LIKELIHOOD INFERENCE IN SOME FINITE MIXTURE MODELS

By

Xiaohong Chen, Maria Ponomareva and Elie Tamer

MAY 2013

COWLES FOUNDATION DISCUSSION PAPER NO. 1895

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Likelihood Inference in Some Finite Mixture Models *

Xiaohong Chen[†] Yale
Maria Ponomareva[‡] NIU
Elie Tamer[§] Northwestern

First Version June 2012 - Current Version May 6, 2013

Abstract

Parametric mixture models are commonly used in applied work, especially empirical economics, where these models are often employed to learn for example about the proportions of various types in a given population. This paper examines the inference question on the proportions (mixing probability) in a simple mixture model in the presence of nuisance parameters when sample size is large. It is well known that likelihood inference in mixture models is complicated due to 1) lack of point identification, and 2) parameters (for example, mixing probabilities) whose true value may lie on the boundary of the parameter space. These issues cause the profiled likelihood ratio (PLR) statistic to admit asymptotic limits that differ discontinuously depending on how the true density of the data approaches the regions of singularities where there is lack of point identification. This lack of uniformity in the asymptotic distribution suggests that confidence intervals based on pointwise asymptotic approximations might lead to faulty inferences. This paper examines this problem in details in a finite mixture model and provides possible fixes based on the parametric bootstrap. We examine the performance of this parametric bootstrap in Monte Carlo experiments and apply it to data from Beauty Contest experiments. We also examine small sample inferences and projection methods.

Keywords: Finite Mixtures, Parametric Bootstrap, Profiled Likelihood Ratio Statistic, Partial Identification, Parameter on the Boundary

*We thank Don Andrews, Ivan Canay, Xu Cheng, Ulrich Mueller and participants at June 2012 Cowles Summer Conference, Princeton Conference on Non-Standard Econometrics, SEA conference, Harvard/MIT, LSE and OSU seminars for comments. We thank Norm Swanson and referee comments that helped improve the paper. We also thank Rosemarie Nagel for providing us with the data.

[†]Department of Economics, Yale University, xiaohong.chen@yale.edu, ph: 203-432-5852. Support from Cowles Foundation for Research in Economics is gratefully acknowledged.

[‡]Department of Economics, Northern Illinois University, mponomareva@niu.edu, ph: 815-753-6434.

[§]Corresponding author: Department of Economics, Northwestern University, tamer@northwestern.edu, ph: 847-491-8218. Support from The National Science Foundation is gratefully acknowledged.

1 Introduction

This paper studies the question of inference, mainly testing and confidence regions, on the mixing probability in the following finite mixture model with two components where the density of the observed data is:

$$p(\cdot;\theta,\delta) = \delta f_\theta + (1-\delta) f_0 \quad \text{with} \quad f_\theta = f(\cdot,\theta) \quad \text{and} \quad f_0 = f(\cdot,\theta_0) \tag{1.1}$$

The mixing probability δ takes values in the closed interval $[0,1]$. We observe a sample of n independent random draws $\{X_i, i = 1,\ldots,n\}$ from the density $p(.;\theta,\delta)$, and are interested in inference on δ in the presence of a nuisance parameter $\theta \in \Theta$, a *compact* subset of $\mathbb{R}^k$. Also, here we assume that θ_0 and the form of $f(.;.)$ are known. The model above is an member of a class of parametric finite mixture models, and in this paper we focus on complications that arise mainly due to the possibility that the true parameters in (1.1) are in the *singularity region* where $\delta * \|\theta - \theta_0\| = 0$. The singularity region leads to two problems: lack of identification (if $\theta = \theta_0$, the model has no information about δ and if $\delta = 0$, the model has no information about θ) and parameters lying on the boundary of the parameter space (when $\delta = 0$). Those two issues create problems for inference based on the maximum likelihood estimators of δ and θ, since the maximum likelihood estimators of the parameters in the singularity region are no longer necessarily consistent, and the asymptotic distribution of the likelihood ratio statistic is no longer standard.

Allowing for cases in which the true parameters can lie in this singularity region is key in mixture models as it is related to learning the number of mixture components in a population, which in many cases is the main object of interest in applications. Each point (δ,θ) in the a singularity region, plotted in Figure 1 below, leads to *the same* density for the observed data, i.e., $p(\cdot) = f_0(\cdot)$. We use a profile likelihood ratio statistic to construct confidence region for δ while treating θ as a nuisance parameter. *The main objective of this paper is to examine the asymptotic behavior of this profiled likelihood ratio statistic when the true model lies close to the singularity region.*

The pointwise asymptotic distribution of this profiled likelihood ratio statistic (or PLR) has a well known limit distribution even when true (δ,θ) belong to the singularity region (see, for example, Liu and Shao (2003)). We complement these results by showing that the limit distribution of this PLR statistic is a discontinuous function of true δ when this true δ is in a close neighborhood and drifting towards the singularity region at a given rate. This discontinuity in the asymptotic distribution of the PLR statistic -or lack of uniformity- when the true parameters are sufficiently close to the singularity region can cause misleading inferences (such as undersized test) when these pointwise limit distributions are

used with finite samples. We first examine the nature of this discontinuity by deriving the asymptotic distribution of the PLR statistic under drifting -towards the singularity region- sequences of true parameters, and second we propose an approach to inference using a parametric bootstrap procedure. The reason to consider the bootstrap in this setup is that the asymptotic limits involve supremum over complicated stochastic process that are hard to deal with. In addition, the parametric bootstrap seems like a natural approach to use in the setup above. We evaluate these issues using some simulations and an empirical application with experimental data.

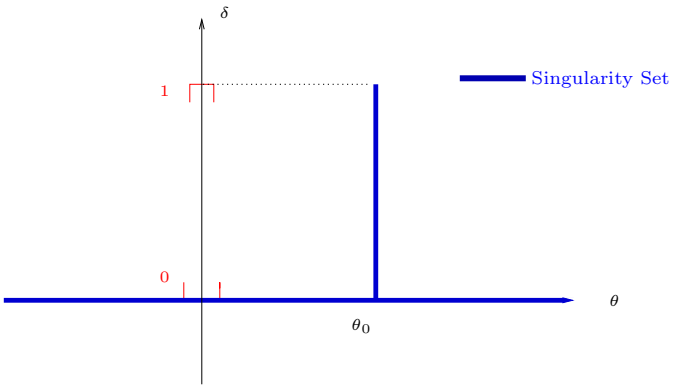

Figure 1: Singularity set in the main model (1.1) : when true δ is equal to zero, then the implied density of the data is f_0 no matter what the value of θ is, i.e., when true δ is zero, the parameter θ is not identified at all. Also, when true θ is equal to θ_0, the implied density of the data is also f_0 and δ is completely not identified.

There are a few well known issues that one must confront in the above setup. The first is the lack of identification in the region of singularity. Our approach delivers confidence sets that maintain coverage whether or not the model is identified. These sets are constructed by inverting the PLR statistic. In the case where the true model belongs to the singularity region, the confidence sets will be the whole parameter space. Another non-standard issue is that in the singularity region some parameters lie on the boundary of the parameter space, which also creates discontinuities in the limiting distribution of the PLR statistic. Due to these two problems, we take an approach of drifting sequences of true parameters (local to the singularity region approach) to derive the limiting distribution of the PLR statistic. In particular, if these sequences stay away from the singularity set, then the PLR statistic has regular χ^2 limits. As these sequences are allowed to approach the singularity set, the PLR statistic admits varying limits depending on the rate at which the true parameter sequence approaches the singularity and the location of the limit point in the singularity set. Critical sequences are the ones where δ approaches the singularity region at the rate of

square root of the sample size. We propose a parametric bootstrap as the method to construct valid confidence set for δ (or joint confidence sets for (δ, θ)). In most cases, the parametric bootstrap mimics the small sample distribution and is shown to be consistent (even in some cases when the limiting mixing probability δ lies on the boundary of the parameter space). The parametric bootstrap is a particularly attractive approach to inference here since using the asymptotic distribution directly may be computationally difficult. In critical cases, we show how the parametric bootstrap can be adjusted in such a way to guarantee the correct uniform coverage.

So, the empirical takeaway from the paper is that when doing inference on the mixing probabilities in the presence of other parameters, a theoretically attractive approach is to build confidence regions by inverting a (profiled) likelihood ratio statistic. Getting critical values is complicated, but the parametric bootstrap seems to do a reasonable job in approximating the small sample distribution.

Although the model we focus on in this paper is simple, it is a prototypical case that highlights the statistical problems that arise when analyzing more complicated models (with more than 2 components and/or vector θ) and so we consider this model in details and discuss extending our methods to more complicated cases (with vector θ's and larger number of mixtures) in the Appendix.

1.1 Motivation, Examples, and Literature

Mixture models are important modeling tools in all areas of applied statistics. See for example McLachlan and Peel (2000). In empirical economics, finite mixtures are used to introduce unobserved heterogeneity. In a nutshell, suppose that an individual or a datum can be one of K types, and each type $k \in \{1, \ldots, K\}$ leads to "behavior" with a density f_k. Then, since we do not observe individuals' types, the likelihood of the observed data is a mixture over these densities with the proportion of types being a main object of interest. An important example of this setup from the econometrics literature is Keane and Wolpin (1997).

In addition, mixture models can arise when analyzing some class of games with multiple Nash equilibria. For example, one equilibrium can involve pure strategies and one in mixed strategies[1] The observed data are proportions of various outcomes where here a given outcome can be observed if 1) it is the pure strategy equilibrium, or 2) if it is on the support of the mixed strategy equilibrium. So, the predicted proportions will be a mixture where the mixing weights are the selection probabilities. See for example Berry and Tamer (2006).

[1]One such game is a 2×2 entry game in which for some values of the payoffs, there are three equilibria, 2 in pure strategies and one in mixed strategies.

In statistics, there is a large and ongoing literature on inference in finite mixture models using the likelihood ratio statistic. Most results in this literature focus on deriving the limit distribution when the true parameter is fixed. These results can allow for lack of identification. See, for example, Liu and Shao (2003), Dacunha-Castelle and Gassiat (1999), Azaïs, Gassiat, and Mercadier (2006), Chernoff and Lander (1995), Chen, Chen, and Kalbfleisch (2004) and others. Pointwise asymptotic distribution of LR statistic under a fixed null hypothesis for finite mixture models and closely related regime switching models have also been studied in econometrics; see, e.g., Cho and White (2007).

In econometrics, the literature on uniform approximation and confidence intervals is motivated by situations where the pointwise asymptotic distribution of a test statistic has discontinuities in its limit distribution. See, e.g., Mikusheva (2007), Romano and Shaikh (2012), Andrews and Cheng (2010), Andrews and Cheng (2011), Andrews, Cheng, and Guggenberger (2011) and references cited therein. Our paper's approach to finite mixtures is motivated by this literature. In particular, Andrews and Cheng (2010) and Andrews and Cheng (2011) provide methods for building valid confidence intervals in moment based and likelihood setups in which some parameters can be non-point identified but they assume the true parameters belong to the interior of the parameter space. We follow their approach in that we consider all possible sequences that approach the region of singularity. A key difference between our model and theirs is that in a mixture model, the singularity region can be such that no parameter is point identified and hence methods used in those papers, which require that at least some parameter be point identified need to be modified.

The main practical results of this paper point towards using the parametric bootstrapped profiled likelihood ratio statistic under the null as a way to conduct inference. We show that in almost all cases, this standard parametric bootstrap approximates the asymptotic distribution of the PLR statistic consistently. There are some sequences for which the parametric bootstrap distribution needs to be modified in a way to guarantee the correct coverage. These problem sequences are related to cases where the nuisance parameters are close to the singularity regions and approach this regions at a particular rate as sample size increases.

The paper is organized as follows. First, we derive the asymptotic distribution of the PLR statistic under drifting sequences in Section 2. Section 3 proposes a parametric bootstrap based method to conduct inference and also constructs confidence sets that are uniformly valid. It also presents Monte Carlo simulations to investigate small sample behaviors of the parametric bootstrap. Section 4 applies our methods to the Beauty Contest data. Section 5 discusses extensions and Section 6 briefly concludes. Appendix A contains all the proofs.

