mu_n\left\{\chi(\alpha(t_n),\alpha_0)s(Z;\alpha(t_n)) - \chi(\alpha,\alpha_0)s(Z;\alpha)\right\} + o_{P_Z}(n^{-1}) \\
= \ & -\frac{2t_n E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)}-1\right)u_n^*(Z,\alpha_0,\lambda)\right] + t_n^2}{2}(1+o(1)) \\
& + t_n \times \mu_n\left\{u_n^*(Z,\alpha_0,\lambda)\right\} + o_{P_Z}(n^{-1}).
\end{aligned}$$

Theorem B.1 *Suppose that Assumptions 4.1 and 4.2 hold and that $\widehat{\alpha}_n$ is a sieve MLE. Then: (1) uniformly over $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$, $\lambda \in U^{d_\phi}$, $\alpha \in \mathcal{B}_n(\alpha_0)$ and $t_n \in \mathcal{T}_n$,*

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}[\ell(Z_i, \alpha(t_n)) - \ell(Z_i, \alpha)] \\ = \; & t_n\left(\mu_n\left\{u_n^*(Z, \alpha_0, \lambda)\right\} - E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right](1 + o(1))\right) \\ & -\frac{t_n^2}{2} + o_{P_Z}(n^{-1}); \end{aligned}$$

(2) Uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi}$,

$$\left|E_0\left[\left(\frac{p(z,\widehat{\alpha}_n)}{p_0(z)} - 1\right) u_n^*(z, \alpha_0, \lambda)\right] - \mu_n\left\{u_n^*(z, \alpha_0, \lambda)\right\}\right| = o_{P_Z}(n^{-\frac{1}{2}}).$$

Thus under assumption 4.3(i), we have: uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi}$,

$$\left|E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right]\right| = O_{P_Z}(n^{-\frac{1}{2}}).$$

(3) Uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi}$,

$$\sup_{\alpha \in \mathcal{B}_n(\alpha_0) \cap \{\alpha : \phi(\alpha) = r_0\}} \left|E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right]\right| = o_{P_Z}(n^{-\frac{1}{2}}).$$

PROOF OF THEOREM B.1: **Result (1)** is already proved before the statement of the theorem.

For **Result (2)**. Let $\varepsilon_n = o(n^{-\frac{1}{2}})$. For any $\alpha \in \mathcal{B}_n(\alpha_0) \subset \mathcal{A}_n$, consider a local perturbation

$$p(z, \alpha(\varepsilon_n)) = p(z, \alpha) \pm \varepsilon_n \times u_n^*(z, \alpha_0, \lambda) \times p_0(z) \in \mathcal{P}_{k(n)}.$$

By assumption 4.2: uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi} \equiv \{\lambda \in \Re^{d_\phi} : |\lambda|_e = 1\}$, and by the definition of $\widehat{\alpha}_n$, we have

$$\begin{aligned} -o_{P_Z}(n^{-1}) \; \leq \; & \frac{1}{n}\sum_{i=1}^{n}\ell(Z_i, \widehat{\alpha}_n) - \frac{1}{n}\sum_{i=1}^{n}\ell(Z_i, \widehat{\alpha}(\varepsilon_n)) \\ = \; & \frac{\pm 2\varepsilon_n E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right] + \varepsilon_n^2}{2}(1 + o(1)) \\ & \mp \varepsilon_n \times \mu_n\left\{u_n^*(Z, \alpha_0, \lambda)\right\} + o_{P_Z}(n^{-1}). \end{aligned}$$

By the definition of $u_n^*(\alpha_0, \lambda)$ and $\varepsilon_n = o_{P_Z}(n^{-\frac{1}{2}})$, we obtain

$$\left|E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right] - \mu_n\left\{u_n^*(Z, \alpha_0, \lambda)\right\}\right| = o_{P_Z}(n^{-\frac{1}{2}}).$$

Since

$$E_0\left[\left(\frac{p(Z,\alpha_{0n}^D)}{p_0(Z)}-1\right)u_n^*(Z,\alpha_0,\lambda)\right]=0,$$

we also have

$$E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n)-p(Z,\alpha_{0n}^D)}{p_0(Z)}\right)u_n^*(Z,\alpha_0,\lambda)\right]=\mu_n(u_n^*(Z,\alpha_0,\lambda))+o_{P_Z}(n^{-\frac{1}{2}})$$

holds uniformly in $\alpha_0\in(\mathcal{A}_I^r,||\cdot||_A)$ and $\lambda\in U^{d_\phi}$.

For **Result (3)**. Under Assumption 4.1.(ii), we have uniformly over $\alpha_0\in\mathcal{A}_I^r\equiv\{\alpha\in\mathcal{A}_I:\phi(\alpha)=r_0\}$:

$$\frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha_{0n}^D-\alpha_0]\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||}=o(n^{-\frac{1}{2}}). \tag{B.1}$$

