**Karun Adusumilli amd Taisuke Otsu**

# Empirical likelihood for random sets

## Article (Accepted version)
## (Refereed)

**Original citation:** Adusumilli, Karun and Otsu, Taisuke (2017) *Empirical likelihood for random sets.* Journal of the American Statistical Association . pp. 1-12. ISSN 0162-1459
DOI: 10.1080/01621459.2016.1188107

# EMPIRICAL LIKELIHOOD FOR RANDOM SETS

KARUN ADUSUMILLI AND TAISUKE OTSU

Abstract. In many statistical applications, the observed data take the form of sets rather than points. Examples include bracket data in survey analysis, tumor growth and rock grain images in morphology analysis, and noisy measurements on the support function of a convex set in medical imaging and robotic vision. Additionally, in studies of treatment effects, researchers often wish to conduct inference on nonparametric bounds for the effects which can be expressed by means of random sets. This article develops concept of nonparametric likelihood for random sets and its mean, known as the Aumann expectation, and proposes general inference methods by adapting the theory of empirical likelihood. Several examples, such as regression with bracket income data, Boolean models for tumor growth, bound analysis on treatment effects, and image analysis via support functions, illustrate the usefulness of the proposed methods.

## 1. Introduction

In many statistical applications, the observed data take the form of sets rather than points. For example, in survey analysis, we often observe bracket data instead of precise measurements. In mathematical morphology, geostatistics, and particle statistics, the observations often take the form of two or three dimensional sets reflecting models for tumor growth or sand rock grains (e.g., Cressie and Hulting, 1992, and Stoyan, 1998, for a review). Also, in the context of medical imaging and robotic vision, researchers sometimes need to infer a convex set from noisy measurements of its support function (Fisher *et al.*, 1997). Furthermore, in studies of treatment effects (e.g., Balke and Pearl, 1997, and

The authors would like to thank Hiroaki Kaido and the Associate Editor for helpful comments. The comments by anonymous referees were substantially helpful for revising the paper. Financial support from the ERC Consolidator Grant (SNP 615882) is gratefully acknowledged (Otsu).

Horowitz and Manski, 2000), researchers often wish to conduct statistical inference on nonparametric bounds for the average treatment effects which can be expressed by means of random sets, as shown in Beresteanu, Molchanov and Molinari (2012).

In this article, we develop a nonparametric likelihood concept for the Aumann expectation of a random sample of convex sets - this is a generalization of the conventional mathematical expectation to random sets - and propose general inference methods by adapting the theory of empirical likelihood (Owen, 2001). In particular, by relying upon the isomorphism between a convex set and its support function, we convert the testing problem on the random set to one on its support function which implies a continuum of moment constraints indexed by the direction of the support function. Based on this conversion, we construct two nonparametric likelihood statistics for testing the moment constraints which we term the marked and sieve empirical likelihood statistics. We study the asymptotic properties of these statistics and describe how to compute critical values for testing. Moreover, to enhance the applicability of our methods, we also discuss testing directed hypotheses and projections, along with situations where the random set of interest is not directly observable due to nuisance parameters to be estimated and where inference is based on noisy measurements of the support function.

We demonstrate the usefulness of the proposed methods by four numerical examples. First, we consider the setup of best linear prediction with interval dependent variables. In this case, the set of all possible coefficients for the best linear predictor is characterized by an Aumann expectation involving the interval data. We illustrate our empirical likelihood methods via inference on the parameters for the best linear predictor of interval wages given years of education using the Current Population Survey (CPS) data. Second, we consider a Boolean model for tumor growth studied by Cressie and Hulting (1992) and numerically evaluate the marked and sieve empirical likelihood tests. Third, we employ the empirical example in Balke and Pearl (1997) on the treatment effect of Vitamin A supplementation under imperfect compliance to study the numerical performance of our empirical likelihood based inference on the bounds of the average treatment effect. Finally, based on Fisher *et al.* (1997), we study the problem of testing the shape of a convex set based on noisy measurements of its support function; the results are provided in the web appendix. Both parameter hypothesis and goodness-of-fit testing problems are investigated. In all of the examples, the proposed empirical likelihood tests perform well in terms of size and power.

After early developments in e.g., Kendall (1974) and Matheron (1975), the literature on the probabilistic and statistical theory of random sets is steadily growing (see, Molchanov, 2005, for a modern and comprehensive treatment of random set theory). Most of the statistical literature on random sets focuses on inference via capacity functionals (e.g., Cressie and Hulting, 1992) and support functions (e.g., Fisher *et al.*, 1997) which provide equivalent characterizations of random sets. The population mean of random sets is typically characterized by the so-called Aumann expectation. Beresteanu and Molinari (2008) developed a Wald type test for the Aumann mean of random sets. This paper introduces a nonparametric likelihood-based approach for inference on the Aumann expectation by modifying the empirical likelihood method. Thus, this paper also contributes to the literature on empirical likelihood (see Owen, 2001, for a review) by extending its scope to random sets rather than points. To establish the asymptotic theory, we adapt the theoretical results developed in Hjort, McKeague and van Keilegom (2009) to our context.

Recently, applications of random set methods have been discussed in the context of partial identification and inference in econometrics; see Molchanov and Molinari (2014) for a review of such applications, Tamer (2010) for a review of partial identification in econometrics, and Manski (2003) for a thorough treatment of partial identification. Partial identification concerns the situation wherein a parameter of interest is not point identified but identified only as a set. This could be because of limitations in the data, e.g. interval or categorical data, or because the theoretical models do not provide enough restrictions to identify a unique value for the parameter, e.g. game theoretic models with multiple equilibria. In this context Balke and Pearl (1997) and Horowitz and Manski (2000) made fundamental contributions to partial identification of treatment effects and probability distributions with missing data, respectively. However, these papers did not connect the inference problems on the identified sets to random set theory. Beresteanu and Molinari (2008) were the first to employ random set methods to conduct estimation and inference for partially identified models.

An important application of random set theory is in the context of inference for parameters characterized by moment inequalities. In this setup, the parameters are typically partially identified, and thus the aim is to propose a confidence region that covers the identified set. Examples of this strand of literature include Chernozhukov, Kocatulum and Menzel (2015), Kaido (2012), and Kaido and Santos (2014) among others. See also Andrews and Shi (2015) for an extension to conditional moment inequalities. On the other hand, Canay (2010) developed an empirical likelihood-based inference method for moment inequality models using "standard" probability theory. Our paper is the first to to bring together random set theory and empirical likelihood. Although sharing applications with the moment inequality setup, our approach, which is based on random sets as observations, is fundamentally different. Indeed, there are situations where the moment inequality setup is not directly applicable unlike ours (e.g., the Boolean model and image analysis via support function), and vice versa. In addition, the focus of our paper is on testing, which may have other uses over and above the construction of confidence regions (cf. the Boolean model example). Closer to our setup, Beresteanu and Molinari (2008) were the first to consider tests for expectations of general random sets. Bontemps, Magnac and Maurin (2012) and Chandrasekhar *et al.* (2012) obtained related inferential results in the context of best linear predictors for set identified functions under a variety of extensions but did not consider other formulations of random sets.

This article is organized as follows. Section 2 introduces the basic setup and presents two inference approaches, the marked and sieve empirical likelihood methods. Section 3 discusses various extensions of these approaches for wider applicability. In Section 4, numerical examples are provided. Assumptions and some definitions are presented in the Appendix. All proofs and additional simulation results are contained in the web appendix.

## 2. Methodology

Suppose we observe a set-valued random variable (SVRV) $X : \Omega \mapsto \mathbb{K}^d$, where $\mathbb{K}^d$ is the collection of all non-empty compact and convex subsets of the Euclidean space $\mathbb{R}^d$. The collection $\mathbb{K}^d$ is endowed with the Hausdorff norm defined as $\|A\|_H = \sup\{\|a\| : a \in A\}$ for every set $A$, where $\|\cdot\|$ is the Euclidean norm. Let $\mu$ denote some underlying probability measure on $\Omega$. The mean of the SVRV $X$ is characterized by the Aumann expectation

$$\mathbb{E}[X] = \left\{ \int_\Omega x d\mu : x \in \{x(\omega) \in X(\omega) \text{ a.s. and } \int_\Omega \|x\| \, d\mu < \infty\} \right\},$$

(see, Molchanov, 2005, for details). We restrict our attention to compact and convex valued SVRVs; however, similar results hold for general compact sets since $\mathbb{E}[X] = \mathbb{E}[\mathrm{co}(X)]$ for compact valued $X$ if $\mu$ is non-atomic, with $\mathrm{co}(X)$ denoting the convex hull operation on $X$ (Molchanov, 2005, p. 154). A fundamental statistical question is to test hypotheses on the Aumann expectation of the form:

$$\mathrm{H}_0 : \mathbb{E}[X] = \Theta_0(\nu) \text{ vs. } \mathrm{H}_1 : \mathbb{E}[X] \neq \Theta_0(\nu), \tag{1}$$

based on a random sample of SVRVs $\{X_1, \ldots, X_n\}$, where $\Theta_0(\nu)$ is a hypothetical set that may depend on real-valued nuisance parameters $\nu \in \mathbb{R}^r$. In general, there is no restriction on the relationship between the dimension $d$ of $X$ and $r$ of $\nu$.

To test the null hypothesis $\mathrm{H}_0$, we focus on the dual representation of convex sets by their support functions. Let $\langle \cdot, \cdot \rangle$ denote the inner product and $\mathbb{S}^d$ the unit sphere in $\mathbb{R}^d$. The support function of a set $A \in \mathbb{K}^d$ is defined as $s(A, p) = \sup_{x \in A} \langle p, x \rangle$ for $p \in \mathbb{S}^d$. If $X$ is integrably bounded, the testing problem in (1) is equivalent to (Molchanov, 2005, p. 157)

$$\mathrm{H}_0 : E[s(X, p)] = s(\Theta_0(\nu), p) \text{ for all } p \in \mathbb{S}^d \text{ vs. } \mathrm{H}_1 : E[s(X, p)] \neq s(\Theta_0(\nu), p) \text{ for some } p \in \mathbb{S}^d, \tag{2}$$

where $E[\cdot]$ is the ordinary mathematical expectation with respect to $\mu$. Therefore, inference on the Aumann mean of the random set is equivalent to inference on the support function (or continuum of moment restrictions over $p \in \mathbb{S}^d$). Since this is a testing problem for infinite dimensional parameters without any parametric distributional assumptions on the population $\mu$, it is of interest to develop a nonparametric likelihood inference method. In particular, we adopt the empirical likelihood approach (Owen, 2001) to our testing problem.

2.1. **Marked empirical likelihood.** We now introduce the first empirical likelihood approach to test the hypothesis in (1) for the Aumann expectation of random sets. We assume that a consistent estimator $\hat{\nu}$ for the nuisance parameters $\nu$ is available. Typically

$\nu$ is a smooth function of population moments which can be estimated by the method of moments.

One method to construct a nonparametric likelihood function to test $\mathrm{H}_0$ in (1) is to fix a direction $p \in \mathbb{S}^d$ for the support function defining the equivalent form of $\mathrm{H}_0$ in (2) and employ the empirical likelihood approach. For given $p$, the marked empirical likelihood function under the restriction $E[s(X,p)] = s(\Theta_0(\nu), p)$ is given by

$$\ell_n(p) = \max\left\{ \prod_{i=1}^n n w_i \middle| \sum_{i=1}^n w_i s(X_i, p) = s(\Theta_0(\hat{\nu}), p),\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}. \tag{3}$$

In practice, $\ell_n(p)$ can be computed from its dual form based on the Lagrange multiplier method, that is

$$\ell_n(p) = \prod_{i=1}^n \frac{1}{1 + \lambda\{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}}, \tag{4}$$

where $\lambda$ solves the first-order condition $\sum_{i=1}^n \frac{s(X_i,p) - s(\Theta_0(\hat{\nu}),p)}{1+\lambda\{s(X_i,p)-s(\Theta_0(\hat{\nu}),p)\}} = 0$. Since the direction $p$ is given, the object $\ell_n(p)$ imposes only a single restriction implied from the null $\mathrm{H}_0$. In order to guarantee consistency against any departure from $\mathrm{H}_0$, we need to assess the whole process $\{\ell_n(p) : p \in \mathbb{S}^d\}$ over the range of $\mathbb{S}^d$. Taking the supremum over $p$ leads to the Kolmogorov-Smirnov type test statistic

$$K_n = \sup_{p \in \mathbb{S}^d} \{-2 \log \ell_n(p)\}.$$

Suppose there exists a function $G(p; \nu)$ continuous in $p \in \mathbb{S}^d$ such that

$$\sup_{p \in \mathbb{S}^d} |s(\Theta_0(\hat{\nu}), p) - s(\Theta_0(\nu), p) - G(p; \nu)'(\hat{\nu} - \nu)| = o_p(n^{-1/2}). \tag{5}$$

In Section 4.1, we provide an example of $G(p;\nu)$ for the case of the best linear prediction with an interval valued dependent variable. The asymptotic properties of $K_n$ are summarized in the following theorem.

**Theorem 1.** *Under Assumption M in the Appendix, it holds*

$$K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\{Z(p)-G(p;\nu)'Z_1\}^2}{\mathrm{Var}(s(X,p))}, \quad \textit{under } \mathrm{H}_0, \tag{6}$$

*where* $(Z(p),Z_1')' \sim N(0,V(p))$ *and* $V(p)$ *is the limiting covariance matrix of* $(n^{-1/2}\sum_{i=1}^n\{s(X_i,p)-E[s(X,p)]\}, \sqrt{n}(\hat{\nu}-\nu)')'$. *In addition,* $K_n$ *diverges to infinity under* $\mathrm{H}_1$.