2 Asymptotic Distribution of PLR Statistics

In this paper, we derive large sample distribution of the profiled likelihood ratio statistic under drifting sequences. This is meant to highlight how these limits vary as a function of the location of the true parameter relative to the singularity region. Again, let $f_\theta = f(\cdot, \theta)$ and let $f_0 = f(\cdot, \theta_0)$. The mixture distribution for the mixing probability δ is

$$p(X_i; \theta, \delta) = \delta f(X_i, \theta) + (1 - \delta) f(X_i, \theta_0)$$

Here the mixing probability δ is the parameter we are interested in, θ_0 is known, and θ is the unknown (nuisance) scalar parameter that lies in Θ,[2] a compact subset of $\mathbb{R}$. The Profile Likelihood Ratio test statistic (PLR) for testing the null hypothesis $\delta = \delta_0$ is given by

$$PLR(\delta_0) = \sup_{\theta,\delta} l_n(\theta, \delta) - \sup_{\theta} l_n(\theta, \delta_0)$$

where

$$l_n(\theta, \delta) = 2 \sum_{i=1}^{n} \log \left(1 + \delta \frac{f(X_i, \theta) - f(X_i, \theta_0)}{f(X_i, \theta_0)} \right)$$

A $(1-\alpha)$ confidence set for δ will be based on inverting the PLR test statistic, and so it will have the form

$$C_n(\alpha) = \{\delta_0 \in [0, 1] : \ PLR(\delta_0) \leq c_n(\delta_0, \alpha)\}$$

with an appropriate value for $c_n(\delta_0, \alpha)$. The asymptotic coverage probability of this CI is the probability under $p(\delta, \theta)$ that the test statistic is less than the appropriate critical value and its asymptotic size is the limit of the infimum of this probability over all parameters $(\delta, \theta) \in [0, 1] \times \Theta$. When using large sample asymptotics to approximate the small sample distribution of a statistic - which is the object of interest- an issue that arises is whether this approximation is *uniform* in the underlying true density. Heuristically, this asymptotic approximation is uniform, if there exists a sample size, N^* say, that is large enough such that for any $n \geq N^*$ the asymptotic distribution is guaranteed to lie in a small neighborhood of the true density *for any* density in the class of models considered. So, a lack of uniformity means that for any arbitrary n, the asymptotic approximation can be poor for some density in the class. So, uniformity is equivalent to the lack of some common "N^*" beyond which we get good approximations of the true density no matter where this latter might lie. In standard cases, usual continuity of the asymptotic distribution in the underlying true density (and other regularity conditions) guarantee that the convergence is uniform. But, here, this

[2]To simplify the notation, we assume that θ is scalar. However, the results can be extended to cover the case when θ is a vector of parameters. The only thing that changes in this case is the definition of the covariance function of several gaussian processes defined later in the paper when evaluated at $\theta = \theta_0$.

asymptotic distribution changes abruptly depending on how we approach the singularity set, which in this case is the set where there is complete lack of identification of both δ and θ.

2.1 Fixed Asymptotics

In the case with lack of point identification, a version of MLE consistency was first proved by Redner (1981) in which he showed that the MLE converges to some point in the identified set.[3] In particular, Figure 2 below shows the argsup of the sample log likelihood for two different sample realizations each of size $n = 1000$ when $\delta = 0$.

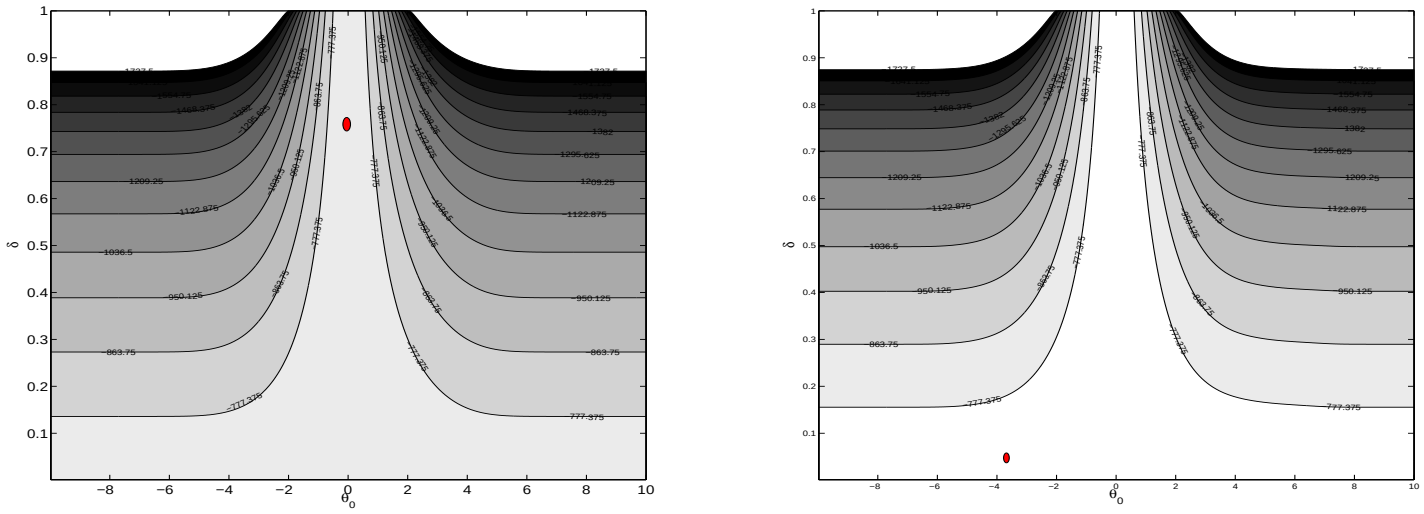

Figure 2: Level sets of two realizations of the expected log likelihood function: the two black dots represent the location of the argsup for that particular sample.

Throughout, we maintain the following assumptions which are commonly made in finite mixture models.

Assumption 2.1 *Let the following hold.*

A1: *(i)* $\Theta \subset \mathbb{R}$ *is compact, and* $\theta_0 \in int(\Theta)$*. (ii) There is a dominating measure* ν *such that for any* $\theta_1, \theta_2 \in \Theta$*:* $f_{\theta_1} = f_{\theta_2}$ ν*-a.e. iff* $\theta_1 = \theta_2$*. (iii)* $\theta \mapsto f(x, \theta)$ *is twice continuously differentiable* ν*-a.e. for any* $x \in \mathcal{X}$*. (iv) There is a function* $B \in (L^2(f_0 \cdot \nu), \|\cdot\|_2)$ *such that* $|f_\theta / f_0|, |f'_\theta / f_0|, |f''_\theta / f_0| \leq B$ *for all* $x \in \mathcal{X}$ *and all* $\theta \in \Theta$*.*

[3]These results were generalized in Chernozhukov, Hong, and Tamer (2007) where a properly specified *set* is shown to converge to this singularity set.

A2: *There exist* $\rho > 0$ *and a constant* M *such that* $E_{\theta,\delta}\left|\left(\frac{f_\theta - f_0}{\delta f_\theta + (1-\delta)f_0}\right)^{2+\rho}\right| < M$ *for all* $\delta \in [0,1]$ *and* $\theta \in \Theta$.

Assumption A1 above ensures that the set $\mathcal{S} = \left\{\frac{(f_\theta - f_0)/f_0}{\|(f_\theta - f_0)/f_0\|_2}, \theta \in \Theta \setminus \{\theta_0\}\right\}$ is Donsker and its closure is compact (see Liu and Shao (2003) for a detailed discussion). In particular, this assumption requires that the parameter space Θ is compact. In the mixture literature, this is an important restriction since without it, the asymptotic distribution of the PLR statistic might become unbounded (see, e.g., Hartigan (1985)). Assumption A2 implies the Lyapounov condition for the sequence $V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{f_{\theta_n} - f_0}{\delta_n f_{\theta_n} + (1-\delta_n)f_0}$, so that the Lindeberg-Lévy CLT for triangular arrays holds for all converging sequences $\{(\theta_n, \delta_n)\}$ in $\Theta \times [0,1]$.

We first state a theorem that provides the asymptotic distribution of the PLR statistic when true $\delta = 0$, i.e., when the true distribution is fixed and equal to f_0. In the main example above, this would be equivalent to having true (δ, θ) be such that $\delta(\theta - \theta_0) = 0$ or (δ, θ) belonging to the singularity set and so θ in this case is not identified. This theorem can be proved via the approach of Liu and Shao (2003) where the PLR is expanded around the true density f_0 in the density space.[4]

Theorem 2.1 *When* $\delta = 0$, *and under Assumption 2.1, the asymptotic distribution of the PLR statistic,* $\sup_{(\delta,\theta)} l_n(\delta,\theta) - \sup_\theta l_n(\theta, 0)$ *is*

$$\sup_{\theta \in \Theta} \max(Z(\theta), 0)^2$$

where $Z(\theta)$ *is a mean zero gaussian process with covariance function* $r(\theta_1, \theta_2)$ *such that for* $\theta_1, \theta_2 \neq \theta_0$,

$$r(\theta_1, \theta_2) = \int \frac{(f_{\theta_1} - f_0)/f_0}{\|(f_{\theta_1} - f_0)/f_0\|_2} \frac{(f_{\theta_2} - f_0)/f_0}{\|(f_{\theta_2} - f_0)/f_0\|_2} f_0 d\nu$$

and

$$r(\theta_1, \theta_0^\pm) = r(\theta_0^\pm, \theta_1) = \pm \int \frac{(f_{\theta_1} - f_0)/f_0}{\|(f_{\theta_1} - f_0)/f_0\|_2} \frac{f_0'/f_0}{\|f_0'/f_0\|_2} f_0 d\nu$$

and

$$r(\theta_0^\pm, \theta_0^\pm) = \int \frac{f_0'/f_0}{\|f_0'/f_0\|_2} \frac{f_0'/f_0}{\|f_0'/f_0\|_2} f_0 d\nu$$

Notice here that the asymptotic distribution depends on the parameter space which is assumed to be compact. Below, we show that the parametric bootstrap in the case when the

[4]We note that the standard Taylor expansions in parameter spaces are not valid here since it is not clear around which parameter vector in the singularity set one should linearize. Instead, by moving to the implied density space, all the parameters in the singularity set correspond to the same density f_0 and hence a functional expansion (around f_0) works.

true model is such that $\delta = 0$ is exact, i.e., its distribution is equal to the distribution of the PLR statistic for all n.

Since our objective is to construct valid confidence interval for δ, the size of such confidence interval is based on the asymptotic behavior of the PLR statistic for all possible true models. And, if the asymptotic distribution is continuous in the underlying true model, then a parametric bootstrap approximation under a fixed model is uniformly consistent. However, as we show below, the asymptotic distribution in the simple mixture model above is not continuous in the underlying DGP, and so the parametric bootstrap might need adjustment. The true models that create problems are drifting true models that approach the singularity set at particular rates. The most problematic sequences are the ones that approach the singularity set at the root n rate. So, the challenge is to adjust the parametric bootstrap in such a way that guarantees coverage against all such drifting sequences. We first derive the asymptotic distribution of the PLR statistic under various kinds of drifting sequences.

2.2 Asymptotics under Drifting Sequences

The Profile Likelihood Ratio test statistic for testing the null hypothesis $\delta = \delta_0$ is

$$PLR(\delta_0) = \sup_{\theta,\delta} l_n(\theta, \delta) - \sup_{\theta} l_n(\theta, \delta_0)$$

The asymptotic size of confidence sets based on inverting the PLR test is determined by its behavior for all possible drifting sequences of true models (θ_n, δ_n). Throughout the rest of the paper, we denote drifting sequences of true models by (θ_n, δ_n), while estimated quantities as $(\tilde{\theta}_n, \tilde{\delta}_n) \in \arg\max_{\theta,\delta} l_n(\theta, \delta)$ and $(\hat{\theta}_n, \delta_n) \in \arg\max_{\theta} l_n(\theta, \delta_n)$. In what follows, we assume that

A3: $\theta_n \to \theta^* \in int(\Theta)$

The assumption A3 restricts the attention only to sequences that belong to the interior of parameter space Θ, and is employed only to simplify the presentation of the results and avoid the discussion of a nuisance parameter being on the boundary [5].

We split all drifting sequences into two main categories:

Definition 2.1 *Define the following two classes of sequences of true models:*

$$\textbf{\textit{Class NI:}}\ \Lambda_{NI} = \{(\theta_n, \delta_n) : \delta_n \|(f_{\theta_n} - f_0)/f_0\|_2 \to 0\}$$
$$\textbf{\textit{Class PI:}}\ \Lambda_{PI} = \{(\theta_n, \delta_n) : \delta_n \|(f_{\theta_n} - f_0)/f_0\|_2 \to d \in (0, \infty)\}$$

[5] Relaxing assumption A3 will add more cases to asymptotic limits in Theorems 2.2 through 2.5 depending on whether the limit belongs to the boundary of Θ or not.

Sequences in Class NI are the sequences that converge to the non-identified singularity set where the true density of the data is $p(\cdot;\theta,0) = p(\cdot;\theta_0,\delta) = f_0$. The other class of sequences, Class PI, contain all models that converge to some point-identified density $p(\cdot;\theta,\delta) \neq f_0$. We also normalize all these sequences by f_0 for convenience (since the formulas below simplify). In addition, we use L_2 norm above since the form of the asymptotic distribution under this norm are familiar (and convenient). To simplify presentation in the rest of the text we denote

$$\eta(\theta_n) \equiv \|(f_{\theta_n} - f_0)/f_0\|_2 = O(1) \;\; and \;\; \eta(\theta^*) \equiv \|(f_{\theta^*} - f_0)/f_0\|_2.$$

2.2.1 Asymptotics for NI Class:

The class NI contains sequences that approach the singularity set at various rates, and so we further split sequences in Class NI into the following three sub-categories (classes):

$$\begin{aligned}
&\textbf{Class NI-0: } \Lambda_{NI}(0) = \{(\theta_n,\delta_n) \in \Lambda_{NI} : \sqrt{n}\delta_n\eta(\theta_n) \to 0\} \\
&\textbf{Class NI-}c\textbf{: } \Lambda_{NI}(c) = \{(\theta_n,\delta_n) \in \Lambda_{NI} : \sqrt{n}\delta_n\eta(\theta_n) \to c \in (0,\infty)\} \\
&\textbf{Class NI-}\infty\textbf{: } \Lambda_{NI}(\infty) = \{(\theta_n,\delta_n) \in \Lambda_{NI} : \sqrt{n}\delta_n\eta(\theta_n) \to \infty\}
\end{aligned} \tag{2.1}$$

Class NI-0 is the class that approach the singularity set at the fastest rate - faster than root n. It includes the model where $\delta_n \equiv \delta = 0$. This is the class of sequences that contain models that are either not point identified, or are weakly identified (in the language of Andrews and Cheng). It turns out that within this class, the asymptotic distribution of the PLR statistic is dominated by the distribution for the case where the true model is such that $\delta_n = 0$ which is given in Theorem 2.1. This distribution can be consistently estimated by the distribution of the parametric bootstrap under the null that $\delta = 0$ (proof below). The class NI-c contains the sequences that converges to the singularity set at just the right rate to create problems because of the presence of the nuisance parameter θ. Depending on the location of θ^*, the limit of θ_n, the asymptotic distribution of PLR differs. Finally, the NI-∞ class contains the sequences that approach the singularity set at a slow rate and so here the distribution of the PLR is standard. We start first with some definitions of gaussian processes that are useful in the limits below.