For all $\alpha\in\mathcal{B}_n(\alpha_0)\cap\{\alpha:\phi(\alpha)=r_0\}$ and for all $\alpha_0\in\mathcal{A}_I^r\equiv\{\alpha\in\mathcal{A}_I:\phi(\alpha)=r_0\}$, under Assumption 4.1.(ii), we have, uniformly over $\alpha_0\in\mathcal{A}_I^r$:

$$\frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha-\alpha_0]\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||}=o(n^{-\frac{1}{2}})$$

$$\begin{aligned}
&\left|E_0\left[\left(\frac{p(Z,\alpha)}{p_0(Z)}-1\right)u_n^*(Z,\alpha_0,\lambda)\right]\right| \\
=&\left|E_0\left[\left(\frac{p(Z,\alpha)-p(Z,\alpha_{0n}^D)}{p_0(Z)}\right)u_n^*(Z,\alpha_0,\lambda)\right]\right| \\
=&\frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha-\alpha_{0n}^D]\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||} \\
\leq&\frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha-\alpha_0]\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||}+\frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha_{0n}^D-\alpha_0]\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||}=o_{P_Z}(n^{-\frac{1}{2}}).
\end{aligned}$$

Thus **Result (3)** is true. *Q.E.D*

Recall that the unconstraint sieve MLE $\widehat{\alpha}_n$ is

$$\widehat{\alpha}_n\in\widehat{\mathcal{A}}_n\equiv\arg\max_{\alpha\in\mathcal{A}_n}\left\{\sum_{i=1}^n\log p(Z_i;\alpha)-o_{P_Z}(1)\right\},$$

and the constraint sieve MLE $\widetilde{\alpha}_n$ is

$$\widetilde{\alpha}_n\in\widetilde{\mathcal{A}}_n\equiv\arg\max_{\{\alpha\in\mathcal{A}_n:\ \phi(\alpha)=r_0\}}\left\{\sum_{i=1}^n\log p(Z_i;\alpha)-o_{P_Z}(1)\right\}.$$

Theorem 4.1 immediately follows from the following Theorem B.2 and assumption 4.3(ii):

Theorem B.2 *Suppose that assumptions 4.1, 4.2 and 4.3(i) hold. Then under null of* $\alpha_0 \in \mathcal{A}_I^r$*, we have*

$$\begin{aligned} LR(r_0) &\equiv 2\left[\sum_{i=1}^n \ell(Z_i, \widehat{\alpha}_n) - \sum_{i=1}^n \ell(Z_i, \widetilde{\alpha}_n)\right] \\ &= \sup_{\alpha_0 \in \mathcal{A}_I^r, \lambda \in U^{d_\phi}} \left[\sqrt{n}\mu_n\left\{u_n^*(\cdot, \alpha_0, \lambda)\right\}\right]^2 + o_{P_Z}(1) \\ &= \sup_{d \in \mathcal{D}_{k(n)}^{eff}} \left[\sqrt{n}\mu_n\left\{d(\cdot)\right\}\right]^2 + o_{P_Z}(1). \end{aligned}$$

Proof of Theorem B.2: The proof consists of several steps. Let $c > 0$ denote a finite constant in the following proof.

Step 1: By the definitions of $\widehat{\alpha}_n$ and $\widetilde{\alpha}_n$, we have:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widehat{\alpha}_n) - \ell(Z_i, \widetilde{\alpha}_n)] \\ \geq\ & \max\left\{\sup_{t_n \in \mathcal{T}_n, \widetilde{\alpha}_n(t_n) \in \mathcal{B}_n(\alpha_0)} \frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widetilde{\alpha}_n(t_n)) - \ell(Z_i, \widetilde{\alpha}_n)] - o_{P_Z}(n^{-1}), 0\right\} \end{aligned}$$

where $\widetilde{\alpha}_n(t_n) \in \mathcal{B}_n(\alpha_0) \subset \mathcal{A}_{k(n)}$ satisfies

$$\frac{p(z, \widetilde{\alpha}_n(t_n))}{p_0(z)} = \frac{p(z, \widetilde{\alpha}_n)}{p_0(z)} + t_n u_n^*(z, \alpha_0, \lambda).$$

By definitions of $\widetilde{\alpha}_n(t_n)$ and $\widetilde{\alpha}_n$, and by Theorem B.1(1), under the null, we have:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widetilde{\alpha}_n(t_n)) - \ell(Z_i, \widetilde{\alpha}_n)] \\ =\ & t_n\left(\mu_n\left\{u_n^*(Z, \alpha_0, \lambda)\right\} - E_0\left[\left(\frac{p(Z, \widetilde{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right](1 + o(1))\right) \\ & -\frac{t_n^2}{2} + o_{P_Z}(n^{-1}). \end{aligned}$$

Since $\widetilde{\alpha}_n \in \mathcal{B}_n(\alpha_0) \cap \{\alpha : \phi(\alpha) = r_0\}$ wpa1, by Theorem B.1(3), we have:

$$E_0\left[\left(\frac{p(Z, \widetilde{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right] = o_{P_Z}(n^{-1/2}).$$

Then, for all $t_n \in \mathcal{T}_n$ we have:

$$\frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widetilde{\alpha}_n(t_n)) - \ell(Z_i, \widetilde{\alpha}_n)] = t_n \times \mu_n\left\{u_n^*(Z, \alpha_0, \lambda)\right\} - \frac{t_n^2}{2} + o_{P_Z}(n^{-1}),$$

which is maximized at $t_n = \mu_n\{u_n^*(Z, \alpha_0, \lambda)\}$, and hence

$$\begin{aligned} & \sup_{t_n \in \mathcal{T}_n} \frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widetilde{\alpha}_n(t_n)) - \ell(Z_i, \widetilde{\alpha}_n)] - o_{P_Z}(n^{-1}) \\ = & \frac{[\mu_n\{u_n^*(Z, \alpha_0, \lambda)\}]^2}{2} + o_{P_Z}(n^{-1}). \end{aligned}$$