By a slight modification of the proof, we can also show that under the local alternative

$$\mathrm{H}_{1n} : E[s(X,p)] = s(\Theta_0(\nu),p) + n^{-1/2}\eta(p) \text{ over } p\in\mathbb{S}^d,$$

for some continuous function $\eta$, the marked empirical likelihood statistic satisfies $K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\{Z(p)-G(p;\nu)'Z_1+\eta(p)\}^2}{\mathrm{Var}(s(X,p))}$. Therefore, the test statistic $K_n$ has non-trivial local power against a local alternative at the parametric rate.

One major advantage of the conventional empirical likelihood approach is that it yields an asymptotically pivotal statistic even for nonparametric objects of interest under complicated data structures. However, the proposed statistic $K_n$ (or other statistics constructed from the process $\{\ell_n(p) : p\in\mathbb{S}^d\}$) does not share such attractiveness, and its limiting distribution contains several unknowns to be estimated. To deal with this problem, Section 2.1.1 proposes a bootstrap procedure to approximate the null distribution of $K_n$. In Section 2.2, we develop an alternative test statistic which is asymptotically pivotal (but requires a choice of a tuning parameter). In the current setup, we are not aware of any test statistic which is both asymptotically pivotal and free from tuning parameters.

We note that lack of pivotalness of process-based tests emerges commonly in the context of goodness-of-fit testing (e.g., Stute, 1997). In the literature on empirical likelihood, Chan *et al.* (2009) propose an integral version of the empirical likelihood statistic to test hypotheses on Lévy processes via characteristic functions and derive a non-pivotal limiting distribution; this is approximated by a bootstrap procedure due to its complicated form. Li (2003) obtained similar results for an empirical likelihood test of survival data. Furthermore, Hjort, McKeague and van Keilegom (2009) provided various extensions of empirical likelihood to the cases of (infinite-dimensional) nuisance parameters and growing numbers of estimating equations. They argued that the empirical likelihood statistic is not necessarily pivotal but can be approximated by bootstrap methods.

Since the marked empirical likelihood statistic is not asymptotically pivotal, one may seek to employ alternative likelihood concepts. For instance, we can generate the likelihood process from the Euclidean likelihood (Owen, 2001, Section 3.15):

$$L_n^E(p) = \max\left\{ -\frac{1}{2}\sum_{i=1}^n (nw_i - 1)^2 \,\middle|\, \sum_{i=1}^n w_i s(X_i, p) = s(\Theta_0(\hat{\nu}), p),\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\},$$

whose dual form is explicitly given by

$$-2L_n^E(p) = \frac{\left(\sum_{i=1}^n \{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}\right)^2}{\sum_{i=1}^n \{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}^2},$$

for each $p$. Inspection of the proof of Theorem 1 shows that $L_n^E(p)$ is asymptotically equivalent to $\log \ell_n(p)$ for each $p$ and the test statistic $K_n^E = \sup_{p\in\mathbb{S}^d}\{-2L_n^E(p)\}$ obeys the same limiting distribution as $K_n$. One practical advantage of the Euclidean likelihood-based statistic $K_n^E$ over $K_n$ is that $K_n^E$ does not require a numerical search for the Lagrange multiplier $\lambda$ as in (4).

2.1.1. *Bootstrap calibration.* The limiting null distribution of the process $\{\ell_n(p) : p \in \mathbb{S}^d\}$ is generally difficult to approximate as it contains parameters to be estimated. Thus, we suggest approximating the distribution of $K_n$ by a bootstrap procedure. Let $\{X_i^*\}_{i=1}^n$ denote the bootstrap draws of $\{X_i\}_{i=1}^n$ with replacement and $\hat{\nu}^*$ the bootstrap counterpart of $\hat{\nu}$.[1] Denote $\bar{s}(p) = n^{-1}\sum_{i=1}^n s(X_i, p)$ and $\hat{V}(p) = n^{-1}\sum_{i=1}^n \{s(X_i, p) - \bar{s}(p)\}^2$. For the bootstrap counterpart of the empirical likelihood function $\ell_n(p)$, we propose

$$\ell_n^*(p) = \max\left\{ \prod_{i=1}^n nw_i \middle| \sum_{i=1}^n w_i\{s(X_i^*, p) - s(\Theta_0(\hat{\nu}^*), p)\} = \{\bar{s}(p) - s(\Theta_0(\hat{\nu}), p)\},\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}. \tag{7}$$

Note that $\ell_n^*(p)$ does not directly mimic the original statistic but rather evaluates the likelihood after recentering by $\bar{s}(p) - s(\Theta_0(\hat{\nu}), p)$. Such a recentering is necessary to account for the effect of the estimated nuisance parameters.[2] Indeed, by Giné and Zinn (1990), after imposing bootstrap analogs of Assumption M (i)-(iii), a similar argument to the proof of Theorem 1 implies that $-2\log \ell_n^*(p)$ is approximated by
$\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \{s(X_i^*, p) - \bar{s}(p)\} - \{s(\Theta_0(\hat{\nu}^*), p) - s(\Theta_0(\hat{\nu}), p)\}\right]^2 / \hat{V}(p)$. However, in the absence of recentering, the additional term $\bar{s}(p) - s(\Theta_0(\hat{\nu}), p)$ appears in the numerator which makes the bootstrap invalid. This is reminiscent of Stute, Gonzalez-Manteiga and Quindimil (1998) who showed inconsistency of the classical bootstrap in the context of model checks for regression. Using the quadratic expansion above, standard arguments

[1] If $\nu$ is a smooth function of means, then $\hat{\nu}^*$ is given by replacing the moments with the bootstrap counterparts. If $\hat{\nu}$ is an M-estimator, we obtain $\hat{\nu}^*$ through properly recentered estimating equations as in Shorack (1982) and Lahiri (1992).

[2] The idea of recentering estimating equations is developed in Shorack (1982) and Lahiri (1992). It is interesting to see whether such recentering induces a desirable higher-order property in our setup as in Lahiri (1992).

based on Giné and Zinn (1990) enable us to prove the following consistency result for the proposed bootstrap statistic.

**Proposition 1.** *Under Assumptions M and M', the process* $\{\ell_n^*(p) : p \in \mathbb{S}^d\}$ *converges in distribution to the Gaussian process* $\{\{Z(p) - G(p;\nu)'Z_1\}^2/\mathrm{Var}(s(X,p)) : p \in \mathbb{S}^d\}$ *in* $P^*$*-probability, where* $P^*$ *denotes the probability computed under the bootstrap distribution conditional on the data.*

Therefore, the bootstrap critical values of $K_n$ are given by the quantiles of $K_n^* = \sup_{p\in\mathbb{S}^d}\{-2\log \ell_n^*(p)\}$.

2.1.2. *Case of no nuisance parameter.* If there is no nuisance parameter to be estimated (i.e., $\Theta_0(\nu) = \Theta_0$), Assumption M is implied by the sole requirement that $E[\|X\|_H^\xi] < \infty$ for some $\xi > 2$,[3] and the null distribution of $K_n$ becomes

$$K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{Z(p)^2}{E[Z(p)^2]},$$

where $Z$ is a Gaussian process with zero mean and covariance kernel $\mathrm{Cov}(s(X,p), s(X,q))$.

For comparison, let us consider the Wald type statistic of Beresteanu and Molinari (2008) adapted to the case of no nuisance parameters. In this case the statistic is simply $W_n = \sqrt{n} d_H\left(\frac{1}{n}\oplus_{i=1}^n X_i, \Theta_0\right)$, i.e., the contrast provided by the Hausdorff distance between the Minkowski average $\frac{1}{n}\oplus_{i=1}^n X_i$ and the null hypothetical set $\Theta_0$. For convex sets, the Wald type statistic $W_n$ may be alternatively characterized using the support functions as $W_n = \sqrt{n}\sup_{p\in\mathbb{S}^d}\left|\frac{1}{n}\sum_{i=1}^n s(X_i,p) - s(\Theta_0,p)\right|$ (Beresteanu and Molinari, 2008, equation

[3]This follows from the Lipschitz property of the support function, $|s(X,p) - s(X,q)| \le \|X\|_H \|p-q\|$ a.s. for any $p, q \in \mathbb{S}^d$, which ensures that $\{s(X,p) : p \in \mathbb{S}^d\}$ is $\mu$-Donsker by a standard empirical process argument (e.g., van der Vaart, 1998, Example 19.7).

(A.1)). Based on the proof of Theorem 1, we can then see that

$$K_n^{1/2} = \sqrt{n} \sup_{p \in \mathbb{S}^d} E[Z(p)^2]^{-1/2} \left| \frac{1}{n} \sum_{i=1}^{n} s(X_i, p) - s(\Theta_0, p) \right| + o_p(1),$$

under $\mathrm{H}_0$. Therefore, while the Wald type statistic $W_n$ of Beresteanu and Molinari (2008) evaluates the contrast $\frac{1}{n}\sum_{i=1}^{n} s(X_i, p) - s(\Theta_0, p)$ over $p \in \mathbb{S}^d$, the empirical likelihood statistic $K_n$ evaluates the same contrast but normalized by its standard deviation. This normalization ensures that our statistic $K_n$ is invariant to scale transformations (i.e., multiplication of both $\{X_i\}_{i=1}^n$ and $\Theta_0$ by some non-singular matrix independent of $i$), unlike the Wald type statistic $W_n$ which is sensitive to such transforms.[4] In Section 4.1, we illustrate that the lack of invariance of the Wald type statistic can yield different size properties depending on what scaling is used.

When there is no nuisance parameter, it is possible to invert $K_n$ to obtain an approximate confidence region within which the Aumann expectation $\mathbb{E}[X]$ lies with some desired probability. Indeed, using the quadratic approximation for the empirical likelihood process (cf. proof of Theorem 1), it follows that with probability $\alpha$, the support function for the set $\mathbb{E}[X]$ asymptotically satisfies $s(\mathbb{E}[X], p) \leq n^{-1}\sum_{i=1}^{n} s(X_i, p) + \sqrt{\frac{\hat{c}_\alpha}{n}}\hat{V}(p)^{1/2}$ for all $p \in \mathbb{S}^d$, where $\hat{c}_\alpha$ is the bootstrap estimate of the $\alpha$-th quantile of the limiting distribution of $K_n$. Based on the right hand side of this inequality, we can thus recover the confidence region that covers $\mathbb{E}[X]$ with the desired probability level $\alpha$.

2.2. **Sieve empirical likelihood.** Another way to construct an empirical likelihood for testing $\mathrm{H}_0$ in (2) is to incorporate the continuum of moment conditions $E[s(X, p)] =$

[4]For the identified set $\Theta_0 = \{\theta : E[m(\theta)] \leq 0\}$ defined by a finite number of moment inequalities, Chernozhukov, Kocatulum and Menzel (2015) proposed a confidence region that is invariant to arbitrary one-to-one mappings of the form $\tau : \Theta_0 \to \Psi$. However, their construction does not apply in general to our setup which is concerned with testing $\mathbb{E}[X_i] = \Theta_0$ implying the continuum of moment inequalities. In contrast, invariance of $K_n$ is restricted to particular transformations (i.e., multiplication of both $\{X_i\}_{i=1}^n$ and $\Theta_0$ by some non-singular matrix independent of $i$).

$s(\Theta_0(\nu), p)$ for all $p \in \mathbb{S}^d$ into a vector of moments with growing dimension. Let $k = k_n$ be a sequence of positive integers satisfying $k \to \infty$ as $n \to \infty$, and choose points (or sieve) $\{p_1, \ldots, p_k\}$ from $\mathbb{S}^d$ so that in the limit they form a dense subset of $\mathbb{S}^d$. By plugging in the nuisance parameter estimator $\hat{\nu}$, the sieve empirical likelihood function under the restrictions $E[s(X, p_j)] = s(\Theta_0(\nu), p_j)$ for $j = 1, \ldots, k$ is defined as

$$l_n = \max\left\{\prod_{i=1}^n nw_i \middle| \sum_{i=1}^n w_i s(X_i, p_j) = s(\Theta_0(\hat{\nu}), p_j) \text{ for } j = 1, \ldots, k,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1\right\}. \tag{8}$$

If there is no nuisance parameter (i.e., $\Theta_0(\nu) = \Theta_0$), we can simplify the proof of Theorem 2 below to show that $(-2\log l_n - k)/\sqrt{2k} \xrightarrow{d} N(0,1)$ under the null $\mathrm{H}_0 : \mathbb{E}[X] = \Theta_0$. When there are nuisance parameters, the statistic $l_n$ containing $\hat{\nu}$ is not internally studentized (i.e., $(-2\log l_n - k)/\sqrt{2k}$ does not converge to the standard normal) due to the variance of $\hat{\nu}$. To recover internal studentization, we penalize the dual form of $l_n$ as

$$L_n = \sup_{\lambda \in \Lambda_n} 2\sum_{i=1}^n \log(1 + \lambda' m_k(X_i)) - n\lambda'(\bar{V}_k - \hat{V}_k)\lambda, \tag{9}$$

where $m_k(X_i) = [s(X_i, p_1) - s(\Theta_0(\hat{\nu}), p_1), \ldots, s(X_i, p_k) - s(\Theta_0(\hat{\nu}), p_k)]'$ and $\Lambda_n$,$\bar{V}_k$, and $\hat{V}_k$ are defined in the Appendix. The limiting null distribution of the penalized statistic $L_n$ is obtained as follows.