Let $D(\theta)$ be a zero mean gaussian process with covariance function

$$\rho(\theta_1,\theta_2) = \int \frac{f'_{\theta_1}/f_0}{\|f'_{\theta_1}/f_0\|_2} \frac{f'_{\theta_2}/f_0}{\|f'_{\theta_2}/f_0\|_2} f_0 d\nu$$

Also, let $W(\theta)$ be a zero mean gaussian process with covariance function

$$\omega(\theta_1, \theta_2) = \int \frac{f_{\theta_1} - f_0}{f_0} \frac{f_{\theta_2} - f_0}{f_0} f_0 d\nu$$

and variance $\sigma^2(\theta) = \omega(\theta, \theta)$. In addition, note that

$$W(\theta) = \sigma(\theta) Z(\theta)$$

where $Z(\theta)$ is the zero mean gaussian process with covariance function $r(\theta_1, \theta_2)$ that is defined in Theorem 2.1.

We start by considering sequences in Class NI that are less than $n^{-1/2}$-away from the density f_0. The theorem below gives asymptotic distribution for the PLR test statistic for such sequences.

Theorem 2.2 [**Asymptotic for Class NI-0**] *Let Assumptions A1-A3 hold. Then under* $(\theta_n, \delta_n) \in \Lambda_{NI}(0)$*:*

(i) If $\delta_n \to 0$, $\sqrt{n}\delta_n \to \infty$, *and* $\eta(\theta_n) \to 0$, *then*

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta), 0\}\right)^2 - \left(D(\theta_0)\right)^2$$

(ii) If $\delta_n \to \delta^* \in (0, 1]$ *and* $\sqrt{n}\eta(\theta_n) \to 0$, *then*

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta), 0\}\right)^2 - \left(D(\theta_0)\right)^2$$

(iii) If $\delta_n \to 0$, $\sqrt{n}\delta_n \to \gamma \in (0, \infty)$, *and* $\eta(\theta_n) \to 0$, *then*

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta), 0\}\right)^2 - \sup_\theta \left(2\gamma W(\theta) - \gamma^2 \sigma^2(\theta)\right)$$

(iv) If $\sqrt{n}\delta_n \to 0$, *then*

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta), 0\}\right)^2$$

Remark 2.1 *For sequences in the NI-0 class the following holds: the limit of* $PLR(\delta_n)$ *where* δ_n *and* θ_n *satisfy the conditions in Case (iv) first-order stochastically dominates the limit of* $PLR(\delta_n)$ *for any other sequences in NI-0.*

Remark 2.2 *Notice that in case (iv) above, the limit holds when* $\delta_n = 0$ *which is the case when* δ *is on the boundary of the parameter space and* θ *is not identified.*

Note here that in the first three cases of Theorem 2.2, we have θ_n drifting to θ_0 (possibly at various rates since we are always restricted here to lie in NI-0), but the asymptotic distribution is different depending on whether δ_n is staying sufficiently away from zero (cases (i) and (ii)) or not (case (iii)). Note also that in the cases when δ_n stays away from zero, the asymptotic distribution is the difference between the sup statistic of case (iv) and $D(\theta_0)^2$ which has a chi-squared distribution. Also, if δ_n is converging to zero sufficiently fast (in fact at a faster rate than $\sqrt{n}$), then the asymptotic distribution is the same as if $\delta_n = 0$, *regardless where true* θ *lies.*

Now let's consider the second sub-class in Class NI: the sequences that are exactly $n^{-1/2}$-away from the density f_0. These sequences present problems for the parametric bootstrap.

Theorem 2.3 [Asymptotics for Class NI-c] *Let Assumptions A1-A3 hold. Then under* $(\theta_n, \delta_n) \in \Lambda_{NI}(c)$*, the followings hold.*

(i) If $\delta_n \to 0$, $\sqrt{n}\delta_n \to \infty$ *and* $\eta(\theta_n) \to 0$*, then*

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta) + cr(\theta_0, \theta), 0\}\right)^2 - \left(D(\theta_0) + cr(\theta_0, \theta_0)\right)^2$$

(ii) If $\sqrt{n}\delta_n \to \gamma \in (0, \infty)$ *and* $\eta(\theta_n) \to \eta(\theta^*) > 0$*, then*

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta) + cr(\theta^*, \theta), 0\}\right)^2 - \sup_\theta \left(2\gamma W(\theta) + 2\gamma^2\omega(\theta, \theta^*) - \gamma^2\sigma^2(\theta)\right)$$

As we can see above, the asymptotic distribution for the case when we approach the singularity region at the exact root-n rate depends on c (and on γ and θ^*, since $c = \gamma\sigma(\theta^*)$). Unfortunately, when $\sqrt{n}\delta_n \to \gamma \in (0, \infty)$, this constant c cannot be consistently estimated, since θ^* cannot be consistently estimated under such drifting sequences. The fact that the asymptotic distribution depends on this unknown limit c creates a problem for the parametric bootstrap: the parametric bootstrap would sample data from a mixture with parameters $(\delta, \hat{\theta})$ (where $\hat{\theta}$ is some estimator of θ^*) but $\hat{\theta}$ might not become close at all to θ^* as sample size n increases. Note that in class NI-c case (i), δ_n goes to 0 at a slower than root-n rate (since $\sqrt{n}\delta_n\eta(\theta_n) \to c \in (0, \infty)$ and $\eta(\theta_n) \to 0$), and as a result, we can estimate θ^* consistently, as opposed to case (ii).

Next, we consider the third category of sequences in Class NI: sequences that are more than $n^{-1/2}$-away from the homogeneity density f_0. Sequences in this class are "too far" from the singularity set, even though they drift towards it, and so for practical reasons, PLR behaves as though the true model is away from the singularity set.

Theorem 2.4 [Asymptotics for Class NI-∞] *Let Assumptions A1-A3 hold. Then under* $(\theta_n, \delta_n) \in \Lambda_{NI}(\infty)$ *the followings hold:*

(i) If $\delta_n \to \delta^* \in [0,1)$*, then*

$$PLR(\delta_n) \Rightarrow \chi^2_1$$

(ii) If $\delta_n \to 1$ *and* $\sqrt{n}(1-\delta_n) \to \infty$*, then*

$$PLR(\delta_n) \Rightarrow \chi^2_1$$

(iii) If $\delta_n \to 1$*,* $\sqrt{n}(1-\delta_n) \to \gamma \in [0,\infty)$ *and* $\eta(\theta_n) \to 0$*, then*

$$PLR(\delta_n) \Rightarrow (\max\{N(0,1) - c_0, 0\})^2$$

where $c_0 = \gamma \|f_0'/f_0\|_2 \geq 0$ *and* $\|\cdot\|_2$ *is the norm in* $L^2(f_0 \cdot \nu)$*.*

Notice here that these sequences, though converging to the singularity region, approach that region at such a slow rate that the PLR statistic converges to the standard distribution as though the parameters were point identified. All three asymptotic distributions here are dominated by the asymptotic distribution in (iv) of Theorem 2.2 (Azaïs, Gassiat, and Mercadier (2006) show that due to a covariance structure of the gaussian process $Z(\theta)$, $\sup_\theta(\max\{Z(\theta),0\})^2 = (\sup_\theta Z(\theta))^2$, and for any fixed θ, $Z(\theta)$ is a standard normal random variable).

Finally, we derive the asymptotic distribution of the PLR test statistic for the sequences in the Point Identified class.

Theorem 2.5 [Asymptotic for Class PI] *Let Assumptions A1-A3 hold. Then under* $(\theta_n, \delta_n) \in \Lambda_{PI}$*, we have* $\eta(\theta_n) \to \eta(\theta^*) > 0$*.*

(i) If $\delta_n \to \delta^* \in (0,1)$*, then*

$$PLR(\delta_n) \Rightarrow \chi^2_1$$

(ii) If $\delta_n \to 1$ *and* $\sqrt{n}(1-\delta_n) \to \infty$*, then*

$$PLR(\delta_n) \Rightarrow \chi^2_1$$

(iii) If $\delta_n \to 1$ *and* $\sqrt{n}(1-\delta_n) \to \gamma \in [0,\infty)$*, then*

$$PLR(\delta_n) \Rightarrow (\max\{N(0,1) - c_*, 0\})^2$$

where $c_* = \gamma \|(f_{\theta^*} - f_0)/f_{\theta^*}\|_2^* \geq 0$ *and* $\|\cdot\|_2^*$ *is the norm in* $L^2(f_{\theta^*} \cdot \nu)$*.*

Remark 2.3 *The limit of* $PLR(\delta_n)$ *where* δ_n *and* θ_n *satisfy the conditions in Case (iv) of class NI-0 first-order stochastically dominates the limit of* $PLR(\delta_n)$ *for any sequence in class NI-*∞ *or class PI.*

Note here again that for the class of point identification, we get the standard asymptotic chi squared limits. To conclude, even though in the limit all the sequences of true parameter converge to the singularity region, the asymptotic distribution of PLR varies discontinuously in the underlying parameters. Though we still have pointwise asymptotic limits, this lack of uniform convergence can impact the coverage of confidence regions. We examine this next in the context of the parametric bootstrap.

3 Confidence Sets for Mixing Probability: Uniform Coverage

As we can see, though the PLR statistic has a limit distribution under a given sequence (δ_n, θ_n), this limit distribution is not uniform in the kinds of sequences that we allow. Hence, the question of interest that this paper attempts to answer is to propose confidence intervals that maintain uniform coverage over all sequences. We adopt the parametric bootstrap as a resampling method because intuitively, in a likelihood setup, a parametric bootstrap generates data from a distribution that is closest to the null. Also, a bootstrap based inference is attractive here because the asymptotic distribution of the PLR statistic is complicated and not easy to simulate since it involves the supremum of stochastic processes that might not behave well.

We construct confidence sets for δ via inverting the likelihood ratio test:

$$CS_n(1-\alpha) = \{\delta \in [0,1] : PLR(\delta) \leq c_{n,1-\alpha}(\delta)\} \cup C_0$$

where $C_0 = [0,1]$ if $PLR(0) \leq c_{n,1-\alpha}(0)$ and $C_0 = \varnothing$ otherwise; and $c_{n,1-\alpha}(\delta)$ is the critical value for the significance level α that possibly depends on n and δ. In order to get uniform coverage, we need to find $c_{n,1-\alpha}(\delta)$ such that

$$\liminf_{n\to\infty} \inf_{P_{\delta,\theta} \in \mathcal{P}} P_{\delta,\theta}\{PLR(\delta) \leq c_{n,1-\alpha}(\delta)\} \geq 1-\alpha$$

where $\mathcal{P}$ is the set of mixture densities $p(\cdot, \theta, \delta)$ that obey Assumption 2.1 and where $\theta \in \Theta_r \subset int(\Theta)$ (assumption A3)[6].

As we showed in the previous section, the asymptotic distribution of PLR statistic for various drifting sequences depends on the class the particular sequence belongs to. The theorem below suggest one way to get critical values for the sequences where θ_n can be consistently estimated. This is useful since for these sequences, a straightforward parametric bootstrap provides correct inference.

Theorem 3.1 [**Asymptotics for the Resampling PLR**] *Let $PLR^*(\delta_n)$ be the value of the profile likelihood ratio statistic for testing the null hypothesis that $\delta = \delta_n$ for n independent random draws from the mixture distribution with the density $p^*(x) = \delta_n f(x, \hat{\theta}_n) + (1 - \delta_n) f(x, \theta_0)$ where $\hat{\theta}_n = \arg\sup_{\theta} l_n(\theta, \delta_n)$. Let δ_n be such that $\hat{\theta}_n - \theta_n = o_p(1)$, $\theta_n \to \theta^*$. Also, let $PLR(\delta_n) \Rightarrow \mathbf{Y}(\delta, \theta^*)$ where $\mathbf{Y}(\delta, \theta^*)$ is the corresponding limit in either NI-0(i,ii,iv), NI-c(i), NI-∞ or PI cases above. Then under Assumptions A1-A3, $PLR^*(\delta_n) \Rightarrow \mathbf{Y}(\delta, \theta^*)$.*

This result implies that for certain sequences $\{(\delta_n, \theta_n) : n = 1, 2, \ldots\}$ we can construct critical values based on the random sampling from the mixture with parameters $\delta = \delta_n$ and $\theta_1 = \hat{\theta}_n$. This covers all cases above that either 1) converge to the singularity region at fast rate, 2) are in the singularity region, 3) converge to the singularity region at slow rates, or 4) are away from the singularity region (point identified). This parametric bootstrap is simple to compute.