Thus

$$\frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widehat{\alpha}_n) - \ell(Z_i, \widetilde{\alpha}_n)] \geq \max\left\{\frac{[\mu_n\{u_n^*(Z, \alpha_0, \lambda)\}]^2}{2} + o_{P_Z}(n^{-1}), 0\right\}.$$

Step 2. By the definition of $\widetilde{\alpha}_n$, we have:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \ell(Z_i, \widetilde{\alpha}_n) & \geq \sup_{t_n \in \{t \in \mathcal{T}_n, \widehat{\alpha}_n(t) \in \mathcal{B}_n(\alpha_0), \phi(\widehat{\alpha}_n(t)) = r_0\}} \frac{1}{n}\sum_{i=1}^n \ell(Z_i, \widehat{\alpha}_n(t_n)) - o_{P_Z}(n^{-1}) \\ & \geq \frac{1}{n}\sum_{i=1}^n \ell(Z_i, \widehat{\alpha}_n(t_n^*)) - o_{P_Z}(n^{-1}) \end{aligned}$$

for any $t_n^* \in \{t \in \mathcal{T}_n, \widehat{\alpha}_n(t) \in \mathcal{B}_n(\alpha_0), \phi(\widehat{\alpha}_n(t)) = r_0\}$.

By definition of $\widehat{\alpha}_n$, we have: uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and for all $\lambda \in U^{d_\phi}$,

$$\begin{aligned} 0 & \leq \frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widehat{\alpha}_n) - \ell(Z_i, \widetilde{\alpha}_n)] \\ & \leq \max\left\{\frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widehat{\alpha}_n) - \ell(Z_i, \widehat{\alpha}_n(t_n^*))] + o_{P_Z}(n^{-1}), 0\right\}, \end{aligned}$$

where $\widehat{\alpha}_n(t_n^*) \in \mathcal{B}_n(\alpha_0) \subset \mathcal{A}_{k(n)}$ satisfies

$$\phi(\widehat{\alpha}_n(t_n^*)) = \phi(\alpha_0) = r_0 = \phi(\widetilde{\alpha}_n) \text{ for all } \alpha_0 \in \mathcal{A}_I^r, \tag{B.2}$$

and

$$\frac{p(Z, \widehat{\alpha}_n(t_n^*))}{p_0(Z)} = \frac{p(Z, \widehat{\alpha}_n)}{p_0(Z)} + t_n^* \times u_n^*(Z, \alpha_0, \lambda). \tag{B.3}$$

By Theorem B.1(1)(2), we have: for all $t_n^* \in \mathcal{T}_n$ satisfying (B.3),

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n [\ell(Z_i, \widehat{\alpha}_n) - \ell(Z_i, \widehat{\alpha}_n(t_n^*))] \\ = & -t_n^*\left(\mu_n\{u_n^*(Z, \alpha_0, \lambda)\} - E_0\left[\left(\frac{p(Z, \widehat{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z, \alpha_0, \lambda)\right](1 + o(1))\right) \\ & + \frac{t_n^{*2}}{2} + o_{P_Z}(n^{-1}) \\ = & \frac{t_n^{*2}}{2} + o_{P_Z}(n^{-1}), \end{aligned}$$

we could let

$$t_n^* = -E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z,\alpha_0,\lambda)\right] + \varepsilon_n^* \tag{B.4}$$

for some $\varepsilon_n^* = o_{P_Z}(n^{-1/2})$ to be specified later. Then

$$\frac{1}{n}\sum_{i=1}^{n}[\ell(Z_i,\widehat{\alpha}_n) - \ell(Z_i,\widehat{\alpha}_n(t_n^*))] = \frac{\left[\mu_n\left\{u_n^*(Z,\alpha_0,\lambda)\right\}\right]^2}{2} + o_{P_Z}(n^{-1})$$

and

$$0 \leq \frac{1}{n}\sum_{i=1}^{n}[\ell(Z_i,\widehat{\alpha}_n) - \ell(Z_i,\widetilde{\alpha}_n)] \leq \frac{\left[\mu_n\left\{u_n^*(Z,\alpha_0,\lambda)\right\}\right]^2}{2} + o_{P_Z}(n^{-1}).$$

It remains to find an $\varepsilon_n^* = o_{P_Z}(n^{-1/2})$ satisfying (B.4), (B.2) and (B.3). By Restriction (B.2), $\widehat{\alpha}_n(t_n^*) \in \mathcal{B}_n(\alpha_0) \cap \{\alpha : \phi(\alpha) = r_0\}$ wpa1, by Theorem B.1(3), we have: uniformly in $\alpha_0 \in \mathcal{A}_I^r$ and $\lambda \in U^{d_\phi}$,

$$E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n(t_n^*))}{p_0(Z)} - 1\right) u_n^*(Z,\alpha_0,\lambda)\right] = o_{P_Z}(n^{-1/2}).$$

This, (B.3) and (B.4) together imply that such an $\varepsilon_n^* = o_{P_Z}(n^{-\frac{1}{2}})$ exists:

$$\begin{aligned} o_{P_Z}(n^{-\frac{1}{2}}) &= E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n(t_n^*))}{p_0(z)} - 1\right) u_n^*(Z,\alpha_0,\lambda)\right] \\ &= E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)} - 1\right) u_n^*(Z,\alpha_0,\lambda)\right] + t_n^* = \varepsilon_n^* \end{aligned}$$