**Theorem 2.** *Under Assumption S in the Appendix, it holds that* $(L_n - k)/\sqrt{2k} \xrightarrow{d} N(0,1)$ *under* $\mathrm{H}_0$. *In addition,* $(L_n - k)/\sqrt{2k}$ *diverges to infinity under* $\mathrm{H}_1$.

By adapting the proof of Theorem 2, we can show that under the local alternative

$$\mathrm{H}_{1n} : E[s(X, p)] = s(\Theta_0(\nu), p) + a_n \eta(p) \text{ over } p \in \mathbb{S}^d,$$

for some continuous function $\eta$, where $a_n = k^{1/4}/\sqrt{n\eta_k'\dot{V}_k\eta_k}$ and $\eta_k = (\eta(p_1), \ldots, \eta(p_k))'$, the sieve empirical likelihood statistic satisfies $(L_n - k)/\sqrt{2k} \xrightarrow{d} N(2^{-1/2}, 1)$. Therefore, the test statistic $(L_n - k)/\sqrt{2k}$ has non-trivial local power against a local alternative at the $a_n$-rate. Also, we note that similar to the marked empirical likelihood statistic $K_n$, both $l_n$ and $L_n$ are invariant to scale transformations (i.e., multiplication of both $\{X_i\}_{i=1}^n$ and $\Theta_0$ by some non-singular matrix independent of $i$).

Compared to the marked empirical likelihood statistic studied in Section 2.1, the sieve empirical likelihood statistic $L_n$ is asymptotically pivotal but requires choosing the sieve $\{p_1, \ldots, p_k\}$. A natural choice for locations of the sieve $\{p_1, \ldots, p_k\}$ is a grid of equidistant angle values in $\mathbb{S}^d$. The main remaining problem for practical implementation is choosing the tuning parameter $k$. In the literature on empirical likelihood, several statistics have been proposed possessing the same feature (i.e., asymptotically pivotal but depending on smoothing parameters), see for instance Fan, Zhang and Zhang (2001), Chen, Härdle and Li (2003), and Fan and Zhang (2004). Following the insight of Fan, Zhang and Zhang (2001) and Fan and Zhang (2004), one may choose $k$ to be the maximizer $\arg\max_{k\in[n^c, n^{c'}]}(L_n - k)/\sqrt{2k}$ for some constants $c' \geq c > 0$. This results in a multi-scale test whose critical value can be obtained by bootstrap. For goodness-of-fit testing of parametric regression models, Fan and Huang (2001) showed adaptive minimaxity of such a test. A thorough analysis of mutli-scale testing in our setup is beyond the scope of this paper.

## 3. Discussion and extensions

3.1. **Test for directed hypotheses.** It is possible to extend the methodology of marked empirical likelihood to test directed hypotheses of the form[5]

$$\mathrm{H}_0 : \Theta_0(\nu) \subseteq \mathbb{E}[X] \text{ vs. } \mathrm{H}_1 : \Theta_0(\nu) \nsubseteq \mathbb{E}[X]. \tag{10}$$

Beresteanu and Molinari (2008) were the first to develop a Wald type test for this problem. Here we propose empirical likelihood tests. By analogy with the testing problem in (1), the above is equivalent to testing the continuum of moment inequalities

$$\mathrm{H}_0 : s(\Theta_0(\nu), p) \leq E[s(X, p)] \text{ for all } p \in \mathbb{S}^d \text{ vs. } \mathrm{H}_1 : s(\Theta_0(\nu), p) > E[s(X, p)] \text{ for some } p \in \mathbb{S}^d.$$

For a given direction $p$ and preliminary estimator $\hat{\nu}$, the moment inequality restriction can be used to form the directed-marked empirical likelihood function

$$\vec{\ell}_n(p) = \max\left\{ \prod_{i=1}^n n w_i \,\middle|\, s(\Theta_0(\hat{\nu}), p) \leq \sum_{i=1}^n w_i s(X_i, p),\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\},$$

which can be equivalently written in the dual form as (see, Canay, 2010)

$$\vec{\ell}_n(p) = \min_{\lambda \leq 0} \prod_{i=1}^n \frac{1}{1 + \lambda\{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}}.$$

Therefore, the directed hypothesis in (10) can be tested by assessing the process $\{\vec{\ell}_n(p) : p \in \mathbb{S}^d\}$. In particular, we propose the directed Kolmogorov-Smirnov type statistic $\vec{K}_n = \sup_{p \in \mathbb{S}^d}\{-2 \log \vec{\ell}_n(p)\}$. By similar arguments as in the proof of Theorem 1 (in particular, by modifying the proof of Hjort, McKeague and van Keilegom 2009, Theorem 2.1), we can show that $\vec{K}_n \xrightarrow{d} \sup_{p \in \mathbb{S}^d} \frac{\min\{Z(p) - G(p;\nu)' Z_1, 0\}^2}{\mathrm{Var}(s(X,p))}$ under $\mathrm{H}_0$. The same also applies for

[5]The null for the opposite direction $\mathrm{H}_0 : \mathbb{E}[X] \subseteq \Theta_0(\nu)$ can be treated analogously.

testing the hypothesis $\mathrm{H}_0 : \theta_0 \in \mathbb{E}[X]$ for a singleton $\theta_0 \in \mathbb{R}^d$. In this case, we simply set $s(\Theta_0(\nu), p) = s(\Theta_0(\hat{\nu}), p) = p'\theta_0$.

It may be possible to extend the construction of the sieve empirical likelihood statistic to test the directed hypotheses in (10) by replacing the equality constraints $\sum_{i=1}^n w_i s(X_i, p_j) = s(\Theta_0(\hat{\nu}), p_j)$ in (8) with the inequalities $\sum_{i=1}^n w_i s(X_i, p_j) \geq s(\Theta_0(\hat{\nu}), p_j)$ for $j = 1, \ldots, k$. If $k$ is fixed, we can apply the results of Canay (2010) to investigate its asymptotic properties. However, for the case of $k \to \infty$, the asymptotic analysis of the statistic is very different and is beyond the scope of this paper.

3.2. **Linear transform and projection.** Our empirical likelihood approach can be easily modified to test hypotheses on a linear transform $R\mathbb{E}[X]$ of the Aumann mean, where $R$ is an $l \times d$ constant matrix with $l < d$ and full row rank. The first test for such hypotheses was proposed by Beresteanu and Molinari (2008) who employed a Wald type statistic based on the Hausdorff metric. Here we provide empirical likelihood based alternatives. Since the null hypothesis $\mathrm{H}_0^R : R\mathbb{E}[X] = R\Theta_0(\nu)$ is equivalent to $\mathrm{H}_0^R : E[s(X, R'q)] = s(\Theta_0(\nu), R'q)$ for all $q \in \mathbb{S}^l$, this motivates the use of the marked empirical likelihood function $\ell_n(R'q)$ for $q \in \mathbb{S}^l$, and the Kolmogorov-Smirnov type statistic $K_n^R = \sup_{q \in \mathbb{S}^l}\{-2\log \ell_n(R'q)\}$ for testing the null. By the invariance property, the latter is simply $K_n^R = \sup_{p \in \Delta}\{-2\log \ell_n(p)\}$, where $\Delta = \{R'q/\|R'q\| : q \in \mathbb{S}^l\}$ is a subset of $\mathbb{S}^d$. Thus, the test statistic $K_n^R$ for the linear transform is given by taking the supremum over a particular subset $\Delta \subset \mathbb{S}^d$ rather than the whole set $\mathbb{S}^d$ as is the case with $K_n$. A modification of Theorem 1 then implies $K_n^R \xrightarrow{d} \sup_{p \in \Delta} \frac{\{Z(p) - G(p;\nu)'Z_1\}^2}{\mathrm{Var}(s(X,p))}$ under $\mathrm{H}_0^R$. It is also possible to extend the sieve empirical likelihood approach to test $\mathrm{H}_0^R$ by choosing a sieve on $\Delta$.

Now let us discuss one of the most important examples: testing for the projection of $\mathbb{E}[X]$ to one of its components. We argue that in this case the sieve empirical likelihood (with profiling out for $\nu$) is particularly attractive. Suppose we are interested in the first component (i.e., $R = [1, 0, \dots, 0]$). In this case, the null hypothesis $\mathrm{H}_0^R : R\mathbb{E}[X] = R\Theta_0(\nu)$ reduces to the two moment constraints $\mathrm{H}_0^R : E[s(X, R'q)] = s(\Theta_0(\nu), R'q)$ for $q = \pm 1$. Let $\nu$ be defined through the estimating equations $E[m(z_i, \nu)] = 0$ for observables $z_i$.[6] Then the sieve empirical likelihood reduces to the conventional empirical likelihood:

$$l_n(\nu) = \max\left\{ \prod_{i=1}^n nw_i \left| \sum_{i=1}^n w_i \begin{pmatrix} s(X_i, R') - s(\Theta_0(\nu), R') \\ s(X_i, -R') - s(\Theta_0(\nu), -R') \\ m(z_i, \nu) \end{pmatrix} = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right. \right\}.$$

By Qin and Lawless (1994), mild regularity conditions guarantee Wilks' theorem, that is $-2\max_\nu\{\log l_n(\nu)\} \xrightarrow{d} \chi_2^2$ under $\mathrm{H}_0^R$. In this case, we recommend internalizing the nuisance parameters $\nu$ and profiling them out because the statistic $l_n(\hat{\nu})$ with a preliminary estimator $\hat{\nu}$ is not asymptotically pivotal in general. See Section 3.3 below for further discussion.

3.3. **Profile likelihood.** In Section 2, we considered empirical likelihood statistics where the nuisance parameters $\nu$ are replaced with a preliminary estimator $\hat{\nu}$. This approach is particularly practical when the dimension of $\nu$ is high. On the other hand, as explained in the last subsection, there are some situations where profiling out $\nu$ may be desirable to achieve asymptotic pivotalness. Here we discuss some such extensions for profiling out $\nu$. Again, suppose throughout that $\nu$ is defined by some estimating equations $E[m(z_i, \nu)] = 0$ for observables $z_i$.

[6]When $\nu$ is defined by a smooth function of means, it can be treated as in Owen (2001, Section 3.4).

The marked profile empirical likelihood can be defined as $\ell_n^P(p) = \max_\nu \ell_n(p,\nu)$, where

$$\ell_n(p,\nu) = \max\left\{\prod_{i=1}^n nw_i \,\middle|\, \sum_{i=1}^n w_i \begin{pmatrix} s(X_i,p) - s(\Theta_0(\nu),p) \\ m(z_i,\nu) \end{pmatrix} = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1\right\}.$$

There is a computational drawback of this approach: it requires optimization with respect to $\nu$ for each $p$. Although the technical arguments would be more involved than the plug-in case, by extending the argument in Qin and Lawless (1994, Corollary 5) we can obtain the limiting distribution of the process $\ell_n^P(p)$. In particular, defining $g_i(p,\nu) = [s(X_i,p) - s(\Theta_0(\nu),p), m(z_i,\nu)]'$, we can show that $\sup_{p\in\mathbb{S}^d}\{-2\log \ell_n^P(p)\}$ will converge to $\sup_{p\in\mathbb{S}^d}\{\tilde{Z}(p)'\tilde{Z}(p)\}$, where $\tilde{Z}(p)' = [Z(p), Z_1']\left(I - S(p)\left(S(p)'\Omega(p)^{-1}S(p)\right)^{-1} S(p)'\right)\Omega(p)^{-1/2}$, with $Z_1$ denoting the limiting distribution of $n^{-1/2}\sum_{i=1}^n m(z_i,\nu_0)$, $S(p) = \begin{bmatrix} G(p;\nu_0)' \\ E[\partial m(z_i,\nu_0)/\partial\nu'] \end{bmatrix}$ (here $G(p;\nu_0)'$ is as defined in (5) and the existence of $E[\partial m(z_i,\nu_0)/\partial\nu']$ is assumed), and $\Omega(p) = \mathrm{Var}(g_i(p,\nu_0))$. We note the limiting distribution is still not pivotal, and the critical value needs to be approximated by bootstrap.

Similarly, the sieve profile empirical likelihood can be defined as $l_n^P = \max_\nu l_n(\nu)$, where

$$l_n(\nu) = \max\left\{\prod_{i=1}^n nw_i \,\middle|\, \sum_{i=1}^n w_i \begin{pmatrix} s(X_i,p_1) - s(\Theta_0(\nu),p_1) \\ \vdots \\ s(X_i,p_k) - s(\Theta_0(\nu),p_k) \\ m(z_i,\nu) \end{pmatrix} = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1\right\}.$$

Compared to the marked profile empirical likelihood $\ell_n^P(p)$, the sieve statistic $l_n^P$ is more tractable because it requires optimization with respect to $\nu$ only once. Additionally, by arguing as in Donald, Imbens and Newey (2003, Theorems 6.3-6.4), it can be shown that the null distribution is standard normal, i.e. $(l_n^P - k)/\sqrt{2k} \xrightarrow{d} N(0,1)$ under certain conditions. Thus, the profile statistic $l_n^P$ is asymptotically pivotal without the need for penalization as in (9).