Remark 3.1 *Notice here that for example, even when some parameters lie on the boundary of the parameter space, such as cases when $\delta = 1$ the parametric bootstrap for the PLR is pointwise consistent. The only condition required in Theorem 3.1 is that $\hat{\theta}_n - \theta_n = o_p(1)$. So, this bootstrap procedure is consistent whether or not true δ lies on the boundary as long as the MLE of θ is consistent.*

3.1 Least Favorable Critical Values for $\sqrt{n}$ Sequences

For sequences in NI-0(iii) and NI-c(ii) classes such that $\sqrt{n}\delta_n \to \gamma \in (0, \infty)$ we have that $\hat{\theta}_n - \theta_n = O_p(1)$ rather than $o_p(1)$. That is, the resampling scheme with $\delta = \delta_n$ and $\theta_1 = \hat{\theta}_n = \arg\sup l_n(\delta_n, \theta)$ may lead to under- or over-coverage. That is, if a sequence of mixing probabilities δ_n goes to zero at the rate $n^{-1/2}$, the resampling scheme in Theorem 3.1 may lead to incorrect coverage. It is possible to get an idea as to how different the

[6]Here we formally consider only uniformity over a subset of the interior of a compact parameters space Θ.

two distributions are in the NI-c case. The asymptotic distribution of the bootstrapped $PLR^*(\delta_n)$ can be characterized in the following way. Let F_{γ,θ^*} be the the distribution of

$$\xi = \arg\sup_{\theta} \left(2\gamma\sigma(\theta)\left[Z(\theta) + \gamma\sigma(\theta^*)r(\theta,\theta^*) - \gamma\sigma(\theta)\right]\right)$$

(here we used the relationship between $W(\theta)$ and $Z(\theta)$: $W(\theta) = \sigma(\theta)Z(\theta)$). Then the asymptotic distribution of $PLR^*(\delta_n)$ for $\sqrt{n}\delta_n \to \gamma > 0$ and $\theta_n \to \theta^* \neq \theta_0$ is for $\xi \sim F_{\gamma,\theta^*}$:

$$\sup_{\theta}\left[\max\{Z(\theta) + \gamma\sigma(\xi)r(\theta,\xi), 0\}\right]^2 - \sup_{\theta}\left[2\gamma\sigma(\theta)\left(Z(\theta) + \gamma\sigma(\xi)r(\theta,\xi) - \frac{1}{2}\gamma\sigma(\theta)\right)\right] \tag{3.1}$$

Compare that to the limit of $PLR(\delta_n)$:

$$\begin{aligned}\sup_{\theta}\left[\max\{Z(\theta) + \gamma\sigma(\theta^*)r(\theta,\theta^*), 0\}\right]^2 - \\ \sup_{\theta}\left[2\gamma\sigma(\theta)\left(Z(\theta) + \gamma\sigma(\theta^*)r(\theta,\theta^*)\sigma(\theta) - \frac{1}{2}\gamma\sigma(\theta)\right)\right]\end{aligned} \tag{3.2}$$

These distributions are different and one can simulate them in particular examples to examine how different they are.

Example 3.1 (Mixture of Normals) *Here we consider the following mixture model:* $\Theta = [-5, 5]$, $f_0(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-1)^2}{2}}$, $f_{\theta^*}(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\theta^*)^2}{2}}$ *and* $\delta_n = \gamma/\sqrt{n}$. *We simulate distributions in (3.1) and (3.2) for the following values:* $n = 100,000$, $\gamma = .1$ *and the following choices for* θ^**:*

(a) $\theta^* = -2.1$*;*

(b) $\theta^* = 0.1$*;*

(c) $\theta^* = 1.6$.

Here one can see that the bootstrap distribution can lie above the asymptotic distribution (Figure 4), below the asymptotic distribution(Figure 5), or even coincide with the asymptotic distribution (Figure 6). In all three cases the distribution of $\hat{\theta}$ *(the constrained argsup of the log-likelihood function) does not place a large probability mass in the area around the true parameter* θ^*.

One way to alleviate this problem is to use the least-favorable critical value combined with pre-testing. Let $c_{1-\alpha}(\delta, \theta^*)$ denote the $(1-\alpha)$ quantile of the asymptotic distribution of PLR statistic for a random sample of size n from the mixture distribution $p(\cdot) = (1-\delta)f(\cdot,\theta_0) + \delta f(\cdot,\theta^*)$. We define the least favorable critical value as

$$c^{LF}_{n,1-\alpha}(\delta) = \sup_{\theta^*} c_{1-\alpha}(\delta,\theta^*)$$

Then by construction (taking into account Remarks 2.1 and 2.3) we have

$$\liminf_{n\to\infty} \inf_{\delta,\theta} P_{\delta,\theta}\{PLR(\delta) \leq c^{LF}_{n,1-\alpha}(\delta)\} \geq 1-\alpha$$

When nuisance parameter θ is multi-dimensional, finding the least-favorable critical value may be a difficult task. One way to approach the optimization problem to obtain the least favorable critical values is to generate randomly a set of M θ's, and under the null of $\delta = \delta_n$, use the parametric bootstrap that draws multiple times a random sample of size n from the mixture with density $\delta_n f(\cdot,\theta) + (1-\delta_n) f(\cdot,\theta_0)$ to get critical values for each of the randomly generated θ's, and then pick the largest critical value. As long as M increases with the sample size n, this procedure can be used to approximate the asymptotic least favorable critical value $c^{LF}_{n,1-\alpha}(\delta)$.

When θ is multidimensional, this procedure might be computationally demanding. We suggest an easy-to-calculate upper bound on the least favorable critical value. First note that the asymptotic distribution of the PLR statistic for a given sequence $\sqrt{n}\delta_n \to \gamma$ and $\theta_n \to \theta^*$ in class NI-c is stochastically dominated by the distribution of $\left(\sup_{\theta\in\Theta}(Z(\theta) + cr(\theta,\theta^*))\right)^2$. Then we can define $(\theta_{max}, \theta^*_{max}) = \arg\max_{\theta,\theta^*} \sigma(\theta^*)|r(\theta,\theta^*)|$. This is a deterministic rather than a stochastic optimization problem, and numerous methods are available to solve it numerically. Then the $(1-\alpha)$ quantile of the distribution of

$$\left(\sup_{\theta\in\Theta}(Z(\theta) + \sqrt{n}\delta_n \sigma(\theta^*_{max})|r(\theta_{max},\theta^*_{max})|)\right)^2$$

bounds the least favorable critical value $c^{LF}_{n,1-\alpha}(\delta)$ from above. However, this bound is not sharp.

An even simpler upper bound can be obtained by noticing that for any $\theta_1, \theta_2 \in \Theta$,

$$|r(\theta_1,\theta_2)| \leq 1$$

Therefore, the $(1-\alpha)$ quantile of the distribution of

$$\left(\sup_{\theta\in\Theta}(Z(\theta) + \sqrt{n}\delta_n \sup_{\theta\in\Theta}\sigma(\theta))\right)^2$$

bounds the least favorable critical value $c^{LF}_{n,1-\alpha}(\delta)$ from above. Again, this upper bound is not sharp, but is relatively easy to calculate, especially when θ is multidimensional. This bound may be too large in many applications, depending on the size of the parameter space Θ.

Example 3.2 (Mixture of Normals, *cont.*) *In the mixture of normals, with* $\Theta = [-5, 5]$*, the least favorable critical value computed for* $\delta_n = 0.1/\sqrt{n}$ *is 4.16. The least-favorable critical value was calculated via sampling from mixture models with* $\theta_0 = 0$ *and* θ^* *chosen on a grid on [5,5] with .1 step, and calculating the asymptotic 5% critical value for each grid point. The least favorable critical value was calculated as the largest asymptotic critical value on the grid. Figure 7 illustrates the behavior of asymptotic critical values as* θ^* *moves away from* $\theta_0 = 0$*.*

Using the same (least favorable) critical value for all candidate values of δ will result in a confidence set that is too conservative (i.e., too large) according to Remarks 2.1 and 2.3. Below, we provide a double pre-testing approach to alleviate this problem. Namely, we identify sequences of mixing probabilities δ that go to zero slower than root-n and sequences that go to zero faster than root-n. For those two types of sequences, we can construct critical values using results in Theorem 3.1 and Theorem 2.2 respectively. However, for the sequences that are exactly root-n away from zero, we still have to rely on the least-favorable critical value.

3.1.1 Pre-Testing Based Critical Values

The LF critical value leads to a conservative coverage for the sequences in NI-∞ and PI classes, as well as some sequences in NI-0 and NI-c classes. In order to improve the LF critical value, we suggest the "pre-testing"[7] for δ_n that goes to zero at $n^{-1/2}$ rate or faster. Let $\{\tau_n\}$ be a deterministic sequence such that $\tau_n \to 0$ and $\sqrt{n}\tau_n \to \infty$. Possible choices are $\tau_n = \log n/\sqrt{n}$ or $\tau_n = \log\log n/\sqrt{n}$. Then we can define

$$c_{n,1-\alpha}(\delta) = \begin{cases} c^{LF}_{n,1-\alpha}(\delta) & \text{if } \delta < \tau_n \\ c^{*}_{n,1-\alpha}(\delta) & \text{if } \delta \geq \tau_n \end{cases} \tag{3.3}$$

where $c^*_{n,1-\alpha}(\delta)$ is the $(1-\alpha)$ quantile of $PLR^*(\delta)$ defined in Theorem 3.1. This way we control for sequences of mixing probabilities that are no more than root-n away from zero.

The above definition of the pre-testing based critical values ignores the fact that for sequences of mixing probabilities that are strictly less than root-n away from zero, the asymptotic distribution of the PLR statistic does not depend on the unknown θ^*. To take that into account, and make the confidence sets less conservative, we can also use pre-testing

[7]This is not an actual pre-testing in the sense that the usual pre-testing based critical values for the test of a null hypothesis may be based on the results of another test ("pre-test"). Rather, in our case "pre-testing" means selecting the appropriate convergence rate category (and therefore critical values) to test a null hypothesis about δ; and this selection rule is based on a hypothesized value of the mixing probability δ, as well as sample size n.

to separate those sequences: let $\tau_n^U, \tau_n^L \to 0$ so that $\sqrt{n}\tau_n^U \to \infty$ and $\sqrt{n}\tau_n^L \to 0$ and define the pre-testing critical value as follows:

$$c_{n,1-\alpha}(\delta) = \begin{cases} c_{n,1-\alpha}^{LF}(\delta) & \text{if } \tau_n^L < \delta < \tau_n^U \\ c_{n,1-\alpha}^{*}(\delta) & \text{if } \delta \geq \tau_n^U \\ c_{n,1-\alpha}^{*}(0) & \text{if } \delta \leq \tau_n^L \end{cases} \tag{3.4}$$

The above procedure is more reasonable than just computing the LF critical values since it zeroes in on sequences that might be in the problem region. We can summarize the uniform coverage results in the following theorem.

Theorem 3.2 [Uniform Coverage with Pre-Testing Based CVs] *Let $PLR(\delta_n)$ be the value of the profile likelihood ratio statistic for testing the null hypothesis that $\delta = \delta_n$, and let $c_{n,1-\alpha}(\delta)$ be defined as in (3.4). Then under Assumptions A1-A3,*

$$\liminf_{n\to\infty} \inf_{P_{\delta,\theta}\in\mathcal{P}} P_{\delta,\theta}\{PLR(\delta) \leq c_{n,1-\alpha}(\delta)\} \geq 1-\alpha.$$

4 Empirical Illustration: Beauty Contest Experiments

In the experimental economics literature, one posits a set of finite behavioral types whereas each member of a population is a type k individual, and a question of interest in this literature is the design of experiments that enables an analyst to infer the proportion of various types using the responses (or behaviors) from the choices made (or the experimental data). See the work of Stahl and Wilson (1995), Bosch-Domenech, Montalvo, Nagel, and Satorra (2002), Costa-Gomes, Crawford, and Broseta (2001), and Kline (2012). In particular, Bosch-Domènech, Montalvo, Nagel, and Satorra (2010) and Bosch-Domenech, Montalvo, Nagel, and Satorra (2002) use experimental data from Beauty Contest games to try to infer the various proportions of behavior. In these games, players guess a number between one and a hundred and the winner is the player whose guess is closest to 2/3 of the mean of the players. Guessing a zero is the unique Nash equilibrium of this game. Below, we plot the histogram of responses from over 8000 data points to this game. We clearly see that the density of the data (See Figure 3) is a mixture of types, and that inference on the number of types is interesting. The PLR based approach to inference based on the parametric bootstrap will be valid whether true parameters are on the boundary and whether the model is point identified which is important in these models.