Step 3. Combining Steps 1 and 2, we obtain:

$$\begin{aligned} LR(r_0) &= 2\left[\sum_{i=1}^{n}\ell(Z_i,\widehat{\alpha}_n) - \sum_{i=1}^{n}\ell(Z_i,\widetilde{\alpha}_n)\right] \\ &= \sup_{\alpha_0\in\mathcal{A}_I^r,\lambda\in U^{d_\phi}}\left[\sqrt{n}\mu_n\left\{u_n^*(Z,\alpha_0(\widehat{\alpha}_n),\lambda)\right\}\right]^2 + o_{P_Z}(1) \\ &= \sup_{d\in\mathcal{D}_{k(n)}^{eff}}\left[\sqrt{n}\mu_n\left\{d(\cdot)\right\}\right]^2 + o_{P_Z}(1). \end{aligned}$$

Now the conclusion follows. *Q.E.D.*

C Proof of Theorem 5.1

Proof of Theorem 5.1: Denote $\ell^\omega(Z,\alpha) \equiv \omega \log p(Z,\alpha)$. By assumption 5.1, we have $E_{ZW}\left((\omega_i - 1)v_n^*(z,\alpha_0,\lambda)\right) = 0$ and

$$E_{ZW}[\ell^\omega(Z_i,\alpha) - \ell^\omega(Z_i,\alpha_0)] = E_Z[\ell(Z_i,\alpha) - \ell(Z_i,\alpha_0)] = -K(\alpha_0,\alpha).$$

$$\ell^{\omega}(Z,\alpha)-\ell^{\omega}(Z,\alpha_0)=\omega\log\left(1+\left[\frac{p(Z,\alpha)}{p(Z,\alpha_0)}-1\right]\right)=\omega\log\left[1+\chi(\alpha,\alpha_0)s(Z;\alpha)\right]$$
$$\begin{aligned}&= \omega\chi(\alpha,\alpha_0)s(Z;\alpha)-\frac{\omega}{2}[\chi(\alpha,\alpha_0)s(Z;\alpha)]^2\left\{1-\mathrm{rem}(\chi(\alpha,\alpha_0)s(Z;\alpha))\right\}\\ &\equiv \omega\chi(\alpha,\alpha_0)s(Z;\alpha)+\omega R(Z;\alpha,\alpha_0).\end{aligned}$$

Thus

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\alpha)-\ell^{\omega}(Z_i,\alpha_0)]\\ &= -K(\alpha_0,\alpha)+\mu_n\left\{\omega\chi(\alpha,\alpha_0)s(Z;\alpha)\right\}+\mu_n\left\{\omega R(Z;\alpha,\alpha_0)\right\}\\ &= -\frac{1}{2}[\chi(\alpha,\alpha_0)]^2(1+o(1))+\mu_n\left\{\omega\chi(\alpha,\alpha_0)s(Z;\alpha)\right\}+\mu_n\left\{\omega R(Z;\alpha,\alpha_0)\right\}.\end{aligned}$$

We now consider perturbation in probability density sieve space: for all $\alpha\in\mathcal{B}_n(\alpha_0)$ and $t_n\in\mathcal{T}_n$

$$p(z,\alpha(t_n))\equiv p(z,\alpha)+t_n u_n^*(z,\alpha_0,\lambda)p_0(z).$$

Thus

$$\begin{aligned}&\mu_n\left\{\omega\chi(\alpha(t_n),\alpha_0)s(Z;\alpha(t_n))\right\}-\mu_n\left\{\omega\chi(\alpha,\alpha_0)s(Z;\alpha)\right\}\\ &= \mu_n\left\{\omega\left[\frac{p(Z,\alpha(t_n))}{p(Z,\alpha_0)}-\frac{p(Z,\alpha)}{p(Z,\alpha_0)}\right]\right\}=t_n\times\mu_n\left\{\omega u_n^*(Z,\alpha_0,\lambda)\right\}.\end{aligned}$$

Therefore, uniformly over $\alpha_0\in(\mathcal{A}_I,||\cdot||_A)$, $\lambda\in U^{d_\phi}$, $\alpha\in\mathcal{B}_n(\alpha_0)$ and $t_n\in\mathcal{T}_n$, we have:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\alpha(t_n))-\ell^{\omega}(Z_i,\alpha)]\\ &= \frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\alpha(t_n))-\ell^{\omega}(Z_i,\alpha_0)]-\frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\alpha)-\ell^{\omega}(Z_i,\alpha_0)]\\ &= -\frac{2t_nE_Z\left[\left(\frac{p(Z,\alpha)}{p_0(z)}-1\right)u_n^*(Z,\alpha_0,\lambda)\right]+t_n^2}{2}(1+o(1))\\ &\quad +t_n\times\mu_n\left\{\omega u_n^*(Z,\alpha_0,\lambda)\right\}+o_{P_{ZW}}(n^{-1}).\end{aligned}$$