3.4. **Inference based on estimated random sets.** In some applications, the random set of interest $X$ is not directly observable because it contains some parameters to be estimated. For example, in the context of treatment effect analysis in experimental studies, Balke and Pearl (1997) proposed nonparametric bounds on the average treatment effect when the treatment assignment is random but subject compliance is imperfect. In a general form, Balke and Pearl's (1997) bound on the average treatment (ATE) can essentially be written as

$$\max_{1\le j\le J_L} \frac{E[g^j_{Li}]}{E[h^j_{Li}]} \le ATE \le \max_{1\le j\le J_U} \frac{E[g^j_{Ui}]}{E[h^j_{Ui}]}, \tag{11}$$

where $g^j_{Li}$ $(j = 1, \ldots, J_L)$ and $g^j_{Ui}$ $(j = 1, \ldots, J_U)$ are observable scalar random variables. By applying the "smooth-max" approximation (Chernozhukov, Kocatulum and Menzel, 2015), these bounds can be approximated by $\sum_{j=1}^{J_L} w^j_A E[g^j_{Ai}]/E[h^j_{Ai}]$ with $w^j_A = e^{\varrho E[g^j_{Ai}]/E[h^j_{Ai}]} / \left(\sum_{j=1}^{J_A} e^{\varrho E[g^j_{Ai}]/E[h^j_{Ai}]}\right)$ for $A = L$ and $U$. Indeed, the approximation error satisfies $\left|\sum_{j=1}^{J_A} w^j_A E[g^j_{Ai}]/E[h^j_{Ai}] - \max_{1\le j\le J_A} E[g^j_{Ai}]/E[h^j_{Ai}]\right| = O(\varrho^{-1})$ for $A = L$ and $U$. Thus by choosing $\varrho$ large enough, the bounds on the ATE given above are well approximated by the Aumann expectation $\mathbb{E}[X_i(\gamma)]$ of the SVRV

$$X_i(\gamma) = \left[\sum_{j=1}^{J_L} w^j_L g^j_{Li}/E[h^j_{Li}], \sum_{j=1}^{J_U} w^j_U g^j_{Ui}/E[h^j_{Ui}]\right],$$

where $\gamma = (E[g^1_{Li}], \ldots, E[g^{J_L}_{Li}], E[h^1_{Li}], \ldots, E[h^{J_L}_{Li}], E[g^1_{Ui}], \ldots, E[g^{J_U}_{Ui}], E[h^1_{Ui}], \ldots, E[h^{J_U}_{Ui}])'$. In this case, the SVRV of interest $X_i(\gamma)$ is not observable because it contains unknown parameters $\gamma$.

In order to test null hypotheses of the form $\mathrm{H}_0 : \mathbb{E}[X(\gamma)] = \Theta_0(\nu)$, the marked empirical likelihood function $\ell_n(p)$ in (3) can be modified by replacing $X_i$ with the estimated counterpart $X_i(\hat{\gamma})$, where $\hat{\gamma}$ is an estimator of $\gamma$. By imposing assumptions analogous to Assumption M (i)-(iii) to deal with the estimation error of $X_i(\hat{\gamma}) - X_i(\gamma)$ along with the assumption $\sup_{p\in\mathbb{S}^d} E[|s(X_i(\gamma_m), p) - s(X_i(\gamma), p)|^2] \to 0$ for all $\gamma_m \to \gamma$, we can show that

$$K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\{Z(p) - G(p;\nu)'Z_1 + \Gamma(p;\gamma)'Z_2\}^2}{\mathrm{Var}(s(X(\gamma), p))},$$

where $(Z(p), Z_1', Z_2')' \sim N(0, \tilde{V}(p))$, $\tilde{V}(p)$ is the limiting covariance matrix of $(n^{-1/2}\sum_{i=1}^n \{s(X_i, p) - E[s(X, p)]\}, \sqrt{n}(\hat{\nu} - \nu)', \sqrt{n}(\hat{\gamma} - \gamma))'$, and $\Gamma(p;\gamma)$ is a function such that

$$|E[s(X(\hat{\gamma}), p)] - E[s(X(\gamma), p)] - \Gamma(p;\gamma)'(\hat{\gamma} - \gamma)| = o_p(n^{-1/2}).$$

To obtain a critical value for testing, we can adapt the bootstrap procedure presented in Proposition 1 (by replacing $X_i^*$ and $\bar{s}(p)$ in (7) with $X_i^*(\hat{\gamma}^*)$ and $n^{-1}\sum_{i=1}^n s(X_i(\hat{\gamma}), p)$, respectively). The asymptotic validity of this bootstrap procedure can be shown under the additional condition: $\sup_{p\in\mathbb{S}^d} |\bar{s}(X_i(\hat{\gamma}^*), p) - \bar{s}(X_i(\hat{\gamma}), p) - \Gamma(p;\gamma)'(\hat{\gamma}^* - \hat{\gamma})| = o_{p^*}(n^{-1/2})$ with probability approaching 1.

It is also possible to employ the sieve empirical likelihood statistic by replacing $X_i$ in (8) with the estimated set $X_i(\hat{\gamma})$. Recall, in Section 2.2 we were able to incorporate nuisance parameters into the sieve statistic by linearizing the term $s(\Theta_0(\hat{\nu}), p) - s(\Theta_0(\nu), p)$ and incorporating the effect of the resulting additional terms via penalization (see the Appendix for more details). We can proceed similarly for the case of estimated sets if we impose the following assumption enabling linearization of $\bar{s}(X_i(\hat{\gamma}), p) - \bar{s}(X_i(\gamma), p)$ as

$$\sup_{p\in\mathbb{S}^d} |\bar{s}(X_i(\hat{\gamma}), p) - \bar{s}(X_i(\gamma), p) - \bar{\Gamma}(p;\gamma)'(\hat{\gamma} - \gamma)| = o_p(n^{-1/2}),$$

where $\bar{\Gamma}(.;.)$ is the derivative of $\bar{s}(X_i(\gamma), p)$ with respect to $\gamma$ satisfying some regularity properties akin to Assumption S (iii) (i.e., (i) $\bar{\Gamma}(p;\gamma)$ converges uniformly in both $p$ and

$\nu$ to a non-stochastic $\Gamma(p;\gamma)$ satisfying $\sup_{p\in\mathbb{S}^d}\|\Gamma(.,\gamma)\|<\infty$ and (ii) for all $\tilde{\gamma}$ in some neighborhood of $\gamma$, $\sup_{p\in\mathbb{S}^d}\left\|\bar{\Gamma}(p;\tilde{\gamma})-\bar{\Gamma}(p;\gamma)\right\| \leq M\|\tilde{\gamma}-\gamma\|^{\alpha}$ for some $\alpha\geq 2/3$ and $M<\infty$ independent of $\tilde{\gamma}$). By a straightforward modification of the penalty term in (9), we can obtain a corresponding result to Theorem 2 for the case of estimated random sets.

Alternatively, it is possible to employ a profile likelihood approach as in section (3.3); this is particularly attractive for tests on low dimensional projections of the set $\Theta_0(\nu)$.

3.5. **Measurements on support function.** In medical imaging and robotic vision, researchers sometimes directly observe measurements of the support function of a convex set of interest (see, Fisher *et al.*, 1997). When noiseless measurements of $\{s(X_i,\cdot)\}_{i=1}^n$ are available, the marked empirical likelihood method can be applied immediately to hypothesis testing. Another common statistical question in image analysis of convex shaped data is to recover a set of interest from noisy measurements of its support function. In this problem, we observe the pairs $\{s_i,p_i\}_{i=1}^n$, where $s_i=s(\Theta,p_i)+\epsilon_i$ with error $\epsilon_i$ and $p_i\in\mathbb{S}^d$. Fisher *et al.* (1997) developed an estimation method for $\Theta$ by estimating the support function $s(\Theta,\cdot)$ nonparametrically. Our empirical likelihood approach can be adapted to test the hypothesis that $\Theta$ takes a particular shape $\Theta_0$, such as a circle or ellipse. The marked empirical likelihood function under the restriction $E[s_i|p_i=p]=s(\Theta_0,p)$ may be constructed as

$$\tilde{\ell}_n(p)=\max\left\{\prod_{i=1}^n nw_i\left|\sum_{i=1}^n w_iK_b(p_i-p)\{s_i-s(\Theta_0,p)\}=0,\ w_i\geq 0,\ \sum_{i=1}^n w_i=1\right.\right\}, \quad (12)$$

where $K_b(\cdot)$ is a kernel function depending on the smoothing parameter $b$. For example, the Cramér-von Mises type statistic, given by $T_n=\int_{p\in\mathbb{S}^d}-2\log\tilde{\ell}_n(p)dp$, can be shown to be asymptotically normal under the null after certain normalizations as in Chen, Härdle and Li (2003). Alternatively, following Härdle and Mammen (1993), a wild bootstrap method (i.e., resampling $s_i^* = s(\Theta_0, p_i) + v_i^* \hat{\epsilon}_i$ with $\hat{\epsilon}_i = s_i - s(\Theta_0, p)$ and $v_i^*$ ~two-point distribution) can be applied to obtain the critical value.

Simulation results, presented in the web appendix, demonstrate reasonable size and power properties for our empirical likelihood test.

## 4. Examples

4.1. **Best linear prediction with interval valued dependent variable.** We first consider the issue of best linear prediction with interval valued dependent variables. In particular, we employ the setup of Beresteanu and Molinari (2008), follow their argument, and use the characterization they provide. See also Bontemps, Magnac and Maurin (2012) for an extension to instrumental variable regression.

In usual regression models, we are mostly interested in the best linear relationship between a dependent variable $y$ and independent variables $x$, which can be estimated by the least squares method. On the other hand, if $y$ is unobservable but we observe the interval $[y_L, y_U]$ to which $y$ belongs almost surely, it would be of interest to conduct inference on the set of the least squares coefficients $\Upsilon = \{\arg\min_\theta \int \{y-(1, x')\theta\}^2 d\mu \text{ for some } \mu \in \mathcal{M}\}$, where $\mathcal{M}$ is the set of distributions of $(y, x)$ compatible with $y \in [y_L, y_U]$ almost surely. There are numerous examples of interval data, including data on wealth (e.g., the Health and Retirement Study) and income (e.g., the Current Population Survey), top coding in surveys, and ordered categorical measurements (e.g., age, expenditure, GPA, and so on). By using the Aumann expectation for the random set $W = \begin{pmatrix} [y_L, y_U] \\ [xy_L, xy_U] \end{pmatrix} \subset$

$\mathbb{R}^{\dim(x)+1}$, the set of least square coefficients may be written as $\Upsilon = \Sigma^{-1}\mathbb{E}[W]$, where $\Sigma = E\begin{pmatrix} 1 & x' \\ x & xx' \end{pmatrix}$ (see, Beresteanu and Molinari, 2008, Proposition 4.1).[7]

We note that if there is no intercept in the regression and $x$ is scalar (or there is only an intercept), then the set of best linear predictors is the interval $\Upsilon = [E[xy_L]/E[x^2], E[xy_U]/E[x^2]]$. Thus, inference on $\Upsilon$ may be conducted by the conventional empirical likelihood for the vector of parameters $(E[xy_L], E[xy_U], E[x^2])$ or via regressions of $y_L$ and $y_U$ on the scalar $x$. However, if the regression model contains an intercept or $x$ is a vector, then the set $\Upsilon$ is multi-dimensional and neither the conventional empirical likelihood for $(E[(1,x')y_L], E[(1,x')y_U], \Sigma)$ nor regressions of $y_L$ and $y_U$ on $(1,x')$ are sufficient for characterizing it completely. Intuitively this is because, as can be seen from the characterization of the support function of $\Upsilon$ given below, we also need to consider situations where some observations of $y$ take the value $y_L$ while the others take $y_U$.

For the following theoretical results we shall suppose that $x$ is a continuous random variable which ensures $\Upsilon$ is strictly convex. Regarding the support function, the null hypothesis $\mathrm{H}_0 : \Upsilon = \Upsilon_0$ for a strictly convex $\Upsilon_0$ can be written as $\mathrm{H}_0 : E[s(W,p)] = s(\Sigma\Upsilon_0, p)$ for all $p \in \mathbb{S}^d$, where $s(W,p) = [y_L + (y_U - y_L)\mathbb{I}\{(1,x')p \geq 0\}](1,x')p$ and $d = \dim(x)+1$. This is equivalent to the general setup of Section 2 if one defines $\Theta_0(\nu) = \Sigma\Upsilon_0$, where the nuisance parameter $\nu = \mathrm{vec}(\Sigma)$ is estimated by its sample counterpart $\mathrm{vec}(\hat{\Sigma})$. Furthermore, since $s(\Sigma\Upsilon_0, p) = s(\Upsilon_0, \Sigma p)$, the support function of the set $\Sigma\Upsilon_0$ can be computed from that of $\Upsilon_0$. Let $\nabla s(\Upsilon_0, p)' = [y_L + (y_U - y_L)\mathbb{I}\{(1,x')p \geq 0\}](1,x')$ be the Fréchet derivative of $s(\Upsilon_0, p)$ with respect to $p$, and define $G(p;\nu) = p \otimes \nabla s(\Upsilon_0, \Sigma p)$, where $\otimes$ represents the Kronecker product. Note that $G(p;\nu)'$ is the pointwise derivative of $s(\Sigma\Upsilon_0, p)$ ($s(\Theta_0(\nu_0))$ in the terminology of Section 2) with respect to $\nu = \mathrm{vec}(\Sigma)$.

[7]Chandrasekhar *et al.* (2012) extended this model further to allow for $y_L$ and $y_U$ to be nonparametrically estimable functions. Although it is beyond the scope of this paper, it would be interesting to extend our empirical likelihood approach to such situations.

In this setup, the null distributions of the empirical likelihood statistics are obtained as follows.