Similar to Bosch-Domènech, Montalvo, Nagel, and Satorra (2010), we split the outcome space $[0, 100]$ into the following segments: $[0, 14]$, $[14, 30]$ and $[30, 40]$ and consider the following types of players:

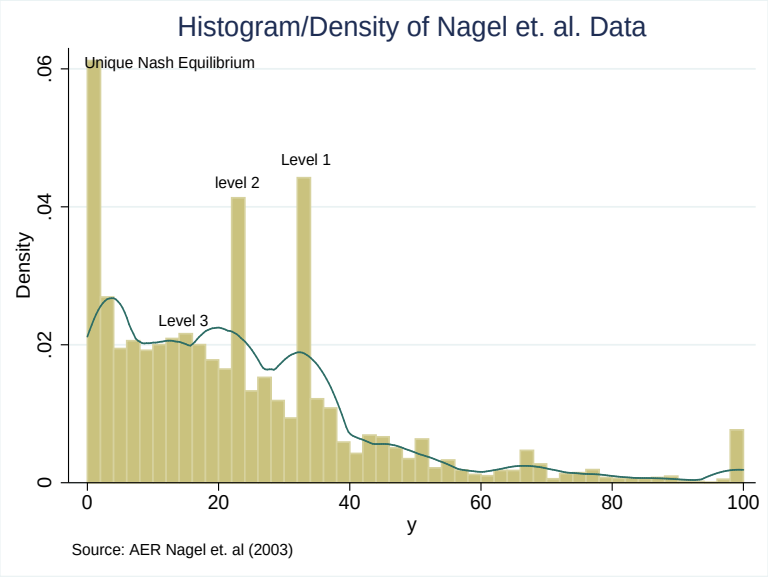

Figure 3: Histogram of responses to Beauty Contest Games: Note that 0 is the unique NE of this game, but other types with various "levels" can provide different guesses. The data is indicative of the presence of a mixture of types (a kernel density estimator is superimposed to highlight the shape of the density).

- $L-0$: those are the players that do not rationalize. We assume that they are uniformly distributed on all three segments: $B(1,1)$.
- $L-1$: players with level-1 rationality. They are distributed according to some Beta distribution on $[30,40]$: $B(\alpha_1,\beta_1)$. That is, on $[30,40]$ we observe a mixture of $L-1$ and $L-0$ players.
- $L-2$: players with level-2 rationality. They are distributed according to some Beta distribution on $[14,30]$: $B(\alpha_2,\beta_2)$ That is, on $[14,30]$ we observe a mixture of $L-2$ and $L-0$ players.
- $L-\infty$: players that play exactly Nash equilibrium (0) or randomize around Nash equilibrium. They are distributed according to some Beta distribution on $[0,14]$: $B(\alpha_\infty,\beta_\infty)$ so that on $[0,14]$ we observe a mixture of $L-\infty$ and $L-0$ players.

According to the results in Bosch-Domènech, Montalvo, Nagel, and Satorra (2010), estimated α_1 and β_1 are both above 350, which makes this case less interesting for our purpose since $B(1,1)$ (uniform) and $B(\alpha_1,\beta_1)$ distributions are too far. Therefore, we focus on $L-2$ and $L-\infty$ cases. Using our procedure, we constructed the following confidence sets for the proportion of non-$L-0$ players: for $L-2$, the set is $[0.4,0.6]$, and for $L-\infty$ it is $[0.3,0.5]$.

5 Extensions: Finite Sample Inference and Projections

It is also possible to conduct *finite sample inference* via simulations following the approach of Dufour (2006). Suppose we want to test the null that $H_0 : \delta = \delta^*$ against the alternative that the null is false, we can also use the profiled likelihood ratio statistic

$$PLR_n(\delta^*) = \sup_{\theta,\delta} l_n(\theta, \delta) - \sup_{\theta} l_n(\theta, \delta^*)$$

The way to compute the critical value via simulation is as follows. Fix a value θ^* for θ. Let $\{X_i^*, i = 1, \ldots, n\}$ be random draws from the density $p^*(\cdot) = \delta^* f(\cdot, \theta^*) + (1 - \delta^*) f(\cdot, \theta_0)$ and define $LR_n^*(\theta^*, \delta^*)$ as the likelihood ratio test statistic calculated based on this sample. Note that the distribution of this LR statistic is fully specified and we can draw as many samples from the density $p^*(.)$ and use the empirical distribution of $LR_n^*(\theta^*, \delta^*)$ to approximate its true distribution assuming that θ^* is the true θ. Let $c_{n,\alpha,R}^*(\theta^*, \delta^*)$ be the $(1 - \alpha)$-quantile of empirical distribution of $LR_n^*(\theta^*, \delta^*)$ based on R random samples of size n. Since, we do not know the true value of θ^*, we can take the largest critical value, i.e., $c_{n,\alpha,R}^*(\delta^*) = \sup_{\theta^*} c_{n,\alpha,R}^*(\delta^*, \theta^*)$. Then, the confidence interval

$$CS_{n,1-\alpha} = \{\delta \in [0, 1] : PLR_n(\delta) \leq c_{n,\alpha,R}^*(\delta)\}$$

will be such that

$$\inf_{P_{\delta,\theta} \in \mathcal{P}} P_{\delta,\theta}\{PLR_n(\delta) \leq c_{n,\alpha,R}^*(\delta)\} = 1 - \alpha$$

as R increases. This is an exact small sample confidence interval. This CS is hard to compute, especially when θ has many components, since the computation will require calculating critical values on the grid in Θ. One might do random draw of θ^* from Θ to simplify the computation.

Projections: Similarly, if we assume that the model is correctly specified. We want to test a simple hypothesis $H_0 : \theta = \theta^*, \delta = \delta^*$ against the alternative H_1 : "H_0 is false" using the likelihood ratio test statistic

$$LR_n(\theta^*, \delta^*) = \sup_{\theta,\delta} l_n(\theta, \delta) - l_n(\theta^*, \delta^*)$$

Let $\{X_i^*, i = 1, \ldots, n\}$ be random draws from the density $p^*(\cdot) = \delta^* f(\cdot, \theta^*) + (1 - \delta^*) f(\cdot, \theta_0)$ and define $LR_n^*(\theta^*, \delta^*)$ as the likelihood ratio test statistic calculated based on this sample. Note that if there is no misspecification and the null hypothesis is true, $LR_n(\theta^*, \delta^*)$ and $LR_n^*(\theta^*, \delta^*)$ have the same distribution for any sample size n. As above, let $c_{n,\alpha,R}^*(\theta^*, \delta^*)$ be

the $(1-\alpha)$-quantile of empirical distribution of $LR_n^*(\theta^*, \delta^*)$ based on R random samples of size n. Then under the null hypothesis,

$$\lim_{R\to\infty} P\{LR_n(\theta^*, \delta^*) \leq c^*_{n,\alpha,R}(\theta^*, \delta^*)\} = 1-\alpha$$

for any sample size n.

Based on this result, one can construct projection-based confidence set for δ in the following way:

$$CS^P_{n,1-\alpha} = \{\delta \in [0,1] : LR_n(\delta, \theta) < c^*_{n,\alpha,R}(\theta, \delta) \text{ for some } \theta \in \Theta\}$$

This projection-based confidence set $CS^P_{n,1-\alpha}$ has uniform coverage by construction (since it is based on a point-wise testing procedure that has exact coverage in finite samples rather than asymptotically), but it is likely to be conservative as compared to an asymptotic CS.

6 Conclusion

This paper examines a canonical finite mixture model and addresses the question of how to build valid confidence sets for the mixing parameters. These confidence sets must remain valid 1) under all sequences of true parameters, 2) and when the true parameters are such that the model is not identified. Using large sample approximations in mixture models where there is reason to believe that the parameters are in -or close to- the singularity regions presents difficulties due to the presence of discontinuities. In addition, we propose a parametric bootstrap that tries to address 1) computational issues with dealing directly with the complicated asymptotic distribution, and 2) the uniformity issues in that the parametric bootstrap is adjusted in cases where we suspect that we are close to the singularity region. The methodology proposed in the paper can be extended to cover the uniformity in nuisance parameter θ over the whole parameter space Θ rather than any subset Θ_r in its interior[8]. Finally, the asymptotic results for the canonical finite mixture model above can be generalized to 2 component mixtures with unknown parameter (θ_0 is unknown) and also to mixtures of three or more distributions. We give in the Appendix a heuristic description of an extension of the above methods to a mixture of two components with unknown parameters (A.2) and to mixtures of three distributions (A.3) where joint confidence intervals are considered.

[8]This will require modifying the parametric bootstrap procedure to take care of discontinuities of asymptotic distribution of PLR statistic when the nuisance parameter approaches the boundary of parameter space. This can be done via pre-testing whether the limit belongs to the boundary or not.

A Appendix

A.1 Proof of Theorems

In this subsection we denote $l_n(\hat{\theta}, \delta_n) = \sup_{\theta} l_n(\theta, \delta_n)$ and $l_n(\tilde{\theta}, \tilde{\delta}) = \sup_{\theta,\delta} l_n(\theta, \delta)$.

Proof. [**Theorem 2.2**]

Case (i): Let $\delta_n \to 0$, $\sqrt{n}\delta_n \to \infty$, and $f_{\theta_n} \to f_0$. Then (by Wong and Shen (1995)) $\|f_{\hat{\theta}} - f_0\|_2 = o_p(1)$, so that θ_n is consistently estimated by $\hat{\theta}$. Also, $\sqrt{n}\delta_n\|(f_{\hat{\theta}} - f_0)/f_0\|_2 = O_p(1)$. Then under $(f_0 \cdot \nu)^{\otimes n}$ and according to Lemma A.1

$$l_n(\hat{\theta}, \delta_n) = 2\sum_{i=1}^{n} \log\left(1 + \delta_n \frac{f_{\hat{\theta}} - f_0}{f_0}\right) \Rightarrow (D(\theta_0))^2$$

Since $[f_0 \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous, we have

$$PLR(\delta_n) \Rightarrow \sup_{\theta} \left(\max\{Z(\theta), 0\}\right)^2 - (D(\theta_0))^2$$

under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$, where the supremum part follows from Azaïs, Gassiat, and Mercadier (2006).

Case (ii): Let $\delta_n \to \delta^* \in (0, 1]$ and $f_{\theta_n} \to f_0$. In this case, $\hat{\theta} \xrightarrow{p} \theta_0$, and

$$l_n(\hat{\theta}, \delta_n) = 2\sum_{i=1}^{n} \log\left(1 + \delta_n \frac{f_{\hat{\theta}} - f_0}{f_0}\right) \Rightarrow (D(\theta_0))^2$$

Since $[f_0 \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous, we have

$$PLR(\delta_n) \Rightarrow \sup_{\theta} \left(\max\{Z(\theta), 0\}\right)^2 - (D(\theta_0))^2$$

under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$.

Case (iii): Now let $\delta_n \to 0$, $\sqrt{n}\delta_n \to \gamma \in (0, \infty)$, and $f_{\theta_n} \to f_0$. In this case $\hat{\theta} - \theta_n = O_p(1)$. Then

$$l_n(\theta, \delta_n) = 2\sum_{i=1}^{n} \delta_n \frac{f_\theta - f_0}{f_0} - \sum_{i=1}^{n} \left(\delta_n \frac{f_\theta - f_0}{f_0}\right)^2 + o_p(1)$$

where $o_p(1)$ is uniform in $\theta \in \Theta$. Then we have under $[f_0 \cdot \nu]^{\otimes n}$

$$l_n(\theta, \delta_n) = 2\gamma W(\theta) - \gamma^2 \sigma^2(\theta) + o_p(1)$$

Since $[f_0 \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous,

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta), 0\}\right)^2 - \sup_\theta \left(2\gamma W(\theta) - \gamma^2 \sigma^2(\theta)\right)$$

under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$.
Processes $D(\theta)$ and $Z(\theta)$ have the following covariance function:

$$Cov(Z(\theta_1), D(\theta_2)) = \int \frac{(f_{\theta_1} - f_0)/f_0}{\|(f_{\theta_1} - f_0)/f_0\|_2} \frac{f'_{\theta_2}/f_0}{\|f'_{\theta_2}/f_0\|_2} f_0 d\nu \text{ if } \theta_1 \neq \theta_0$$

and

$$Cov(Z(\theta_0^{\pm}), D(\theta_2)) = \pm \int \frac{f'_0/f_0}{\|f'_0/f_0\|_2} \frac{f'_{\theta_2}/f_0}{\|f'_{\theta_2}/f_0\|_2} f_0 d\nu$$

Case (iv): Let $\delta_n \to 0$ and $\sqrt{n}\delta_n \to 0$. Since Θ is compact, $\|(f_\theta - f_0)/f_0\|_2$ is bounded uniformly over $\theta \in \Theta$. Therefore, uniformly in $\theta \in \Theta$, $l_n(\theta, \delta_n) = o_p(1)$ under $[f_0 \cdot \nu]^{\otimes n}$. That is, under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$,

$$PLR(\delta_n) \Rightarrow \sup_\theta \left(\max\{Z(\theta), 0\}\right)^2.$$

■

The following lemma is used to prove Theorem 2.2:

Lemma A.1 *Let $\sqrt{n}\delta_n \|(f_{\theta_n} - f_0)/f_0\|_2 \to 0$ and $f_{\theta_n} \to f_0$. If $\hat{\theta} \xrightarrow{p} \theta_0$, then under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$*

$$l_n(\hat{\theta}, \delta_n) = \sum_{i=1}^n \log\left(1 + \delta_n \frac{f_{\hat{\theta}} - f_0}{f_0}\right) \Rightarrow (D(\theta_0))^2$$