Define the unconstraint weighted sieve MLE $\widehat{\alpha}_n^{\omega}$ as

$$\begin{aligned}\widehat{\alpha}_n^{\omega} &\in \widehat{\mathcal{A}}_n^{\omega}\equiv\arg\max_{\alpha\in\mathcal{A}_n}\left\{\sum_{i=1}^{n}\omega_i\log p(Z_i;\alpha)-o_{P_{ZW}}(1)\right\}\\ &= \arg\max_{\alpha\in\mathcal{A}_n}\left\{\sum_{i=1}^{n}\ell^{\omega}(Z_i,\alpha)-o_{P_{ZW}}(1)\right\},\end{aligned}$$

and the constraint weighted sieve MLE $\widetilde{\alpha}_n^{\omega}$ as:

$$\begin{aligned}\widetilde{\alpha}_n^{\omega} &\in \widetilde{\mathcal{A}}_n^{\omega} \equiv \arg \max_{\{\alpha \in \mathcal{A}_n:\ \phi(\alpha)=\widehat{r}\}} \left\{ \sum_{i=1}^{n} \omega_i \log p(Z_i;\alpha) - o_{P_{ZW}}(1) \right\} \\ &= \arg \max_{\{\alpha \in \mathcal{A}_n:\ \phi(\alpha)=\widehat{r}\}} \left\{ \sum_{i=1}^{n} \ell^{\omega}(Z_i,\alpha) - o_{P_{ZW}}(1) \right\}.\end{aligned}$$

Theorem C.1 *Suppose that Assumptions 4.1 and 4.2 hold and that $\widehat{\alpha}_n$ is a sieve MLE. Then: (1) uniformly over $\alpha_0 \in (\mathcal{A}_I, ||\cdot||_A)$, $\lambda \in U^{d_\phi}$, $\alpha \in \mathcal{B}_n(\alpha_0)$ and $t_n \in \mathcal{T}_n$,*

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \omega_i [\ell(Z_i, \alpha(t_n)) - \ell(Z_i,\alpha)] \\ =\ & t_n \left(\mu_n \{\omega u_n^*(Z,\alpha_0,\lambda)\} - E_Z\left[\left(\frac{p(Z,\alpha)}{p_0(Z)} - 1 \right) u_n^*(Z,\alpha_0,\lambda) \right] (1+o(1)) \right) \\ & -\frac{t_n^2}{2} + o_{P_{ZW}}(n^{-1});\end{aligned}$$

(2) Uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi}$,

$$\left| E_Z\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega})}{p_0(Z)} - 1 \right) u_n^*(Z,\alpha_0,\lambda) \right] - \mu_n\{\omega u_n^*(Z,\alpha_0,\lambda)\} \right| = o_{P_{ZW}}(n^{-\frac{1}{2}}).$$

$$\left| E_Z\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega})}{p_0(Z)} - \frac{p(Z,\widehat{\alpha})}{p_0(Z)} \right) u_n^*(Z,\alpha_0,\lambda) \right] - \mu_n\{(\omega-1) u_n^*(Z,\alpha_0,\lambda)\} \right| = o_{P_{ZW}}(n^{-\frac{1}{2}}).$$

Thus under assumptions 4.3(i) and 5.1, we have: uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi}$,

$$\left| E_Z\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega})}{p_0(Z)} - \frac{p(Z,\widehat{\alpha})}{p_0(Z)} \right) u_n^*(Z,\alpha_0,\lambda) \right] \right| = O_{P_Z}(n^{-\frac{1}{2}}).$$

(3) Uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi}$,

$$\sup_{\alpha \in \mathcal{B}_n(\alpha_0) \cap \{\alpha:\phi(\alpha)=\phi(\widehat{\alpha})\}} \left| E_Z\left[\left(\frac{p(Z,\alpha)}{p_0(Z)} - \frac{p(Z,\widehat{\alpha})}{p_0(Z)} \right) u_n^*(Z,\alpha_0,\lambda) \right] \right| = o_{P_{WZ}}(n^{-\frac{1}{2}}).$$

Proof of Theorem C.1: **Result (1)** is already proved before the statement of the theorem.

For **Result (2)**. Let $\varepsilon_n = o(n^{-\frac{1}{2}})$. For any $\alpha \in \mathcal{B}_n(\alpha_0) \subset \mathcal{A}_n$, consider a local perturbation

$$p(\cdot, \alpha(\varepsilon_n)) = p(\cdot,\alpha) \pm \varepsilon_n u_n^*(\cdot,\alpha_0,\lambda) p_0(\cdot) \in \mathcal{P}_{k(n)}.$$

By assumption 4.2: uniformly over $\alpha_0 \in (\mathcal{A}_I^r, ||\cdot||_A)$ and $\lambda \in U^{d_\phi}$, and by the definition of $\widehat{\alpha}_n^\omega$, we have

$$\begin{aligned} -o_{P_Z}(n^{-1}) &\leq \frac{1}{n}\sum_{i=1}^n \ell^\omega(Z_i, \widehat{\alpha}_n^\omega) - \frac{1}{n}\sum_{i=1}^n \ell^\omega(Z_i, \widehat{\alpha}_n^\omega(\varepsilon_n)) \\ &= \frac{\pm 2\varepsilon_n E_Z\left[\left(\frac{p(Z,\widehat{\alpha}_n^\omega)}{p_0(Z)} - 1\right) u_n^*(Z,\alpha_0,\lambda)\right] + \varepsilon_n^2}{2}(1+o(1)) \\ &\quad \mp \varepsilon_n \times \mu_n\left\{\omega u_n^*(Z,\alpha_0,\lambda)\right\} + o_{P_{ZW}}(n^{-1}). \end{aligned}$$