**Proposition 2.** *Consider the setup of this subsection. Assume that* $\{y_{Li}, y_{Ui}, x_i\}_{i=1}^n$ *is i.i.d., where the distribution of* $x_i$ *is absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}^{d-1}$*, and* $\Sigma$ *is full rank.*

***(i)*** *Suppose* $E[\|(y_{Li}, y_{Ui}, x_i' y_{Li}, x_i' y_{Ui})\|^\xi] < \infty$ *for some* $\xi > 2$*,* $E[\|x_i\|^4] < \infty$*, and* $\mathrm{Var}(y_{Li}|x_i), \mathrm{Var}(y_{Ui}|x_i) \geq \underline{\sigma}^2$ *a.s. for some* $\underline{\sigma}^2 > 0$*. Then* $K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\tilde{Z}(p)^2}{\mathrm{Var}(s(W_i,p))}$ *under* $\mathrm{H}_0$*, where* $\tilde{Z}(\cdot) = Z(\cdot) - G(\cdot;\nu)'\Gamma$ *is the Gaussian process implied from* $(Z(p), \Gamma)' \sim N(0, \tilde{V}(p))$ *and* $\tilde{V}(p)$ *is the covariance matrix of the vector* $(s(W_i, p), \{z_i - vec(\Sigma)\}')$*.*

***(ii)*** *Suppose* $E[\|(y_{Li}, y_{Ui}, x_i' y_{Li}, x_i' y_{Ui})\|^\xi] < \infty$ *for some* $\xi \geq 4$*,* $E[\|x_i\|^4] < \infty$*, and* $\nabla s(\Upsilon_0, p)$ *is locally Hölder continuous of order* $\alpha \geq 2/3$ *over the domain* $\mathbb{S}^d$*. Also assume* $k \to \infty$ *and* $(k^5 \dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$*, where* $\dot{\phi}_k$ *is defined in Appendix. Then* $\frac{L_n - k}{\sqrt{2k}} \xrightarrow{d} N(0,1)$ *under* $\mathrm{H}_0$*.*

The assumptions are similar to those of Beresteanu and Molinari (2008, Theorem 4.3). These results are obtained by verifying the conditions in Theorems 1 and 2. The critical values for the marked empirical likelihood test may be obtained by the bootstrap procedure presented in Proposition 1.

We now evaluate the finite sample performance of our test statistic by conducting inference on the returns to education on (log) wages using data from the Current Population Survey (CPS). We use data from the March 2009 wave of the CPS on white males aged between 20 and 50 who earn at least \$1000/year. This gives 18017 observations on wages and education. Analogous to the construction in Beresteanu and Molinari (2008), the wage data (in thousands of dollars) is artificially bracketed and top-coded in terms of the following brackets (the top coding value is \$100 million):

[1, 5], [5, 7.5], [7.5, 10], [10, 12.5], [12.5, 15], [15, 20], [20, 25], [25, 30], [30, 35], [35, 40], [40,50], [50, 60], [60, 75], [75, 100], [100, 150], [150, 100000]

Thus, the variables $(y_{Li}, y_{Ui,} x_i)$ correspond to lower and upper bounds of log wages and education, respectively. We draw 5000 samples of size $n$ =100, 200, 500, 1000, and 2000 from the 'true' population (consisting of 18017 observations from the CPS) and conduct inference for $\Upsilon$, the set of intercept and slope coefficients consistent with the population data. Table 1 reports the rejection frequencies of the marked empirical likelihood test under the nominal 5% rejection level. This is compared with Wald-type test statistics based on the Hausdorff distances $nd_H\left(\hat{\Sigma}^{-1}\frac{1}{n}\oplus_{i=1}^{n}W_i, \Upsilon_0\right)^2$ and $nd_H\left(\frac{1}{n}\oplus_{i=1}^{n}W_i, \hat{\Sigma}\Upsilon_0\right)^2$ (called Wald 1 and 2, respectively). The first Wald-type test was proposed by Beresteanu and Molinari (2008). For both the marked empirical likelihood and Wald tests, the critical values are obtained by the bootstrap calibrations outlined in Section 2 with 399 repetitions. In Table 1 it is seen that the marked empirical likelihood test has good size control and performs better than both Wald tests for smaller sample sizes. As explained previously, the Wald statistic is not invariant to multiplication of the sets by a constant matrix unlike the empirical likelihood tests; this drawback is evident in the different sizes for the two Wald tests.[8] The statistics vary considerably along $p$; for some directions $p = (\cos\vartheta, \sin\vartheta)'$ with $\vartheta = \left(0, \frac{\pi}{3}, \frac{\pi}{4}, \frac{2\pi}{3}, \frac{\pi}{2}\right)$, the critical values of Wald 1, marked EL, and $\hat{V}(\hat{\Sigma}p)$ are $(5.3\times10^{-2}, 2.0\times10^{-5}, 1.4\times10^{-4}, 2.5\times10^{-4})$, $(10, 6.8, 4.3, 2.0, 0.14)$, and $(7.5, 337.4, 610.0, 870.8, 1.1\times10^{3})$, respectively.

We can also adapt the construction of the confidence set based on $K_n$, described in Section 2, to the present context. We exploit the invariance property of $K_n$ which ensures

[8]As expected, however, the marked empirical likelihood test is computationally more expensive than the Wald test. In particular, for sample size $n = 1000$, the marked empirical likelihood test with 399 bootstrap repetitions has an average run time of 5.7 seconds as compared to 0.6 seconds for the Wald test.

that with probability $\alpha$ the inequalities $s(\Upsilon, p) \leq n^{-1}\sum_{i=1}^{n} s(\hat{\Sigma}^{-1}X_i, p) + \sqrt{\frac{\hat{c}_\alpha}{n}}\hat{V}(\hat{\Sigma}^{-1}p)^{1/2}$ hold asymptotically for all $p \in \mathbb{S}^d$, where $\hat{c}_\alpha$ estimates the $\alpha$-th quantile of the limiting distribution of $K_n$. In particular, we can obtain $\hat{c}_\alpha$ by the bootstrap procedure presented in Section 3.4. Figure 1 displays the 95% confidence region thus obtained for a sample size of $n = 1000$, along with the 'true' population region and also the confidence region from the Wald-type test proposed in Beresteanu and Molinari (2008). It can be seen that the confidence region based on $K_n$ covers an area that is much less ($< 5\%$) than the one based on the Wald test.

We can also employ our inferential procedures to obtain confidence intervals for the best linear predictor of the (log) wage $y$ given some education $x$. This is equivalent to providing a confidence region for the projection $R\Upsilon_0$ where $R = (1, x)$. To this end, we can use the results from Section 3.4 on estimated random sets by exploiting the fact $E[s(\Sigma^{-1}W, R'q)] = s(\Upsilon, R'q)$, where setting $q = 1$ and $-1$ gives the upper and lower bounds for the confidence interval. Table 2 reports the estimated prediction intervals for the cases when $x = 12$ (corresponding to high school education) and $x = 16$ (corresponding to undergraduate degree). For computational reasons we report the results for profile likelihood using the Euclidean likelihood function (c.f. Section (2.1)). The profile likelihood is used to obtain a joint confidence set for the upper and lower bounds of the interval, from which we obtain a necessarily conservative confidence interval by taking the worst possible value for each of the components. Nevertheless, the length of the confidence interval is comparable to, or smaller, than those based on the Marked EL and Wald statistics.

In the web appendix, we report additional numerical results to compare the marked empirical likelihood confidence region displayed in Figure 1 with the one based on the method by Chernozhukov, Kocatulum and Menzel (2015).

TABLE 1. Rejection frequencies of the marked empirical likelihood and Wald tests at the nominal 5% level

| $n$ | Size (Marked EL) | Size (Wald 1) | Size (Wald 2) |
|---|---|---|---|
| 100 | 0.038 | 0.098 | 0.107 |
| 200 | 0.049 | 0.073 | 0.081 |
| 500 | 0.057 | 0.069 | 0.059 |
| 1000 | 0.053 | 0.057 | 0.059 |
| 2000 | 0.050 | 0.056 | 0.058 |

TABLE 2. 95% confidence intervals for the best linear predictor of (log) wage $y$ given education $x$ using profile likelihood, marked Empirical Likelihood and Wald statistics

| Education | True Region | Profile Lik. | Marked EL | Wald |
|---|---|---|---|---|
| High school degree | [3.549, 3.931] | [3.454, 3.999] | [3.456, 3.995] | [3.465, 3.983] |
| Undergraduate degree | [4.020, 4.915] | [3.967, 5.051] | [3.906, 5.003] | [3.873, 4.976] |

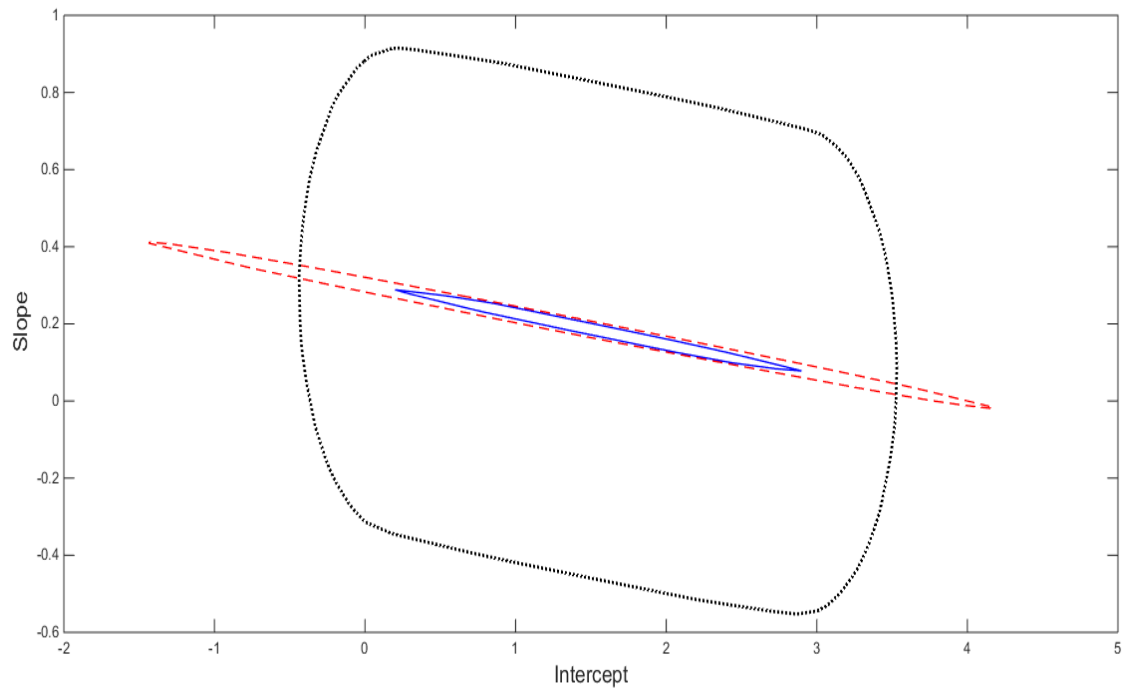


FIGURE 1. The population identification region (solid line) and the corresponding 95% confidence regions using the marked empirical likelihood statistic (dashed line) and the Wald statistic (dotted line) for sample size $n = 1000$.

4.2. **Boolean model.** In the context of mathematical morphology, geostatistics, and particle statistics, researchers often observe a series of two or three dimensional random sets, such as tumors and sand or rock grains (see, Stoyan, 1998, for a review). One of the most popular models to explain the growth pattern of these shapes is the Boolean model, where the random set is generated as $X = \cup_j \{W_j \oplus \{g_j\} : g_j \in G\}$ based on i.i.d. copies of random sets $W_j \subset \mathbb{R}^d$ ($j = 1, 2, \ldots$), and a point process $G$ in $\mathbb{R}^d$ for the foci $\{g_j : j \in \mathbb{N}\}$. For example, Cressie and Hulting (1992) developed a Boolean model to describe the growth of tumor shapes by specifying $G$ to be a Poisson process with constant intensity function $\lambda$ over a unit circle support. For simplicity, we shall assume that $W_j = W$ is a non-random ball with unknown radius $R$. We note that taking $R$ to be non-stochastic is not too strong a requirement in this instance. Indeed, as seen in Cressie and Hulting (1992), the variance of $R$ is an order of magnitude smaller than its mean. We thus consider $\gamma = (R, \lambda)$ as parameters of the tumor growth process which differ for normal and malignant tissues; consequently, we wish to conduct inference on these joint parameters.

To estimate $\gamma = (R, \lambda)$, Cressie and Hulting (1992) focused on the hitting probability (or capacity functionals). Alternatively, we can conduct inference using the Aumann expectation. More precisely, given the hypothesized parameter value $\gamma_0$, we can numerically evaluate the Aumann expectation $\Theta(\gamma_0) = \mathbb{E}[X(\gamma_0)]$. Then based on the sample $\{X_1, \ldots, X_n\}$ of tumor shapes of patients, the hypothesis $\mathrm{H}_0 : \gamma = \gamma_0$ can be tested via our methods for $\mathbb{E}[X] = \Theta(\gamma_0)$, specifically the marked (Section 2.1) and sieve empirical likelihood (Section 2.2) statistics.

We note that $X$ may not be convex in this example. However, as long as $X$ is compact valued, the Aumann expectation $\mathbb{E}[X]$ emerges as the almost sure limit of the Minkowski average of the sample $\{X_1, \ldots, X_n\}$. Therefore, the Aumann expectation can be intuitively interpreted as the 'average' shape of the observed sets. Furthermore, since the underlying probability measure is non-atomic in this example, it holds that $\mathbb{E}[X] = \mathbb{E}[\mathrm{co}(X)]$

(see, the discussion in Section 2). So, even though $X$ is non-convex, our inferential procedures continue to hold after applying the convex hull operation (note: the support function remains unchanged since $s(X,p) = s(\mathrm{co}(X),p)$ for any compact $X$).