Proof. Since $\hat{\theta} - \theta_0 = o_p(1)$, we can use Taylor series expansion around $\theta = \theta_0$: under $[f_0 \cdot \nu]^{\otimes n}$

$$\begin{aligned} l_n(\theta, \delta_n) &= 2\sum_{i=1}^n \log\left(1 + \delta_n \frac{f_\theta - f_0}{f_0}\right) \\ &= 2\sum_{i=1}^n \left(\delta_n \frac{f_\theta - f_0}{f_0}\right) - \sum_{i=1}^n \left(\delta_n \frac{f_\theta - f_0}{f_0}\right)^2 + o_p(1) \\ &= 2\sum_{i=1}^n \delta_n \frac{f'_0(\theta - \theta_0)}{f_0} - \sum_{i=1}^n \left(\delta_n \frac{f'_0(\theta - \theta_0)}{f_0}\right)^2 + o_p(1) \end{aligned}$$

Taking supremum over θ yields

$$l_n(\hat{\theta}, \delta_n) = \frac{\left(\sum_{i=1}^{n} f_0'/f_0\right)^2}{\sum_{i=1}^{n} (f_0'/f_0)^2} + o_p(1)$$

Under $[f_0 \cdot \nu]^{\otimes n}$, $l_n(\hat{\theta}, \delta_n) \Rightarrow (D(\theta_0))^2$, and since $[f_0 \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous, we have $l_n(\hat{\theta}, \delta) \Rightarrow (D(\theta_0))^2$ under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$. ■

Proof. [**Theorem 2.3**]

Case (i): Let $\|(f_{\theta_n} - f_0)/f_0\|_2 \to 0$. In this case (see Wong and Shen (1995)), $\hat{\theta} - \theta_n \xrightarrow{p} 0$ and $\hat{\theta} - \theta_0 \xrightarrow{p} 0$. Then under $[f_0 \cdot \nu]^{\otimes n}$

$$\begin{aligned} l_n(\hat{\theta}, \delta_n) &= 2\sum_{i=1}^{n} \delta_n \frac{f_0'}{f_0}(\hat{\theta} - \theta_0) - \sum_{i=1}^{n} \left(\delta_n \frac{f_0'}{f_0}(\hat{\theta} - \theta_0)\right)^2 + o_p(1) \\ &\Rightarrow (D(\theta_0))^2 \end{aligned}$$

By Le Cam's third lemma, under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$,

$$l_n(\hat{\theta}, \delta_n) \Rightarrow (D(\theta_0) + cr(\theta_0, \theta_0))^2$$

Finally, since $[f_0 \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous,

$$PLR(\delta_n) \Rightarrow \sup_{\theta} \left(\max\{Z(\theta) + cr(\theta_0, \theta), 0\}\right)^2 - (D(\theta_0) + cr(\theta_0, \theta_0))^2$$

Case (ii): Now let $\sqrt{n}\delta_n \to \gamma \in (0, \infty)$ and $f_{\theta_n} \to f_{\theta^*} \neq f_0$. Then under $[f_0 \cdot \nu]^{\otimes n}$, $\hat{\theta} - \theta_n = O_p(1)$. Also, following similar argument as in Case (iii) of Theorem 2.2,

$$\begin{aligned} l_n(\theta, \delta_n) &= 2\sum_{i=1}^{n} \delta_n \frac{f_\theta - f_0}{f_0} - \sum_{i=1}^{n} \left(\delta_n \frac{f_\theta - f_0}{f_0}\right)^2 + o_p(1) \\ &\Rightarrow (2\gamma W(\theta) - \gamma^2 \sigma^2(\theta)) \end{aligned}$$

Since $[f_0 \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous, then according to Le Cam's third lemma, under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$,

$$l_n(\theta, \delta_n) \Rightarrow (2\gamma W(\theta) + 2\gamma c\mu(\theta, \theta^*) - \gamma^2 \sigma^2(\theta))$$

where

$$\mu(\theta, \theta^*) = \int \frac{f_\theta - f_0}{f_0} \frac{(f_{\theta^*} - f_0)/f_0}{\|(f_{\theta^*} - f_0)/f_0\|_2} f_0 d\nu.$$

Note that here $c = \gamma\|(f_{\theta^*} - f_0)/f_0\|_2$. Thus we can also write

$$l_n(\theta, \delta_n) \Rightarrow (2\gamma W(\theta) + 2\gamma^2\omega(\theta, \theta^*) - \gamma^2\sigma^2(\theta)).$$

Finally, we have under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$

$$PLR(\delta_n) \Rightarrow \sup_{\theta} \left(\max\{Z(\theta) + cr(\theta^*, \theta), 0\}\right)^2 - \sup_{\theta} \left(2\gamma W(\theta) + 2\gamma^2\omega(\theta, \theta^*) - \gamma^2\sigma^2(\theta)\right)$$

■

Proof. [**Theorem 2.4**]

Let $Q_n(\theta, \delta) = \frac{1}{n} l_n(\theta, \delta)$ be the normalized objective function. Its expected value $E(Q_n(\theta, \delta))$ is uniquely maximized at (θ_n, δ_n). Conditions (Lipschitz and higher moments) imply that the Lindeberg-Lévy CLT for triangular arrays holds uniformly in (θ, δ), so that under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$

$$\sup_{\theta,\delta} |\sqrt{n}(Q_n(\theta, \delta) - E(Q_n(\theta, \delta)))| = O_p(1),$$

and $\tilde{\theta} - \theta_n = o_p(1)$, $\tilde{\delta} - \delta_n = o_p(1)$, where $Q_n(\tilde{\theta}, \tilde{\delta}) \geq Q_n(\theta, \delta)$ for all $(\theta, \delta) \in \Theta \times [0, 1]$.

Case (i): Let $\theta_n \to \theta^*$ and $\delta_n \to \delta^* \in [0, 1)$ (here $\theta^* = \theta_0$ if $\delta^* > 0$). Taylor series expansion around (θ_n, δ_n) yields:

$$PLR(\delta_n) = \frac{\left(\sum_{i=1}^{n} \frac{(f_{\theta_n} - f_0)/f_0}{1+\delta_n(f_{\theta_n} - f_0)/f_0}\right)^2}{\sum_{i=1}^{n} \left(\frac{(f_{\theta_n} - f_0)/f_0}{1+\delta_n(f_{\theta_n} - f_0)/f_0}\right)^2} + o_p(1)$$

For any (θ, δ), we define

$$Z_I(\theta, \delta) = \begin{cases} \lim_{n\to\infty} \frac{\sum_{i=1}^{n} f_0'/f_0}{\left(\sum_{i=1}^{n} (f_0'/f_0)^2\right)^{1/2}} & \text{if } \theta = \theta_0 \\ \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \frac{(f_\theta - f_0)/f_0}{1+\delta(f_\theta - f_0)/f_0}}{\left(\sum_{i=1}^{n} \left(\frac{(f_\theta - f_0)/f_0}{1+\delta(f_\theta - f_0)/f_0}\right)^2\right)^{1/2}} & \text{if } \theta \neq \theta_0 \end{cases} \tag{A.1}$$

Under $[f_0 \cdot \nu]^{\otimes n}$, $Z_I(\theta_0, \delta^*)$ and $Z_I(\theta^*, 0)$ are standard normal random variables (here we again used Lindeberg-Lévy CLT for triangular arrays). Since $[f_0 \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous, we have that under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$

$$PLR(\delta_n) \Rightarrow \chi_1^2$$

Case (ii): Now let $\delta_n \to 1$ and $\sqrt{n}(1-\delta_n) \to \infty$. Since $\delta_n \to 1$, it must be the case that $\theta_n \to \theta^* = \theta_0$. Then similar to case (i), we have

$$PLR(\delta_n) = \frac{\left(\sum_{i=1}^{n} \frac{(f_{\theta_n}-f_0)/f_0}{1+\delta_n(f_{\theta_n}-f_0)/f_0}\right)^2}{\sum_{i=1}^{n}\left(\frac{(f_{\theta_n}-f_0)/f_0}{1+\delta_n(f_{\theta_n}-f_0)/f_0}\right)^2} + o_p(1) \Rightarrow Z_I(\theta_0,1)^2$$

Case (iii): Let $\sqrt{n}(1-\delta_n) \to \gamma \in [0,\infty)$. Again in this case we have that $\theta_n \to \theta^* = \theta_0$. Taking into account boundary conditions ($\delta_n \to 1$), we have that under $[f_{\theta_0}\cdot\nu]^{\otimes n}$,

$$PLR(\delta_n) = \left(\max\left\{\frac{\sum_{i=1}^{n} \frac{(f_{\theta_n}-f_0)/f_0}{1+\delta_n(f_{\theta_n}-f_0)/f_0}}{\left(\sum_{i=1}^{n}\left(\frac{(f_{\theta_n}-f_0)/f_0}{1+\delta_n(f_{\theta_n}-f_0)/f_0}\right)^2\right)^{1/2}}, 0\right\}\right)^2 + o_p(1)$$

Then under $[f_{\theta_0}\cdot\nu]^{\otimes n}$,

$$PLR(\delta_n) \Rightarrow (\max\{Z_I(\theta_0,1),0\})^2$$

Since $[f_{\theta_0}\cdot\nu]^{\otimes n}$ and $[(f_0+\delta_n(f_{\theta_n}-f_0))\cdot\nu]^{\otimes n}$ are mutually contiguous, we have that under $[(f_0+\delta_n(f_{\theta_n}-f_0))\cdot\nu]^{\otimes n}$,

$$PLR(\delta_n) \Rightarrow (\max\{Z_I(\theta_0,1)-c_0,0\})^2$$

where $c_0 = \gamma\|f_0'/f_0\|_2$, $\|\cdot\|_2$ is the norm in $L^2(f_0\cdot\nu)$, and $Z_I(\theta_0,1)$ is a standard normal random variable. ■

Proof. [**Theorem 2.5**] This is a point identified case except that $\theta^* \neq \theta_0$, so that we can directly use the theory of extremum estimators to show that under $[(f_0+\delta^*(f_{\theta^*}-f_0))\cdot\nu]^{\otimes n}$, $\hat{\theta}-\theta^* = o_p(1)$, $\tilde{\theta}-\theta^* = o_p(1)$, and $\tilde{\delta}-\delta^* = o_p(1)$ is all three cases.

Case (i): Let $\delta_n \to \delta^* \in (0,1)$ and $f_{\theta_n} \to f_{\theta^*} \neq f_0$. Using Taylor series expansion around (δ^*,θ^*) for $l_n(\tilde{\theta},\tilde{\delta})$ and around θ^* for $l_n(\hat{\theta},\delta_n)$, we have

$$l_n(\tilde{\theta},\tilde{\delta}) - l_n(\hat{\theta},\delta_n) = 2\sum_{i=1}^{n}\frac{(f_{\theta^*}-f_0)/f_0}{1+\delta^*(f_{\theta^*}-f_0)/f_0}(\tilde{\delta}-\delta^*) - \sum_{i=1}^{n}\frac{((f_{\theta^*}-f_0)/f_0)^2}{(1+\delta^*(f_{\theta^*}-f_0)/f_0)^2}(\tilde{\delta}-\delta^*)^2 + o_p(1)$$

That is, under $[(f_0+\delta^*(f_{\theta^*}-f_0))\cdot\nu]^{\otimes n}$

$$PLR(\delta_n) = \frac{\left(\sum_{i=1}^{n} \frac{(f_{\theta^*}-f_0)/f_0}{1+\delta^*(f_{\theta^*}-f_0)/f_0}\right)^2}{\sum_{i=1}^{n}\left(\frac{(f_{\theta^*}-f_0)/f_0}{1+\delta^*(f_{\theta^*}-f_0)/f_0}\right)^2} + o_p(1) \Rightarrow \chi_1^2$$

Since $[(f_0 + \delta^*(f_{\theta^*} - f_0) \cdot \nu)]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous,

$$PLR(\delta_n) \Rightarrow \chi^2_1$$

under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$.

Case (ii): Now let $\delta_n \to 1$ and $\sqrt{n}(1 - \delta_n) \to \infty$. Similar arguments can be used to show that under $[(f_0 + \delta_n(f_{\theta_n} - f_0)) \cdot \nu]^{\otimes n}$,

$$PLR(\delta_n) = \frac{\left(\sum_{i=1}^{n} \frac{(f_{\theta_n} - f_0)/f_0}{1+\delta_n(f_{\theta_n} - f_0)/f_0}\right)^2}{\sum_{i=1}^{n} \left(\frac{(f_{\theta_n} - f_0)/f_0}{1+\delta_n(f_{\theta_n} - f_0)/f_0}\right)^2} + o_p(1) \Rightarrow \chi^2_1$$

Case (iii): Let $\sqrt{n}(1 - \delta_n) \to \gamma \in [0, \infty)$. Then taking into account boundary conditions, we have that under $[f_{\theta^*} \cdot \nu]^{\otimes n}$, we have

$$PLR(\delta_n) = \left(\max \left\{ \frac{\sum_{i=1}^{n} \frac{(f_{\theta^*} - f_0)/f_0}{1+\delta_n(f_{\theta^*} - f_0)/f_0}}{\left(\sum_{i=1}^{n} \left(\frac{(f_{\theta^*} - f_0)/f_0}{1+\delta_n(f_{\theta^*} - f_0)/f_0}\right)^2\right)^{1/2}}, 0 \right\} \right)^2 + o_p(1)$$

Then using the definition of $Z_I(\theta, \delta)$ in the proof of Theorem 2.4 in equation (A.1), we have that under $[f_{\theta^*} \cdot \nu]^{\otimes n}$,

$$PLR(\delta_n) \Rightarrow (\max\{Z_I(\theta^*, 1), 0\})^2$$

Under $[f_{\theta^*} \cdot \nu]^{\otimes n}$, $Z_I(\theta^*, 1)$ is a standard normal random variable. Since $[f_{\theta^*} \cdot \nu]^{\otimes n}$ and $[(f_0 + \delta_n(f_{\theta^*} - f_0)) \cdot \nu]^{\otimes n}$ are mutually contiguous, we have that under $[(f_0 + \delta_n(f_{\theta^*} - f_0)) \cdot \nu]^{\otimes n}$,

$$PLR(\delta_n) \Rightarrow (\max\{Z_I(\theta^*, 1) - c_*, 0\})^2$$

where $c_* = \gamma \|(f_{\theta^*} - f_0)/f_{\theta^*}\|_2^*$, $\|\cdot\|_2^*$ is the norm in $L^2(f_{\theta^*} \cdot \nu)$, and $Z_I(\theta^*, 1)$ is a standard normal random variable. ■

Proof. [**Theorem 3.1**] Let's start with the NI-0 class: in cases (i), (ii) and (iv) (where θ_n is consistently estimated by $\hat{\theta}_n$, the limiting distribution of the PLR statistic is continuous in $\theta^* = \lim_{n\to\infty} \theta_n$. Since $\hat{\theta}_n - \theta^* = o_p(1)$, then continuity in θ^* implies that $\sup_u |P\{PLR(\delta_n) \leq u\} - P\{PLR^*(\delta_n) \leq u\}| \to 0$.