By the definition of $u_n^*(Z,\alpha_0,\lambda)$ and $\varepsilon_n = o_{P_Z}(n^{-\frac{1}{2}})$, we obtain

$$\left|E_Z\left[\left(\frac{p(Z,\widehat{\alpha}_n^\omega)}{p_0(Z)} - 1\right) u_n^*(Z,\alpha_0,\lambda)\right] - \mu_n\left\{\omega u_n^*(Z,\alpha_0,\lambda)\right\}\right| = o_{P_Z}(n^{-\frac{1}{2}}).$$

For **Result (3)**. For all $\alpha \in \mathcal{B}_n(\alpha_0) \cap \{\alpha : \phi(\alpha) = \phi(\widehat{\alpha})\}$ and for all $\alpha_0 \in \mathcal{A}_I^r \equiv \{\alpha \in \mathcal{A}_I : \phi(\alpha) = r_0\}$, under Assumption 4.1.(ii), we have, uniformly over $\alpha_0 \in \mathcal{A}_I^r$:

$$\frac{\left|\lambda'\left(\phi(\alpha) - \phi(\alpha_0) - \frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha - \alpha_0]\right)\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||} = o(n^{-\frac{1}{2}})$$

$$\frac{\left|\lambda'\left(\phi(\widehat{\alpha}) - \phi(\alpha_0) - \frac{\partial\phi(\alpha_0)}{\partial\alpha}[\widehat{\alpha} - \alpha_0]\right)\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||} = o(n^{-\frac{1}{2}})$$

For all $\alpha \in \mathcal{B}_n(\alpha_0) \cap \{\alpha : \phi(\alpha) = \phi(\widehat{\alpha})\}$ we have

$$\frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha - \widehat{\alpha}]\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||} = \frac{\left|\lambda'\left(\phi(\alpha) - \phi(\widehat{\alpha}) - \frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha - \widehat{\alpha}]\right)\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||} = o(n^{-\frac{1}{2}})$$

$$\left|E_Z\left[\left(\frac{p(Z,\alpha)}{p_0(Z)} - \frac{p(Z,\widehat{\alpha})}{p_0(Z)}\right) u_n^*(Z,\alpha_0,\lambda)\right]\right| = \frac{\left|\lambda'\frac{\partial\phi(\alpha_0)}{\partial\alpha}[\alpha - \widehat{\alpha}]\right|}{||v_n^*(\cdot,\alpha_0,\lambda)||} = o_{P_{ZW}}(n^{-\frac{1}{2}}).$$

Thus **Result (3)** is true. *Q.E.D*

Theorem C.2 *Suppose that assumptions 4.1, 4.2 and 4.3(i) hold. Then under null of $\alpha_0 \in \mathcal{A}_I^r$, we have*

$$\begin{aligned} LR^\omega(\widehat{r}) &\equiv 2\left[\sum_{i=1}^n \ell^\omega(Z_i,\widehat{\alpha}_n^\omega) - \sum_{i=1}^n \ell^\omega(Z_i,\widetilde{\alpha}_n^\omega)\right] \\ &= \sup_{\alpha_0\in\mathcal{A}_I^r, \lambda\in U^{d_\phi}} \left[\sqrt{n}\mu_n\left\{(\omega-1)u_n^*(\cdot,\alpha_0,\lambda)\right\}\right]^2 + o_{P_{ZW}}(1) \\ &= \sup_{d\in\mathcal{D}_{k(n)}^{eff}} \left[\sqrt{n}\mu_n\left\{(\omega-1)d(\cdot)\right\}\right]^2 + o_{P_{ZW}}(1). \end{aligned}$$

PROOF OF THEOREM C.2: The proof consists of several steps. Let $c > 0$ denote a finite constant in the following proof.

Step 1: By the definitions of $\widehat{\alpha}_n^\omega$ and $\widetilde{\alpha}_n^\omega$, we have:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n [\ell^\omega(Z_i, \widehat{\alpha}_n^\omega) - \ell^\omega(Z_i, \widetilde{\alpha}_n^\omega)] \\
\geq\ & \max\left\{ \sup_{t_n \in \mathcal{T}_n, \widetilde{\alpha}_n^\omega(t_n) \in \mathcal{B}_n(\alpha_0)} \frac{1}{n}\sum_{i=1}^n [\ell^\omega(Z_i, \widetilde{\alpha}_n^\omega(t_n)) - \ell^\omega(Z_i, \widetilde{\alpha}_n^\omega)] - o_{P_{ZW}}(n^{-1}), 0 \right\}
\end{aligned}$$

where $\widetilde{\alpha}_n^\omega(t_n) \in \mathcal{B}_n(\alpha_0) \subset \mathcal{A}_{k(n)}$ satisfies

$$\frac{p(Z, \widetilde{\alpha}_n^\omega(t_n))}{p_0(Z)} = \frac{p(Z, \widetilde{\alpha}_n^\omega)}{p_0(Z)} + t_n u_n^*(Z, \alpha_0, \lambda).$$