We present some Monte Carlo simulation based on Cressie and Hulting (1992) to evaluate the finite sample performance of our test statistics. In particular, we simulate the data from the estimated parameter values for $\gamma$ obtained in Cressie and Hulting (1992, Table 3) with 5000 Monte Carlo replications for the sample sizes $n$ =100, 200, and 500. Numerical evaluation of $\Theta(\gamma_0)$ is achieved by averaging over 5000 draws of the process generated using the parameter value $\gamma_0$.

Table 3 reports the rejection frequencies of the marked empirical likelihood test under the nominal 5% rejection level. The null hypothesis is $\mathrm{H}_0 : \gamma_0 = (1.342, 4.046)$. We consider three types of alternatives $\mathrm{H}_1^a : (1.342, 4.5)$, $\mathrm{H}_1^b : (1.320, 4.046)$, and $\mathrm{H}_1^c : (1.320, 4.5)$. The critical values for this test are obtained by implementing the bootstrap procedure outlined in Section 3.4 with 99 repetitions. With respect to CPU seconds, the average computing time to obtain the bootstrap critical values is 4.85 for 399 repetitions and 1.84 for 99 repetitions. With additional parallel processing, we expect that these times may be further reduced. The first column indicates the test statistic has good size control over the sample sizes. The second and third columns show that the statistic is sensitive to slight changes in $R$ and, to a lesser extent, changes in $\lambda$. This is consistent with the standard deviations of the estimates in Cressie and Hulting (1992, Table 3) which are large for $\lambda$ compared to $R$. The fourth column reports the power properties of the test when changing both $R$ and $\lambda$. In this case, these changes somewhat cancel each other out in the net effect (lower radius vs. higher number of foci), which explains why the alternative $\mathrm{H}_1^c$ is harder to reject.

TABLE 3. Rejection frequencies of the marked empirical likelihood test at the nominal 5% level

| $n$ | $H_0$ | $H_1^a$ | $H_1^b$ | $H_1^c$ |
|---|---|---|---|---|
| 100 | 0.059 | 0.259 | 0.389 | 0.074 |
| 200 | 0.063 | 0.461 | 0.679 | 0.099 |
| 500 | 0.071 | 0.856 | 0.973 | 0.173 |

TABLE 4. Rejection frequencies of the sieve empirical likelihood test at the nominal 5% level

| $n, k$ | $H_0$ | $H_1^a$ | $H_1^b$ | $H_1^c$ |
|---|---|---|---|---|
| 100, 3 | 0.058 | 0.328 | 0.475 | 0.078 |
| 100, 5 | 0.074 | 0.383 | 0.584 | 0.088 |
| 100, 10 | 0.102 | 0.393 | 0.521 | 0.121 |
| 200, 3 | 0.059 | 0.569 | 0.789 | 0.095 |
| 200, 5 | 0.066 | 0.648 | 0.890 | 0.101 |
| 200,10 | 0.085 | 0.581 | 0.847 | 0.110 |
| 500, 3 | 0.070 | 0.944 | 0.993 | 0.176 |
| 500, 5 | 0.082 | 0.974 | 0.999 | 0.202 |
| 500, 10 | 0.090 | 0.940 | 0.999 | 0.174 |

Table 4 reports analogous results for the sieve empirical likelihood test. We construct the sieve from a grid of equidistant angle values corresponding to directions of the support function. We report outcomes for different values of sieve size $k$ =3, 5, and 10. The critical values for the test are based on a $\chi^2_k$ calibration since, for the sample sizes and values of $k$ considered, the theoretical normal approximation is found to be too rough. We see that the sieve empirical likelihood dominates the marked empirical likelihood in terms of power for all values of $k$ while having comparable size control for smaller values of $k$.

So far we have considered inference for the joint hypothesis involving both parameters $R$ and $\lambda$. By using our empirical likelihood tests with nuisance parameters, it is also possible to test the single parameter hypothesis $H_0 : \lambda = \lambda_0$ by plugging-in an estimated value for $R$ (e.g. the one in Cressie and Hulting, 1992).

4.3. **Treatment effect.** We consider the problem of inference for nonparametric bounds on average treatment effects in the presence of imperfect compliance. In particular, we conduct a simulation study based on the Vitamin A Supplementation example in Balke and Pearl (1997, Section 4.1). Briefly, the study consisted of administering doses of Vitamin A in a randomized trial to check for the effect on mortality. While the assignment to control and treatment groups was random, there were a substantial number of subjects who did not consume the treatment even when assigned to the treatment group. In the absence of any further assumptions on the relationship between compliance and response, Balke and Pearl (1997) obtained the sharpest possible bounds on the average treatment effect, which are of the form described in (11). Using the marked empirical likelihood statistic with estimated random sets proposed in Section 3.4, we can provide ways to conduct inference and construct confidence intervals for such bounds.

We use data simulated from the estimated joint probability distributions obtained in Balke and Pearl (1997, Tables 1 and 2) with 5000 Monte Carlo replications for each of the sample sizes $n$ =500, 1000, 2500, and 5000. Note that the numerical example in Balke and Pearl (1997) is based on over 20000 observations. We look at the size and power properties of the marked empirical likelihood test statistic under the null of the identified set $\mathrm{H}_0 : \Theta_0 = [-0.1946, 0.0054]$ and the alternative hypotheses obtained by expanding, contracting, and shifting $\Theta_0$ to the left by a value of 0.025 (i.e., $\mathrm{H}_1^a : [-0.2196, 0.0304]$, $\mathrm{H}_1^b : [-0.1696, -0.0196]$, and $\mathrm{H}_1^c : [-0.2196, -0.0196]$, respectively). The critical value for the test is obtained by implementing the bootstrap procedure outlined in Section 3.4 with 399 repetitions. The tuning parameter $\varrho$ for the 'smooth-max' approximation (cf. Section 3.4) employed in this test is chosen to be $\varrho = 1000$.

Table 5 reports the rejection frequencies of the marked empirical likelihood test under the nominal 5% rejection level. We can see that the our testing procedure has excellent size properties across all sample sizes (which are much smaller than the numerical example in Balke and Pearl, 1997). Also, our test has reasonable power properties against the three types of alternatives when the sample size is large enough.

Table 5. Rejection frequencies of the marked empirical likelihood test at the nominal 5% level

| $n$ | $\mathrm{H}_0$ | $\mathrm{H}_1^a$ | $\mathrm{H}_1^b$ | $\mathrm{H}_1^c$ |
|---|---|---|---|---|
| 500 | 0.053 | 0.193 | 0.122 | 0.197 |
| 1000 | 0.054 | 0.342 | 0.288 | 0.335 |
| 2500 | 0.051 | 0.891 | 0.967 | 0.940 |
| 5000 | 0.055 | 0.998 | 1.000 | 0.998 |

A comparison with the Wald statistic of Beresteanu and Molinari (2008) shows that both statistics have similar size and power properties. The marked empirical likelihood test appears on average to have higher power, but the difference is marginal; in particular, the confidence regions are around 3.5% shorter. Because the results are so similar, we do not report the additional simulations here.

## Appendix A. Assumptions and some definitions

Let $\mathbb{G}_n f(\cdot) = n^{-1/2}\sum_{i=1}^n (f(X_i) - E[f(X)])$ be the empirical process. Hereafter "w.p.a.1" means "with probability approaching one". For Theorem 1 and Proposition 1, we impose the following assumptions.

**Assumption M.**

**(i):** *$\{X_i\}_{i=1}^n$ is an i.i.d. sequence of compact and convex SVRSs. The class $\{s(X,p) : p \in \mathbb{S}^d\}$ is a $\mu$-Donsker class with envelope $F$ such that $E[|F|^\xi] < \infty$ for some $\xi > 2$. Also,* $\inf_{p\in\mathbb{S}^d} \mathrm{Var}(s(X,p)) > 0$.

**(ii):** *$\hat{\nu} \xrightarrow{p} \nu$, $\|\Theta_0(\hat{\nu})\|_H = O_p(1)$, and there exists a function $G(p;\nu)$ continuous in $p \in \mathbb{S}^d$ satisfying (5).*

**(iii):** *For every finite collection of points $\{p_1, \ldots, p_J\} \in \mathbb{S}^d$, the vector $(\mathbb{G}_n s(\cdot,p_1), \ldots, \mathbb{G}_n s(\cdot,p_J), \sqrt{n}(\hat{\nu} - \nu))$ converges in distribution to a Gaussian random vector.*

**Assumption M'.** *For the bootstrap probability $P^*$ conditional on the data, it holds w.p.a.1,*

$$\sup_{p \in \mathbb{S}^d} |s(\Theta_0(\hat{\nu}^*), p) - s(\Theta_0(\hat{\nu}), p) - G(p; \nu)'(\hat{\nu}^* - \hat{\nu})| = o_{p^*}(n^{-1/2}).$$

For Theorem 2, we restrict attention to the situation where $\nu = f(E[z])$ is a smooth function of means of $z \in \mathbb{R}^{\dim(z)}$. A consistent estimator of $\nu$ is given by $\hat{\nu} = f(\bar{z})$. We introduce the following notation: Let $m_k(X_i)$, $\tilde{m}_k(X_i)$, $\dot{m}_k(X_i)$, and $\hat{m}_k(X_i)$ be $k$-dimensional vectors whose $j$-th elements are given by

$$\begin{aligned}
m_{k,j}(X_i) &= s(X_i, p_j) - s(\Theta_0(\hat{\nu}), p_j), \quad \tilde{m}_{k,j}(X_i) = s(X_i, p_j) - s(\Theta_0(\nu), p_j), \\
\dot{m}_{k,j}(X_i) &= s(X_i, p_j) - s(\Theta_0(\nu), p_j) - G(p_j; \nu)' \nabla f(E[z])'(z_i - E[z]), \\
\hat{m}_{k,j}(X_i) &= s(X_i, p_j) - s(\Theta_0(\hat{\nu}), p_j) - G(p_j; \hat{\nu})' \nabla f(\bar{z})'(z_i - \bar{z}),
\end{aligned}$$

respectively. Define $\hat{V}_k = n^{-1} \sum_{i=1}^n m_k(X_i) m_k(X_i)'$, $V_k = \mathrm{Var}(\tilde{m}_k(X_i))$, $\dot{V}_k = \mathrm{Var}(\dot{m}_k(X_i))$, $\bar{V}_k = n^{-1} \sum_{i=1}^n \hat{m}_k(X_i) \hat{m}_k(X_i)'$, $\dot{\phi}_k = \lambda_{\min}(\dot{V}_k)$, and $\bar{\phi}_k = \lambda_{\min}(\bar{V}_k)$. The test statistic $L_n$ in (9) is defined as the maximum over a shrinking neighborhood $\Lambda_n = \{\gamma \in \mathbb{R}^k : \|\gamma\| \leq C \bar{\phi}_k^{-3/2} \sqrt{k/n}\}$ for some positive constant $C$. In particular, $C$ is chosen to satisfy $C > \max\{2C' \bar{\phi}_k^{1/2}, 1\}$ where $C'$ is the positive constant obtained from $\|\bar{m}\| \leq C' \sqrt{k/n}$ w.p.a.1. The condition on $C$ ensures that the local maximum $\hat{\gamma}$ lies in the interior of $\Lambda_n$ w.p.a.1 even in the case when $\dot{\phi}_k^{-1}$ is bounded. If $\dot{\phi}_k^{-1}$ diverges to infinity, this additional condition on $C$ may be dispensed with. Note that the optimization in (9) is well defined only in the region $S_n = \{\gamma \in \mathbb{R}^k : \gamma' m_k(X_i) > -1 \text{ for all } i = 1, \ldots, n\}$. However, since our assumptions guarantee $\max_{1 \leq i \leq n} \sup_{\gamma \in \Lambda_n} |\gamma' m_k(X_i)| = o_p(1)$, it holds that $\Lambda_n \subseteq S_n$ w.p.a.1. For Theorem 2, we impose the following assumptions.

**Assumption S.**

**(i):** *Assumption M holds with the envelope function $F$ in Assumption M (i) satisfying $E[|F|^{\xi}] < \infty$ for some $\xi \geq 4$.*

**(ii):** *$\nabla f(\cdot)$ is Hölder continuous of order $\alpha \geq 2/3$ in a neighborhood of $E[z]$. Furthermore, $E[\|z\|^4] < \infty$.*

**(iii):** *For some neighborhood $\mathcal{N}$ of $\nu$, there exists a function $G(\cdot;.) : \mathbb{S}^d \times \mathcal{N} \to \mathbb{R}^{dim(\nu)}$ such that $\sup_{p\in\mathbb{S}^d} \|G(p;\nu_m) - G(p;\nu)\| \to 0$ for all $\nu_m \to \nu$, where $G(p;\nu)$ is defined in Assumption M (ii). Furthermore, for all $\tilde{\nu} \in \mathcal{N}$, $\sup_{p\in\mathbb{S}^d} \|G(p;\tilde{\nu}) - G(p;\nu)\| \leq M \|\tilde{\nu} - \nu\|^{\alpha}$ for some $\alpha \geq 2/3$ and $M < \infty$ independent of $\tilde{\nu}$.*

**(iv):** *$k \to \infty$ and $(k^5 \dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$ as $n \to \infty$.*

## References


[1] Andrews, D. W. K. and X. Shi (2015) Inference based on conditional moment inequalities, *Econometrica*, 81, 609-666.