For the sequences in the class NI-c (i): if $\|f_{\theta_n} - f_0\|_2 \to 0$, then $\hat{\theta}_n - \theta_0 = o_p(1)$ (see the proof of Theorem 2.3). That is, the parametric bootstrap density also satisfies NI-c (i) condition in probability: $\|f_{\hat{\theta}_n} - f_0\|_2 \xrightarrow{p} 0$, so that $\sup_u |P\{PLR(\delta_n) \leq u\} - P\{PLR^*(\delta_n) \leq u\}| \to 0$.

For all sequences in the class NI-∞: since parameter space Θ is compact, for any sequence in NI-∞ we can choose a convergent sub-sequences such that $\|f_{\theta_n} - f_{\theta^*}\| \to 0$ for some $\theta^* \in \Theta$. For those subsequences, $\hat{\theta}_n - \theta^* = o_p(1)$, so that we also have $\sqrt{n}\delta_n\|(f_{\hat{\theta}_n} - f_0)/f_0\|_2$ is unbounded in probability, and $\delta_n\|(f_{\hat{\theta}_n} - f_0)/f_0\|_2 \xrightarrow{p} \lim_{n\to\infty} \delta_n\|(f_{\theta_n} - f_0)/f_0\|_2$. That is, in this case we have $\sup_u |P\{PLR(\delta_n) \leq u\} - P\{PLR^*(\delta_n) \leq u\}| \to 0$. Finally, the result for the class PI can be proved similarly to that for the class NI-∞. ■

Proof. [**Theorem 3.2**] The result follows from the definition of $c_{n,1-\alpha}(\delta)$, sequences τ_n^L and τ_n^U, and uniform convergence result in Theorem 3.1. ■

A.2 Mixture of Two Components with Unknown Parameters

Suppose now that we don't know the parameters in either of the mixing distributions. That is, the density of X is given by

$$p(\cdot, \theta_1, \theta_2, \delta) = (1-\delta)f(\cdot, \theta_1) + \delta f(\cdot, \theta_2)$$

and both θ_1 and θ_2 are unknown (as well as the mixing probability δ). We assume that $\theta_1, \theta_2 \in \Theta$. In order to distinguish f_{θ_1} from f_{θ_2}, we need to impose some restrictions on the parameter space $\Theta \times \Theta$. For simplicity, let's assume again that $\theta_1, \theta_2 \in \mathbb{R}$, and that $\theta_1 \leq \theta_2$. Identification fails when $\delta = 0$, $\delta = 1$ or $\theta_1 = \theta_2$. The log-likelihood for a set of parameters $(\delta, \theta_1, \theta_2)$ is

$$l_n(\delta, \theta_1, \theta_2) = 2\sum_{i=1}^{n} \log((1-\delta)f(X_i, \theta_1) + \delta f(X_i, \theta_2))$$

and the PLR statistic for testing $H_0 : \delta = \delta_0$ is defined as

$$PRL(\delta_0) = \sup_{\delta, \theta_1 \leq \theta_2} l_n(\delta, \theta_1, \theta_2) - \sup_{\theta_1 \leq \theta_2} l_n(\delta_0, \theta_1, \theta_2)$$

Now there are two points at which the asymptotic distribution of $PLR(\delta_0)$ is non-standard: $\delta_0 = 0$ and $\delta_0 = 1$. Both those cases cannot be distinguished from the case where $\theta_1 = \theta_2$. Following Dacunha-Castelle and Gassiat (1999), we can define the set of extended scores at $\delta = 0$ as

$$S_0 = \left\{ d_{\theta_1,\theta_2} = \frac{(f_{\theta_2} - f_{\theta_1})/f_{\theta_1}}{\|(f_{\theta_2} - f_{\theta_1})/f_{\theta_1}\|_2}, \text{ for } \theta_1 \neq \theta_2 \right\} \bigcup \left\{ d^+_{\theta,\theta} = \frac{+f'_\theta/f_\theta}{\|f'_\theta/f_\theta\|_2}, \text{ for } \theta \in \Theta \right\}$$

Similarly, we can define the set of extended scores at $\delta = 1$ as

$$S_1 = \left\{ d_{\theta_1,\theta_2} = \frac{(f_{\theta_2} - f_{\theta_1})/f_{\theta_2}}{\|(f_{\theta_2} - f_{\theta_1})/f_{\theta_2}\|_2}, \text{ for } \theta_1 \neq \theta_2 \right\} \bigcup \left\{ d^-_{\theta,\theta} = \frac{-f'_\theta/f_\theta}{\|f'_\theta/f_\theta\|_2}, \text{ for } \theta \in \Theta \right\}$$

For any $d \in S_l$, $l = 0, 1$, we can define the gaussian processes

$$Z_l(\theta_1, \theta_2) = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} d(X_i, \theta_1, \theta_2)$$

Then under the null hypothesis $H_0 : \delta = 0$ or $\theta_1 = \theta_2$, the unconstrained maximum likelihood converges in distribution to

$$\sup_{\delta, \theta_1 \leq \theta_2} l_n(\delta, \theta_1, \theta_2) \Rightarrow \sup_{\theta_1, \theta_2} \left(\max\{Z_0(\theta_1, \theta_2), 0\}\right)^2$$

and similarly, under the null hypothesis $H_0 : \delta = 1$ or $\theta_1 = \theta_2$,

$$\sup_{\delta, \theta_1 \leq \theta_2} l_n(\delta, \theta_1, \theta_2) \Rightarrow \sup_{\theta_1, \theta_2} \left(\max\{Z_1(\theta_1, \theta_2), 0\}\right)^2$$

We suggest using similar procedure to construct confidence sets for δ: collect all candidate values of δ_0 for which we fail to reject the null hypothesis $H_0 : \delta = \delta_0$, keeping in mind the following (now two) special cases: if $H_0 : \delta = 0$ or $\theta_1 = \theta_2$ is not rejected, $[0, 1]$ is the CS for δ; and if $H_0 : \delta = 1$ or $\theta_1 = \theta_2$ is not rejected, $[0, 1]$ is the CS for δ. Critical values for the test of $H_0 : \delta = \delta_0$can be obtained by a resampling from the restricted mixture distribution

$$\hat{p}(\cdot) = \delta_0 f(\cdot, \hat{\theta}_1) + (1 - \delta_0) f(\cdot, \hat{\theta}_2)$$

where

$$(\hat{\theta}_1, \hat{\theta}_2) = \arg \sup_{\theta_1 \leq \theta_2} l_n(\delta_0, \theta_1, \theta_2)$$

As before, let $\{X_i^* : i = 1, \ldots, n\}$ be iid random draws from mixture distribution with density $\hat{p}(\cdot; \delta) = (1 - \delta) f(\cdot, \hat{\theta}_1) + \delta f(\cdot, \hat{\theta}_2)$, and let $PLR^*(\delta)$ be the PLR statistic calculated for this sample.

Let's fix θ_1 and θ_2 and consider all possible sequences for δ_n. As in the previous case (with θ_0 known), estimator $\hat{\theta}_1$ is inconsistent when $\delta_n = O(n^{-1/2})$ and $\hat{\theta}_2$ is inconsistent when $1 - \delta_n = O(n^{-1/2})$. That is, to construct uniform confidence set for δ, we can modify the procedure from the previous section in the following way: for a given confidence level $(1 - \alpha)$, the $100(1 - \alpha)\%$ uniform confidence set for δ is

$$CS_n(1 - \alpha) = \{\delta \in [0, 1] : PLR(\delta) \leq c_{n,1-\alpha}(\delta)\} \cup C_0$$

where $C_0 = [0, 1]$ if $PLR(0) \leq c_{n,1-\alpha}(0)$ and $C_0 = \varnothing$ otherwise. Let $\tau_n^U = \log \log n / \sqrt{n}$ and $\tau_n^L = 1/(\sqrt{n} \log \log n)$ and define the pre-tested critical value as follows:

$$c_{n,1-\alpha}(\delta) = \begin{cases} c^*_{n,1-\alpha}(0) & \text{if } \delta \leq \tau_n^L \\ c^{LF}_{n,1-\alpha}(\delta) & \text{if } \tau_n^L < \delta < \tau_n^U \\ c^*_{n,1-\alpha}(\delta) & \text{if } \tau_n^U \leq \delta \leq 1 - \tau_n^U \\ c^{LF}_{n,1-\alpha}(\delta) & \text{if } 1 - \tau_n^U \leq \delta \leq 1 - \tau_n^L \\ c^*_{n,1-\alpha}(1) & \text{if } \delta \geq 1 - \tau_n^L \end{cases} \tag{A.2}$$

where $c^*_{n,1-\alpha}(\delta)$ is the $(1-\alpha)$ quantile of the distribution of $PLR^*(\delta)$, and the least favorable critical value $c^{LF}_{n,1-\alpha}(\delta)$ is defined as before as

$$c^{LF}_{n,1-\alpha}(\delta) = \sup_{\theta_1, \theta_2} c_{n,1-\alpha}(\delta, \theta_1, \theta_2)$$

Here $c_{n,1-\alpha}(\delta, \theta_1, \theta_2)$ is the $(1 - \alpha)$ quantile of the distribution of $PLR(\delta)$ for a random sample of size n from a mixture density $p(\cdot) = (1 - \delta) f(\cdot, \theta_1) + \delta f(\cdot, \theta_2)$.

A.3 Mixture of Three Distributions

Assume now that we have a mixture of three components: one known and two unknown. That is,

$$p(\cdot) = (1 - \delta_1 - \delta_2)f(\cdot, \theta_0) + \delta_1 f(\cdot, \theta_1) + \delta_2 f(\cdot, \theta_2)$$

where θ_0 is known, but θ_1 and θ_2 are unknown. Two out of three distributions in this mixture have unknown parameters, therefore we need to restrict parameter space to $\theta_1, \theta_2 \in \Theta$ and $\theta_1 \leq \theta_2$.

We may be interested in two sets of confidence regions for mixing probabilities δ_1, δ_2: either joint confidence set for (δ_1, δ_2) or in individual confidence sets for δ_1 and δ_2. As with the mixture of two distributions, we construct pointwise testing-based confidence sets using both resampling and least favorable critical values to achieve uniform coverage. We treat two cases separately in more details below.

Throughout this section, we use the following notation:

$$l_n(\delta_1, \delta_2, \theta_1, \theta_2) = 2\sum_{i=1}^{n} \log\left((1 - \delta_1 - \delta_2)f(X_i, \theta_0) + \delta_1 f(X_i, \theta_1) + \delta_2 f(X_i, \theta_2)\right)$$

A.3.1 Joint Confidence Sets

We define the joint profile likelihood ratio statistic for (δ_1, δ_2) as

$$PLR_J(\delta_1, \delta_2) = \sup_{\delta_1, \delta_2; \theta_1 \leq \theta_2} l_n(\delta_1, \delta_2, \theta_1, \theta_2) - \sup_{\theta_1 \leq \theta_2} l_n(\delta_1, \delta_2, \theta_1, \theta_2)$$

and construct the joint confidence set for (δ_1, δ_2) as

$$CS_n(1-\alpha) = \{(\delta_1, \delta_2) \in [0,1] \times [0,1] : PLR(\delta_1, \delta_2)) \leq c_{n,1-\alpha}(\delta_1, \delta_2)\} \cup C_{00} \cup C_{01} \cup C_{10} \quad \text{(A.3)}$$

where

- $C_{00} = [0,1] \times [0,1]$ if $PLR(0,0) \leq c_{n,1-\alpha}(0,0)$ and $C_{00} = \varnothing$ otherwise.
- $C_{01} = \{(\delta_1, \delta_2) \in [0,1] \times [0,1], PLR(0, \delta_2) \leq c_{n,1-\alpha}(0, \delta_2)\}$
- $C_{01} = \{(\delta_1, \delta_2) \in [0,1] \times [0,1], PLR(\delta_1, 0) \leq c_{n,1-\alpha}(\delta_1, 0)\}$

Depending on how close δ_1 or δ_2 are to zero, we use as $c_{n,1-\alpha}(\delta_1, \delta_2)$ either resampling critical value $c^*_{n,1-\alpha}(\delta_1, \delta_2)$ or least-favorable critical value $c^{LF}_{n,1-\alpha}(\delta_1, \delta_2)$. let $(\hat{\theta}_1, \hat{\theta}_2) = \arg\sup_{\theta_1 \leq \theta_2} l_n(\delta_1, \delta_2, \theta_1, \theta_2)$. The resampling critical value $c^*_{n,1-\alpha}(\delta_1, \delta_2)$ is the $(1-\alpha)$ quantile of the distribution of $PLR^*_J(\delta_1, \delta_2)$ based on the iid random sample $\{X^*_i : i = 1, \ldots, n\}$ from the mixture distribution $\hat{p} = (1 - \delta_1 - \delta_2)f_0 + \delta_1 f_{\hat{\theta}_1} + \delta_2 f_{\hat{\theta}_2}$.