By definitions of $\widetilde{\alpha}_n^\omega(t_n)$ and $\widetilde{\alpha}_n^\omega$, and by Theorem C.1(1), under the null, we have:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n [\ell^\omega(Z_i, \widetilde{\alpha}_n^\omega(t_n)) - \ell^\omega(Z_i, \widetilde{\alpha}_n^\omega)] \\
=\ & t_n \left(\mu_n\left\{\omega u_n^*(Z, \alpha_0, \lambda)\right\} - E_0\left[\left(\frac{p(Z, \widetilde{\alpha}_n^\omega)}{p_0(Z)} - 1 \right) u_n^*(Z, \alpha_0, \lambda) \right] (1 + o(1)) \right) \\
& - \frac{t_n^2}{2} + o_{P_{ZW}}(n^{-1}) \\
=\ & t_n \left(\mu_n\left\{(\omega - 1) u_n^*(Z, \alpha_0, \lambda)\right\} - E_0\left[\left(\frac{p(Z, \widetilde{\alpha}_n^\omega)}{p_0(Z)} - \frac{p(Z, \widehat{\alpha}_n)}{p_0(Z)} \right) u_n^*(Z, \alpha_0, \lambda) \right] (1 + o(1)) \right) \\
& - \frac{t_n^2}{2} + o_{P_{ZW}}(n^{-1}),
\end{aligned}$$

where the last equality is due to the fact that $t_n \in \mathcal{T}_n$ and by Theorem B.1(2),

$$E_0\left[\left(\frac{p(Z, \widehat{\alpha}_n)}{p_0(Z)} - 1 \right) u_n^*(Z, \alpha_0, \lambda) \right] (1 + o(1)) - \mu_n\left\{u_n^*(Z, \alpha_0, \lambda)\right\} = o_{P_Z}(n^{-1/2}).$$

Since $\widetilde{\alpha}_n^\omega \in \mathcal{B}_n(\alpha_0) \cap \{\alpha : \phi(\alpha) = \phi(\widehat{\alpha}_n)\}$ wpa1, by Theorem C.1(3), we have:

$$E_0\left[\left(\frac{p(Z, \widetilde{\alpha}_n^\omega)}{p_0(Z)} - \frac{p(Z, \widehat{\alpha}_n)}{p_0(Z)} \right) u_n^*(Z, \alpha_0, \lambda) \right] = o_{P_Z}(n^{-1/2}).$$

Then, for all $t_n \in \mathcal{T}_n$ we have:

$$\frac{1}{n}\sum_{i=1}^n [\ell^\omega(Z_i, \widetilde{\alpha}_n^\omega(t_n)) - \ell^\omega(Z_i, \widetilde{\alpha}_n^\omega)] = t_n \times \mu_n\left\{(\omega - 1) u_n^*(Z, \alpha_0, \lambda)\right\} - \frac{t_n^2}{2} + o_{P_{ZW}}(n^{-1}),$$

which is maximized at $t_n = \mu_n\left\{(\omega - 1) u_n^*(Z, \alpha_0, \lambda)\right\}$, and hence

$$\begin{aligned}
&\sup_{t_n \in \mathcal{T}_n} \frac{1}{n}\sum_{i=1}^n [\ell^\omega(Z_i, \widetilde{\alpha}_n^\omega(t_n)) - \ell^\omega(Z_i, \widetilde{\alpha}_n^\omega)] - o_{P_{ZW}}(n^{-1}) \\
=\ & \frac{\left[\mu_n\left\{(\omega - 1) u_n^*(Z, \alpha_0, \lambda)\right\}\right]^2}{2} + o_{P_{ZW}}(n^{-1}).
\end{aligned}$$

Thus

$$\frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega})-\ell^{\omega}(Z_i,\widetilde{\alpha}_n^{\omega})] \geq \max\left\{\frac{\left[\mu_n\left\{(\omega-1)u_n^*(Z,\alpha_0,\lambda)\right\}\right]^2}{2}+o_{P_{ZW}}(n^{-1}),0\right\}.$$

Step 2. By the definition of $\widetilde{\alpha}_n^{\omega}$, we have:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\ell^{\omega}(Z_i,\widetilde{\alpha}_n^{\omega}) &\geq \sup_{t_n\in\{t\in\mathcal{T}_n:\widehat{\alpha}_n^{\omega}(t)\in\mathcal{B}_n(\alpha_0),\phi(\widehat{\alpha}_n^{\omega}(t))=\phi(\widehat{\alpha}_n)\}}\frac{1}{n}\sum_{i=1}^{n}\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}(t_n))-o_{P_{ZW}}(n^{-1})\\ &\geq \frac{1}{n}\sum_{i=1}^{n}\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}(t_n^*))-o_{P_{ZW}}(n^{-1})\end{aligned}$$

for any $t_n^*\in\{t\in\mathcal{T}_n:\widehat{\alpha}_n^{\omega}(t)\in\mathcal{B}_n(\alpha_0),\phi(\widehat{\alpha}_n^{\omega}(t))=\phi(\widehat{\alpha}_n)\}$.