[2] Balke, A. and J. Pearl (1997) Bounds on treatment effects from studies with imperfect compliance, *Journal of the American Statistical Association*, 92, 1171-1176.

[3] Beresteanu, A., Molchanov, I. and F. Molinari (2012) Partial identification using random set theory, *Journal of Econometrics*, 166, 17-32.

[4] Beresteanu, A. and F. Molinari (2008) Asymptotic properties for a class of partially identified models, *Econometrica*, 76, 763-814.

[5] Bontemps, C., Magnac, T. and E. Maurin (2012) Set identified linear models, *Econometrica*, 80, 1129-1155.

[6] Canay, I. A. (2010) EL inference for partially identified models: large deviations optimality and bootstrap validity, *Journal of Econometrics*, 156, 408-425.

[7] Chan, N. H., Chen, S. X., Peng, L. and C. L. Yu (2009) Empirical likelihood methods based on characteristic functions with applications to Levy processes, *Journal of the American Statistical Association*, 104, 1621-1630.

[8] Chandrasekhar, A., Chernozhukov, V., Molinari, F. and P. Schrimpf (2012) Inference for best linear approximations to set identified functions, *arXiv preprint arXiv:1212.5627.*

[9] Chen, S. X., Härdle, W. and M. Li (2003) An empirical likelihood goodness-of-fit test for time series, *Journal of the Royal Statistical Society*, B, 65, 663-678.

[10] Chernozhukov, V., Kocatulum, E. and K. Menzel (2015) Inference on sets in finance, *Quantitative Economics*, 6, 309-358.

[11] Cressie, N. and F. L. Hulting (1992) A spatial statistical analysis of tumor growth, *Journal of the American Statistical Association*, 87, 272-283.

[12] Donald, S. G., Imbens, G. W. and W. K. Newey (2003) Empirical likelihood estimation and consistent tests with conditional moment restrictions, *Journal of Econometrics*, 117, 55-93.

[13] Fan, J. and L.-S. Huang (2001) Goodness-of-fit tests for parametric regression models, *Journal of the American Statistical Association*, 96, 640-652.

[14] Fan, J. and J. Zhang (2004) Sieve empirical likelihood ratio tests for nonparametric functions, *Annals of Statistics*, 32, 1858-1907.

[15] Fan, J., Zhang, C. and J. Zhang (2001) Generalized likelihood ratio statistics and Wilks phenomenon, *Annals of Statistics*, 29, 153-193.

[16] Fisher, N. I., Hall, P., Turlach, B. A. and G. S. Watson (1997) On the estimation of a convex set from noisy data on its support function, *Journal of the American Statistical Association*, 92, 84-91.

[17] Giné, E. and J. Zinn (1990) Bootstrapping general empirical measures, *Annals of Probability*, 18, 851-869.

[18] Härdle, W. and E. Mammen (1993) Comparing nonparametric versus parametric regression fits, *Annals of Statistics*, 21, 1926-1947.

[19] Hjort, N. L., McKeague, I. W. and I. van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079-1111.

[20] Horowitz, J. L. and C. F. Manski (2000) Nonparametric analysis of randomized experiments with missing covariate and outcome data, *Journal of the American Statistical Association*, 95, 77-84.

[21] Kaido, H. (2012) A dual approach to inference for partially identified econometric models, Working Paper. Boston University.

[22] Kaido, H. and A. Santos (2014) Asymptotically efficient estimation of models defined by convex moment inequalities, *Econometrica,* 82, 387-413.

[23] Kendall, D. G. (1974) Foundations of a theory of random sets, in Harding, E. F. and D. G. Kendall (eds.), *Stochastic Geometry*, 322-376, Wiley, New York.

[24] Lahiri, S. N. (1992) Bootstrapping M-estimators of a multiple linear regression parameter, *Annals of Statistics*, 20, 1548-1570.

[25] Li, G. (2003) Nonparametric likelihood ratio goodness-of-fit tests for survival data, *Journal of Multivariate Analysis*, 86, 166-182.

[26] Manski, C. F. (2003) *Partial Identification of Probability Distributions*, Springer.

[27] Matheron, G. (1975) *Random Sets and Integral Geometry*, Wiley, New York.

[28] Molchanov, I. (2005) *Theory of Random Sets*, Springer.

[29] Molchanov, I. and F. Molinari (2014) Applications of random set theory in econometrics, *Annual Review of Economics*, 6, 229-251.

[30] Owen, A. B. (2001) *Empirical Likelihood*, Chapman & Hall/CRC.

[31] Qin, J. and J. Lawless (1994) Empirical likelihood and general estimating equations, *Annals of Statistics*, 22, 300-325.

[32] Shorack, G. (1982) Bootstrapping robust regression, *Communications in Statistics: Theory and Methods*, 11, 961-972.

[33] Stoyan, D. (1998) Random sets: models and statistics, *International Statistical Review*, 1998, 66, 1-27.

[34] Stute, W. (1997) Nonparametric model checks for regression, *Annals of Statistics*, 25, 613-641.

[35] Stute, W., Gonzalez-Manteiga, W. and M.P. Quindimil (1998) Bootstrap approximations in model checks for regression, *Journal of the American Statistical Association*, 93, 141-149.

[36] Tamer, E. (2010) Partial identification in econometrics, *Annual Review of Economics*, 2, 167-195.

[37] van der Vaart, A. W. (1998) *Asymptotic Statistics*, Cambridge University Press.

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK. k.adusumilli@lse.ac.uk

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK. t.otsu@lse.ac.uk

# WEB APPENDIX TO "EMPIRICAL LIKELIHOOD FOR RANDOM SETS"

Abstract. Sections A and B present proofs for Theorems 1 and 2 from the main text, respectively. Section C reports additional numerical results for Section 4.1. Section D provides additional simulation results to illustrate the empirical likelihood test proposed in Section 3.5.

## Appendix A. Proof of Theorem 1

We first derive the limiting distribution of $K_n$ under $\mathrm{H}_0$. By Assumption M (ii),

$$n^{-1/2}\sum_{i=1}^{n}\{s(X_i,p)-s(\Theta_0(\hat{\nu}),p)\}=\mathbb{G}_n s(\cdot,p)-G(p;\nu)'(\hat{\nu}-\nu)+o_p(n^{-1/2}),$$

uniformly over $p\in\mathbb{S}^d$. Assumptions M (i) and (iii) guarantee weak convergence of the process $\{\mathbb{G}_n s(\cdot,p),\sqrt{n}(\hat{\nu}-\nu):p\in\mathbb{S}^d\}$ to $\{Z(p),Z_1:p\in\mathbb{S}^d\}$. Thus, by continuity of $G(p;\nu)$ (Assumption M(ii)), it follows that $n^{-1/2}\sum_{i=1}^{n}\{s(X_i,p)-s(\Theta_0(\hat{\nu}),p)\}$ converges weakly to $Z(p)-G(p;\nu)'Z_1$. Using Assumptions M (i) and (ii) and standard arguments, $\sup_{p\in\mathbb{S}^d}|n^{-1}\sum_{i=1}^{n}\{s(X_i,p)-s(\Theta_0(\hat{\nu}),p)\}^2-\mathrm{Var}(s(X,p))|\xrightarrow{p}0$. From the envelope condition in Assumption M (i) and a Borel-Cantelli lemma argument as in Owen (1988), it holds $\max_{1\le i\le n}\sup_{p\in\mathbb{S}^d}|s(X_i,p)|=o(n^{1/2})$ almost surely. This, along with $\|\Theta_0(\hat{\nu})\|_H=O_p(1)$ (Assumption M (ii)), implies $\max_{1\le i\le n}\sup_{p\in\mathbb{S}^d}|s(X_i,p)-s(\Theta_0(\hat{\nu}),p)|=o_p(n^{1/2})$. Combining these results, the null distribution of $K_n$ follows by a similar argument as in the proof of Hjort, McKeague and van Keilegom (2009, Theorem 2.1).

We now prove the second assertion, $K_n\to\infty$ under $\mathrm{H}_1$. Let $g_i(p,t)=s(X_i,p)-s(\Theta_0(t),p)$ for $t=\nu$ or $\hat{\nu}$. Under $\mathrm{H}_1$, there exists $p^*\in\mathbb{S}^d$ such that $E[g_i(p^*,\nu)]\neq 0$. We prove the case of $E[g_i(p^*,\nu)]>0$ only; the case of $E[g_i(p^*,\nu)]<0$ can be shown in the same manner. Pick any $\delta \in (0, 1/2)$. Observe that

$$\begin{aligned} -\log \ell_n(p^*) &= \sup_{\lambda \in \mathbb{R}} \sum_{i=1}^n \log(1 + \lambda g_i(p^*, \hat{\nu})) \geq \sum_{i=1}^n \log(1 + n^{-(1/2+\delta)} g_i(p^*, \hat{\nu})) \\ &= n^{1/2-\delta} \left\{ \frac{1}{n} \sum_{i=1}^n g_i(p^*, \hat{\nu}) \right\} + n^{-2\delta} \left\{ \frac{1}{2n} \sum_{i=1}^n g_i(p^*, \hat{\nu})^2 \right\} + O_p(n^{-2\delta}), \end{aligned}$$

where the first equality follows from the convex duality and the second equality follows from a Taylor expansion. Since the first term diverges to infinity and the other terms are negligible under Assumptions M (i)-(iii), the conclusion is obtained.

## Appendix B. Proof of Theorem 2

We first derive the limiting distribution of $(L_n - k)/\sqrt{2k}$ under $\mathrm{H}_0$. Define $\dot{g}_i(p) = s(X_i, p) - s(\Theta_0(\nu)) - G(p; \nu)' \nabla f(E[z])'(z_i - E[z])$, $\bar{m}_k = n^{-1} \sum_{i=1}^n m_k(X_i)$, and $\bar{\dot{m}}_k = n^{-1} \sum_{i=1}^n \dot{m}_k(X_i)$. Note that by the mean value theorem (applicable here by Assumption S (iii)), for each $p \in \mathbb{S}^d$ there exists some $\tilde{\nu}_p$ satisfying $\|\tilde{\nu}_p - \nu\| \leq \|\hat{\nu} - \nu\|$ and $s(\Theta_0(\hat{\nu}), p) - s(\Theta_0(\nu), p) = G(p; \tilde{\nu}_p)'(\hat{\nu} - \nu)$. Thus by Assumption S (ii) and the asymptotic expansion $\hat{\nu} - \nu = \nabla f(E[z])' n^{-1} \sum_{i=1}^n (z_i - E[z]) + O_p(n^{-(1+\alpha)/2})$, we have

$$\begin{aligned} \|\bar{m}_k - \bar{\dot{m}}_k\| &\leq \sqrt{k} \sup_{p \in \mathbb{S}^d} \|s(\Theta_0(\hat{\nu}), p) - s(\Theta_0(\nu), p) - G(p; \nu)'(\hat{\nu} - \nu)\| + O_p(\sqrt{k/n^{1+\alpha}}) \\ &\leq \sqrt{k}\, \|\hat{\nu} - \nu\| \sup_{p \in \mathbb{S}^d} \|G(p; \tilde{\nu}_p) - G(p; \nu)\| + O_p(\sqrt{k/n^{1+\alpha}}) = O_p(\sqrt{k/n^{1+\alpha}}) \quad (1) \end{aligned}$$

Also note that

$$\bar{\dot{m}}_k = O_p(\sqrt{k/n}), \qquad \bar{m}_k = O_p(\sqrt{k/n}), \tag{2}$$

where the first statement follows from the fact that the process $\{\dot{g}_i(p); p \in \mathbb{S}^d\}$ is $\mu$-Donsker by Assumption S (i), and the second statement follows by (1). Next, observe that

$$\left\|\hat{V}_k - V_k\right\| \le k \sup_{p,q\in\mathbb{S}^d}\left|\frac{1}{n}\sum_{i=1}^{n}\{\dot{g}_i(p)\dot{g}_i(q) - E[\dot{g}_i(p)\dot{g}_i(q)]\}\right| + O_p\left(\sqrt{\frac{k}{n}}\right) = O_p(k/\sqrt{n}) \quad (3)$$

where the inequality follows from $\sup_{p\in\mathbb{S}^d}|s(\Theta(\hat{\nu}),p) - s(\Theta(\nu),p)| = O_p(n^{-1/2})$ and the equality follows from the fact that the process$\{\dot{g}_i(p)\dot{g}_i(q); p,q\in\mathbb{S}^d\}$ is $\mu$-Donsker. Furthermore, using Assumptions S (i) and (ii) combined with $\|\bar{z} - E[z]\| = O_p(n^{-1/2})$, straightforward algebra ensures that

$$\|\dot{m}_k(X_i) - \hat{m}_k(X_i)\| = O_p(\sqrt{k/n^\alpha})\,\|z_i - E[z]\| + O_p(\sqrt{k/n}).$$