Table 1: Critical values for the joint CS: $c_{n,1-\alpha}(\delta_1, \delta_2) =$

	$\delta_2 \leq \tau^L_{2,n}$	$\tau^L_{2,n} < \delta_2 \leq \tau^U_{2,n}$	$\tau^U_{2,n} < \delta_2$
$\delta_1 \leq \tau^L_{1,n}$	$c^*_{n,1-\alpha}(0,0)$	$c^{LF,J}_{n,1-\alpha}(0,\delta_2)$	$c^*_{n,1-\alpha}(0,\delta_2)$
$\tau^L_{1,n} < \delta_1 \leq \tau^U_{1,n}$	$c^{LF,J}_{n,1-\alpha}(\delta_1,0)$	$c^{LF,J}_{n,1-\alpha}(\delta_1,\delta_2)$	$c^{LF,1}_{n,1-\alpha}(\delta_1,\delta_2)$
$\tau^U_{1,n} < \delta_1$	$c^*_{n,1-\alpha}(\delta_1,0)$	$c^{LF,2}_{n,1-\alpha}(\delta_1,\delta_2)$	$c^*_{n,1-\alpha}(\delta_1,\delta_2)$

The joint least favorable critical value $c^{LF,J}_{n,1-\alpha}(\delta_1, \delta_2)$ is defined as

$$c^{LF,J}_{n,1-\alpha}(\delta_1, \delta_2) = \sup_{\theta_1,\theta_2} c_{n,1-\alpha}(\delta_1, \delta_2, \theta_1, \theta_2)$$

where $c_{n,1-\alpha}(\delta_1, \delta_2, \theta_1, \theta_2)$ is the $(1-\alpha)$ quantile of the distribution of $PLR(\delta_1, \delta_2)$ for a random sample of size n from a mixture density $p = (1-\delta_1-\delta_2)f_0 + \delta_1 f_{\theta_1} + \delta_2 f_{\theta_2}$.

The partial least favorable critical value $c^{LF,j}_{n,1-\alpha}(\delta_1, \delta_2)$ for $j = 1, 2$ is defined as

$$c^{LF,j}_{n,1-\alpha}(\delta_1, \delta_2) = \sup_{\theta_1} c^{j,*}_{n,1-\alpha}(\delta_1, \delta_2, \theta_j)$$

where $c^{j,*}_{n,1-\alpha}(\delta_1, \delta_2, \theta_j)$ is the $(1-\alpha)$ quantile of the distribution of $PLR(\delta_1, \delta_2)$ for a random sample of size n from a mixture density $p = (1-\delta_1-\delta_2)f_0 + \delta_1 f_{\theta_j} + \delta_2 f_{\hat{\theta}_{(-j)}}$.

The choice of $c_{n,1-\alpha}(\delta_1, \delta_2)$ in (A.3) is summarized in Table 1 for some sequences $\tau^L_{j,n}, \tau^U_{l,n} \to 0$ such that $\sqrt{n}\tau^L_{j,n} \to 0$ and $\sqrt{n}\tau^U_{j,n} \to \infty$ for $j = 1, 2$.

A.4 Figures

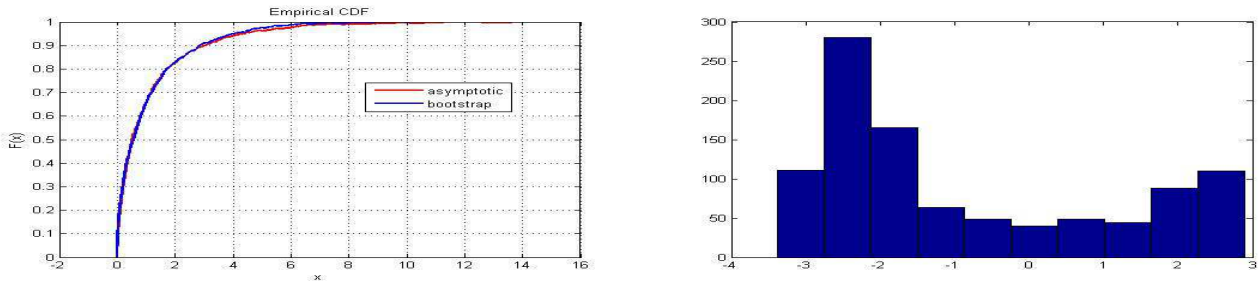

Figure 4: Asymptotic and Bootstrap distributions of PLR statistic (left); and distribution of $\hat{\theta}$ (right) for $\theta^* = -2.1$.

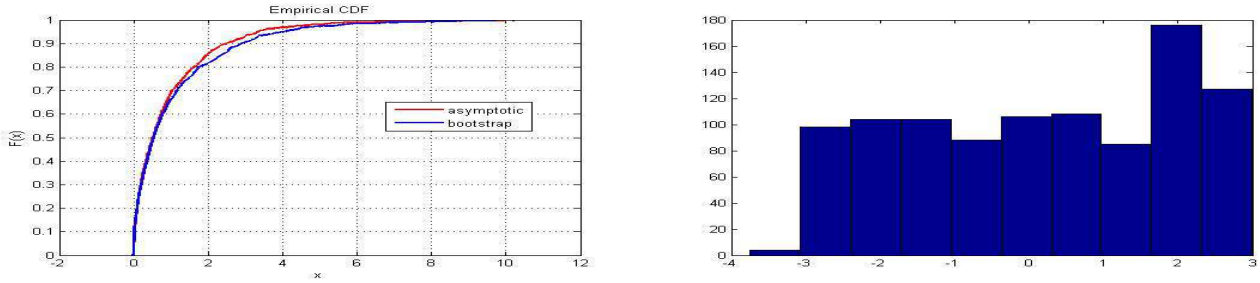

Figure 5: Asymptotic and Bootstrap distributions of PLR statistic (left); and distribution of $\hat{\theta}$ (right) for $\theta^* = 0.1$.

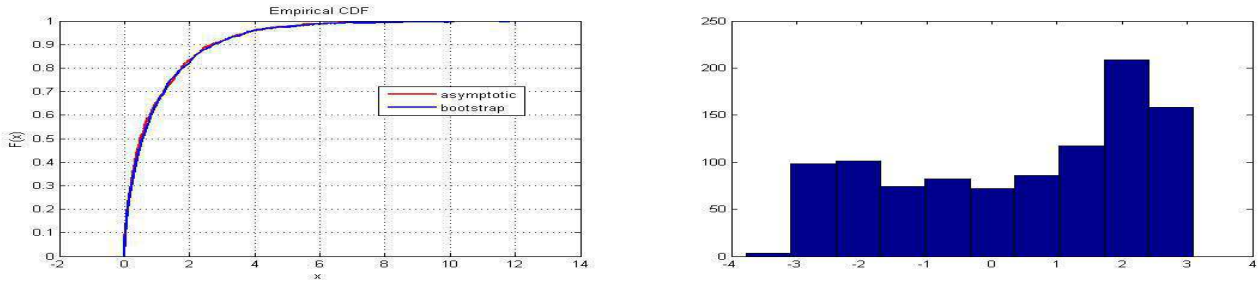

Figure 6: Asymptotic and Bootstrap distributions of PLR statistic (left); and distribution of $\hat{\theta}$ (right) for $\theta^* = 1.6$.

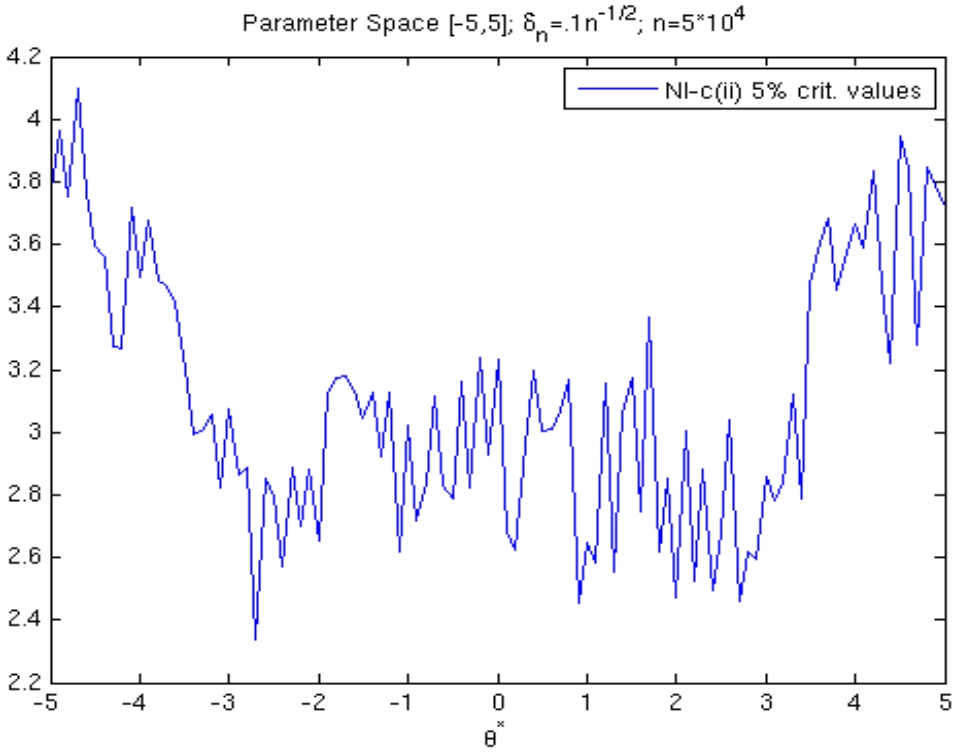

Figure 7: Asymptotic critical values.

References

Andrews, D., and X. Cheng (2010): "Estimation and inference with weak, semi-strong, and strong identification," forthcoming, Econometrica.

——— (2011): "Maximum likelihood estimation and uniform inference with sporadic identification failure," Yale Working Paper.

Andrews, D., X. Cheng, and P. Guggenberger (2011): "Generic results for establishing the asymptotic size of confidence sets and tests," Cowles Foundation Discussion Paper.

Azaïs, J., É. Gassiat, and C. Mercadier (2006): "Asymptotic distribution and local power of the log-likelihood ratio test for mixtures: bounded and unbounded cases," *Bernoulli*, 12(5), 775–799.

Berry, S., and E. Tamer (2006): "Identification in Models of Oligopoly Entry," in *Advances in economics and econometrics: theory and applications, ninth World Congress*, vol. 2, p. 46. Cambridge Univ Pr.

Bosch-Domenech, A., J. Montalvo, R. Nagel, and A. Satorra (2002): "One, two,(three), infinity,...: Newspaper and lab beauty-contest experiments," *The American Economic Review*, 92(5), 1687–1701.

Bosch-Domènech, A., J. Montalvo, R. Nagel, and A. Satorra (2010): "A finite mixture analysis of beauty-contest data using generalized beta distributions," *Experimental Economics*, 13(4), 461–475.

Chen, H., J. Chen, and J. Kalbfleisch (2004): "Testing for a finite mixture model with two components," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1), 95–115.

Chernoff, H., and E. Lander (1995): "Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial," *Journal of Statistical Planning and Inference*, 43(1-2), 19–40.

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

Cho, J. S., and H. White (2007): "Testing for regime switching," *Econometrica*, 75(6), 1671–1720.

Costa-Gomes, M., V. Crawford, and B. Broseta (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69(5), 1193–1235.

Dacunha-Castelle, D., and E. Gassiat (1999): "Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes," *The Annals of Statistics*, 27(4), 1178–1209.

Dufour, J. (2006): "Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics," *Journal of Econometrics*, 133(2), 443–477.

Hartigan, J. (1985): "A failure of likelihood asymptotics for normal mixtures," in *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, vol. 2, pp. 807–810. Wadsworth, Belmont, CA.

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

Kline, B. (2012): "The accuracy of Nash equilibrium in experiments," University of Texas - Austin Working Paper.

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

McLachlan, G., and D. Peel (2000): *Finite mixture models*, vol. 299. Wiley-Interscience.

Mikusheva, A. (2007): "Uniform inference in autoregressive models," *Econometrica*, 75(5), 1411–1452.

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

Romano, J. P., and A. M. Shaikh (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *arXiv preprint arXiv:1204.2762*.

Stahl, D., and P. Wilson (1995): "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10(1), 218–254.

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