By definition of $\widehat{\alpha}_n$, we have: uniformly over $\alpha_0\in(\mathcal{A}_I^r,||\cdot||_A)$ and for all $\lambda\in U^{d_\phi}$,

$$\begin{aligned}0 &\leq \frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega})-\ell^{\omega}(Z_i,\widetilde{\alpha}_n^{\omega})]\\ &\leq \max\left\{\frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega})-\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}(t_n^*))]+o_{P_{ZW}}(n^{-1}),0\right\},\end{aligned}$$

where $\widehat{\alpha}_n^{\omega}(t_n^*)\in\mathcal{B}_n(\alpha_0)\subset\mathcal{A}_{k(n)}$ satisfies

$$\phi(\widehat{\alpha}_n^{\omega}(t_n^*))=\phi(\widehat{\alpha}_n)=\widehat{r}=\phi(\widetilde{\alpha}_n^{\omega}) \text{ for all } \alpha_0\in\mathcal{A}_I^r, \tag{C.1}$$

and

$$\frac{p(Z,\widehat{\alpha}_n^{\omega}(t_n^*))}{p_0(Z)}=\frac{p(Z,\widehat{\alpha}_n^{\omega})}{p_0(Z)}+t_n^*\times u_n^*(Z,\alpha_0,\lambda). \tag{C.2}$$

By Theorem C.1(1)(2), we have: for all $t_n^*\in\mathcal{T}_n$ satisfying (C.2),

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega})-\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}(t_n^*))]\\ =\;& -t_n^*\left(\mu_n\left\{\omega u_n^*(Z,\alpha_0,\lambda)\right\}-E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega})}{p_0(Z)}-1\right)u_n^*(Z,\alpha_0,\lambda)\right](1+o(1))\right)\\ &+\frac{t_n^{*2}}{2}+o_{P_{ZW}}(n^{-1})\\ =\;&\frac{t_n^{*2}}{2}+o_{P_{ZW}}(n^{-1}),\end{aligned}$$

we could let

$$\begin{aligned}t_n^* &= -E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega})}{p_0(Z)}-\frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)}\right)u_n^*(Z,\alpha_0,\lambda)\right]+\varepsilon_n^* &\text{(C.3)}\\ &= -\mu_n\left\{(\omega-1)u_n^*(Z,\alpha_0,\lambda)\right\}+o_{P_{ZW}}(n^{-1/2}) &\text{(C.4)}\end{aligned}$$

for some $\varepsilon_n^* = o_{P_{ZW}}(n^{-1/2})$ to be specified later. Then

$$\frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}) - \ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}(t_n^*))] = \frac{\left[\mu_n\left\{(\omega-1)u_n^*(Z,\alpha_0,\lambda)\right\}\right]^2}{2} + o_{P_{ZW}}(n^{-1})$$

and

$$0 \leq \frac{1}{n}\sum_{i=1}^{n}[\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}) - \ell^{\omega}(Z_i,\widetilde{\alpha}_n^{\omega})] \leq \frac{\left[\mu_n\left\{(\omega-1)u_n^*(Z,\alpha_0,\lambda)\right\}\right]^2}{2} + o_{P_{ZW}}(n^{-1}).$$

It remains to find an $\varepsilon_n^* = o_{P_{ZW}}(n^{-1/2})$ satisfying (C.3), (C.1) and (C.2). By Restriction (C.1), $\widehat{\alpha}_n^{\omega}(t_n^*) \in \mathcal{B}_n(\alpha_0) \cap \{\alpha : \phi(\widehat{\alpha}_n^{\omega}(t_n^*)) = \phi(\widehat{\alpha}_n) = \phi(\widetilde{\alpha}_n^{\omega})\}$ wpa1, by Theorem C.1(3), we have: uniformly in $\alpha_0 \in \mathcal{A}_I^r$ and $\lambda \in U^{d_\phi}$,

$$E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega}(t_n^*))}{p_0(Z)} - \frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)}\right)u_n^*(Z,\alpha_0,\lambda)\right] = o_{P_{ZW}}(n^{-1/2}).$$

This, (C.2) and (C.3) together imply that such an $\varepsilon_n^* = o_{P_{ZW}}(n^{-\frac{1}{2}})$ exists:

$$\begin{aligned} o_{P_{ZW}}(n^{-\frac{1}{2}}) &= E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega}(t_n^*))}{p_0(Z)} - \frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)}\right)u_n^*(Z,\alpha_0,\lambda)\right] \\ &= E_0\left[\left(\frac{p(Z,\widehat{\alpha}_n^{\omega})}{p_0(Z)} - \frac{p(Z,\widehat{\alpha}_n)}{p_0(Z)}\right)u_n^*(Z,\alpha_0,\lambda)\right] + t_n^* = \varepsilon_n^* \end{aligned}$$

Step 3. Combining Steps 1 and 2, we obtain:

$$\begin{aligned} LR^{\omega}(\widehat{r}) &\equiv 2\left[\sum_{i=1}^{n}\ell^{\omega}(Z_i,\widehat{\alpha}_n^{\omega}) - \sum_{i=1}^{n}\ell^{\omega}(Z_i,\widetilde{\alpha}_n^{\omega})\right] \\ &= \sup_{\alpha_0\in\mathcal{A}_I^r,\lambda\in U^{d_\phi}}\left[\sqrt{n}\mu_n\left\{(\omega-1)u_n^*(\cdot,\alpha_0,\lambda)\right\}\right]^2 + o_{P_{ZW}}(1) \\ &= \sup_{d\in\mathcal{D}_{k(n)}^{eff}}\left[\sqrt{n}\mu_n\left\{(\omega-1)d(\cdot)\right\}\right]^2 + o_{P_{ZW}}(1). \end{aligned}$$

Now the conclusion follows. *Q.E.D.*