We can now see that $\bar{V}_k - n^{-1}\sum_{i=1}^n \dot{m}_k(X_i)\dot{m}_k(X_i)'$ is bounded by $2n^{-1}\sum_{i=1}^n\{k^{1/2}\dot{g}_i\delta_i + \delta_i^2\}$, where $\delta_i = \|\dot{m}_k(X_i) - \hat{m}_k(X_i)\|$. Substituting the expression for the latter from the previous equation and noting that our assumptions guarantee $E[\dot{g}_i^2] < \infty$, we obtain $\left\|\bar{V}_k - n^{-1}\sum_{i=1}^n \dot{m}_k(X_i)\dot{m}_k(X_i)'\right\| = O_p(\sqrt{k^2/n^\alpha})$ using the law of large numbers. Moreover, $\left\|n^{-1}\sum_{i=1}^n \dot{m}_k(X_i)\dot{m}_k(X_i)' - \dot{V}_k\right\| = O_p(k/\sqrt{n})$ by analogous weak convergence arguments as used to show (3). Combining these results proves

$$\left\|\bar{V}_k - \dot{V}_k\right\| = O_p(\sqrt{k^2/n^\alpha}). \quad (4)$$

We also make frequent use of the following fact implied by (4) and the rate condition $(k^5\dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$:

$$|\bar{\phi}_k^c - \dot{\phi}_k^c| = o_p(\dot{\phi}_k^c) \quad \text{for each } c \in \mathbb{R}. \quad (5)$$

For the conclusion of this theorem, it is sufficient to show the followings:

$$\frac{L_n(\hat{\nu}) - n\bar{m}_k'\bar{V}_k^{-1}\bar{m}_k}{\sqrt{2k}} \xrightarrow{p} 0, \tag{6}$$

$$\frac{n\bar{m}_k'\bar{V}_k^{-1}\bar{m}_k - k}{\sqrt{2k}} \xrightarrow{d} N(0,1). \tag{7}$$

We first show (6). Let $\hat{\gamma} \in \arg\max_{\gamma\in\Lambda_n} G_n(\gamma)$ and $D_n = \max_{1\le i\le n}\|m_k(X_i)\|$. Also define $G_n^*(\gamma) = n(2\gamma'\bar{m}_k - \gamma'\bar{V}_k\gamma)$, which is maximized at $\gamma^* = \bar{V}_k^{-1}\bar{m}_k$. For (6), it is sufficient to show that $\hat{\gamma}, \gamma^* = O_p(\dot{\phi}_k^{-1}\sqrt{k/n})$, and $\sup_{\gamma\in\Omega_n\subseteq\Lambda_n} k^{-1/2}|G_n(\gamma) - G_n^*(\gamma)| \xrightarrow{p} 0$ where $\Omega_n = \{\gamma \in \mathbb{R}^k : \|\gamma\| \le c\dot{\phi}_k^{-1}\sqrt{k/n}\}$ with $c > 0$ chosen to ensure $\Omega_n$ contains both $\hat{\gamma}$ and $\gamma^*$ w.p.a.1 and $\Omega_n \subseteq \Lambda_n$ (such a $c$ exists by the definition of $\Lambda_n$). Indeed, these are shown by an argument similar to the proof of Hjort, McKeague and van Keilegom (2009, Proposition 4.1) if the following requirements are satisfied under $(k^5\dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$:

$$(n^{-1/2}k^{3/2}\dot{\phi}_k^{-3})D_n = o_p(1), \tag{8}$$

$$\|\gamma^*\| = O_p(\dot{\phi}_k^{-1}\sqrt{k/n}), \tag{9}$$

$$\lambda_{\max}(\hat{V}_k) = O_p(k), \tag{10}$$

$$\hat{\gamma} \text{ exists w.p.a.1 and } \|\hat{\gamma}\| = O_p(\dot{\phi}_k^{-1}\sqrt{k/n}). \tag{11}$$

We first show (8). Using the envelope condition in Assumption S (i) which implies $\sup_{k\in\mathbb{N}} E[\|k^{-1/2}\tilde{m}_k(X_i)\|^\xi] < \infty$, an argument similar to the proof of Hjort, McKeague and van Keilegom (2009, Lemma 4.1) guarantees $(n^{-1/2}k^{3/2}\dot{\phi}_k^{-3})\max_{1\le i\le n}\|\tilde{m}_k(X_i)\| = o_p(1)$ under the rate condition $(k^5\dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$. Furthermore, $\max_{1\le i\le n}\|\tilde{m}_k(X_i) - m_k(X_i)\| \le \sup_{p\in\mathbb{S}^d}|s(\Theta(\hat{\nu}),p) - s(\Theta(\nu),p)| = O_p(n^{-1/2})$, and (8) follows. Next, (9) follows from (2) and (5). To show (10), observe that $\left\|\hat{V}_k - n^{-1}\sum_{i=1}^n \tilde{m}_k(X_i)\tilde{m}_k(X_i)'\right\| = O_p(k/\sqrt{n})$ by Assumption S (ii) and

$\|n^{-1}\sum_{i=1}^n \tilde{m}_k(X_i)\tilde{m}_k(X_i)'\| = O_p(k)$ by $E[\|X_i\|_H^2] < \infty$. Hence, using $\lambda_{\max}(\hat{V}_k) \leq \left\|\hat{V}_k\right\|$ and the triangle inequality, (10) is verified. Finally, for (11), we first note that $\hat{\gamma}$ exists w.p.a.1 since $\Lambda_n \subseteq S_n$ w.p.a.1 and $\Lambda_n$ is a compact set. Thus, letting $b_n = \max_{1\leq i\leq n}\sup_{\gamma\in\Lambda_n}\{1-(1+\gamma' m_k(X_i))^{-2}\}$, an expansion around $\gamma = 0$ yields $0 \leq G_n(\hat{\gamma}) \leq n\{2\hat{\gamma}'\bar{m}_k - \hat{\gamma}'(\bar{V}_k - b_n\hat{V}_k)\hat{\gamma}\}$. Note that

$$b_n = O_p\left(\max_{1\leq i\leq n}\sup_{\gamma\in\Lambda_n}|\gamma' m_k(X_i)|\right) = O_p\left(D_n\sup_{\gamma\in\Lambda_n}\|\gamma\|\right) = o_p(\dot{\phi}_k^{3/2}k^{-1}),$$

where the last equality follows from (5), (8) and the definition of $\Lambda_n$. Consequently, $\lambda_{\min}(\bar{V}_k - b_n\hat{V}_k) \geq \bar{\phi}_k - |b_n|\lambda_{\max}(\hat{V}_k) = \dot{\phi}_k(1+o_p(1))$, where the equality also uses (10) and (5). Thus $\hat{\gamma}'(\bar{V}_k + b_n\hat{V}_k)\hat{\gamma} \geq \|\hat{\gamma}\|^2\,\dot{\phi}_k(1+o_p(1))$, which implies $\|\hat{\gamma}\| \leq 2\dot{\phi}_k^{-1}\|\bar{m}_k\|(1+o_p(1))$. Therefore, by (2) it must be the case that $\hat{\gamma}$ is an interior solution w.p.a.1. (by the choice of $C$ in the definition of $\Lambda_n$) and that $\|\hat{\gamma}\| = O_p(\dot{\phi}_k^{-1}\sqrt{k/n})$. This proves (11). Combining these results, the claim in (6) follows.

We now show (7). We can decompose

$$\begin{aligned}\frac{n\bar{m}_k'\bar{V}_k^{-1}\bar{m}_k - k}{\sqrt{2k}} &= \frac{n\bar{m}_k'(\bar{V}_k^{-1} - \dot{V}_k^{-1})\bar{m}_k}{\sqrt{2k}} + \frac{n(\bar{m}_k - \bar{\dot{m}}_k)'\dot{V}_k^{-1}\bar{m}_k}{\sqrt{2k}} \\ &\quad + \frac{n\bar{\dot{m}}_k'\dot{V}_k^{-1}(\bar{m}_k - \bar{\dot{m}}_k)}{\sqrt{2k}} + \frac{n\bar{\dot{m}}_k'\dot{V}_k^{-1}\bar{\dot{m}}_k - k}{\sqrt{2k}}. \end{aligned} \tag{12}$$

By de Jong and Bierens (1994, Lemma 4a), the first term of (12) is bounded by $nk^{-1/2}\|\bar{m}_k\|^2\bar{\phi}_k^{-1}\dot{\phi}_k^{-1}\left\|\bar{V}_k - \dot{V}_k\right\|$ and is thus negligible using (2),(4) and (5). Next, by (1),(2) and (5) the second term of (12) is bounded by $n\dot{\phi}_k^{-1}\|\bar{m}_k - \bar{\dot{m}}_k\|\,\|\bar{m}_k\|/\sqrt{2k} = O_p(\dot{\phi}_k^{-1}\sqrt{k/n^{\alpha}})$ which is negligible for $\alpha \geq 1/3$. Negligibility of the third term of (12) follows by a similar argument. Finally, note that $E[\dot{m}_k(X_i)] = 0$ and $\mathrm{Var}(\dot{m}_k(X_i)) = \dot{V}_k$. Therefore, arguing as in the proof of de Jong and Bierens (1994, Theorem 1), the last term of (12) converges in distribution to $N(0,1)$ under the rate condition $\dot{\phi}_k^{-4}k^2/n \to 0$. Thus the result in (7) follows.

We now prove the second assertion, $(L_n - k)/\sqrt{2k} \to \infty$ under $\mathrm{H}_1$. Since in the limit the points $\{p_1, \ldots, p_k\}$ form a dense subset of $\mathbb{S}^d$ and the support function is continuous, under $\mathrm{H}_1$ there exists an integer $N$ such that for all $n \geq N$ the set of points includes a direction $p^*$ for which $E[s(X_i, p^*) - s(\Theta_0(\nu), p^*)] \neq 0$. Without loss of generality we prove the case of $E[s(X_i, p^*) - s(\Theta_0(\nu), p^*)] > 0$. Define $g_i(p) = s(X_i, p) - s(\Theta_0(\hat{\nu}), p)$ and $\bar{g}_i(p) = s(X_i, p) - s(\Theta_0(\hat{\nu}), p) - G(p; \hat{\nu})'\nabla f(\bar{z})'(z_i - \bar{z})$. Pick any $\delta \in (0, 0.3)$ and observe

$$\begin{aligned} L_n &\geq 2\sum_{i=1}^n \log(1 + n^{-(1/2+\delta)} g_i(p^*)) + n^{-2\delta}\Big\{\frac{1}{n}\sum_{i=1}^n g_i(p^*)^2 - \frac{1}{n}\sum_{i=1}^n \bar{g}_i(p^*)^2\Big\} \\ &= 2n^{1/2-\delta}\left\{\frac{1}{n}\sum_{i=1}^n g_i(p^*)\right\} - n^{-2\delta}\left\{\frac{1}{n}\sum_{i=1}^n \bar{g}_i(p^*)^2\right\} + O_p(n^{-2\delta}), \end{aligned}$$

for all $n \geq N$, where the inequality follows by setting $\gamma = n^{-(1/2+\delta)}e^* \in \Gamma_n$ w.p.a.1, where $e^*$ is the unit vector that selects the component of $m_k(X_i)$ containing $p^*$, and the equality follows from a Taylor expansion. Now, $n^{-1}\sum_{i=1}^n g_i(p^*) \xrightarrow{p} E[s(X_i, p^*)] - s(\Theta_0(\nu), p^*) \neq 0$ by a suitable law of large numbers and $n^{-1}\sum_{i=1}^n \bar{g}_i(p^*)^2 \xrightarrow{p} E[\dot{g}_i(p^*)^2] < \infty$ by a similar argument used to show (4). Thus, $L_n$ diverges to infinity at the rate $n^{1/2-\delta}$ which implies that $(L_n - k)/\sqrt{2k}$ diverges.

## Appendix C. Additional numerical results for Section 4.1

In this section we report additional numerical results to compare the marked empirical likelihood confidence region obtained in Section 4.1 with the one based on the method by Chernozhukov, Kocatulum and Menzel (2015) (hereafter CKM). As in Section 4.1, we consider the relationship between the unobservable dependent variable $y$ and regressors $x$, where we observe the interval $[y_L, y_U]$ satisfying $y_L \leq y \leq y_U$ almost surely.

Consider the set of coefficients characterized by the conditional moment inequalities

$$\Xi = \{\theta : \ E[y_L|x] \le (1, x')\theta \le E[y_U|x]\}.$$

The set $\Xi$ would be the identified region of interest if we assume $E[y|x] = (1, x')\theta$. It is important to note that the set $\Xi$ is a subset of

$$\Upsilon = \{\arg\min_{\theta} \int \{y - (1, x')\theta\}^2 d\mu \text{ for some } \mu \in \mathcal{M}\},$$

which is the identified region of interest in Section 4.1. Indeed, this can be seen from the fact that $\Upsilon$ is obtained as the set of parameters satisfying $E[(1, x')\{y - (1, x')\theta\}] = 0$.

If all the regressors $x$ are discrete, then $\Xi$ is characterized by a finite number of moment inequalities (see, Andrews and Shi, 2015, for a general case). CKM suggest a general approach to obtain confidence regions in this context by combining the moment inequalities into a single one using the smooth-max approximation (see, Section 3.4). In our numerical example with log wages and education, the education variable takes 13 values and thus provides 26 moment inequalities. Since it is computationally difficult to work with a smooth-max approximation with such a large number of moments, we simplify the problem by partitioning the regressor values into four bins and utilizing the moment inequalities within each bin (corresponding to a total of 8 moment inequalities). In particular, we partition the education variable into the following broad categories: Less than $10^{\text{th}}$ grade ($x \le 10$); High school graduate ($x \in [11, 12]$); Some college or associate degree including vocational training ($x \in [13, 14]$); and Bachelor's degree or higher ($x \ge 15$).

Figure 1 compares the 95% confidence region of CKM for $\Xi$ with that from the marked empirical likelihood for $\Upsilon$. The sample size is $n = 1000$. The tuning parameter $\varrho$ for the 'smooth-max' approximation is chosen to be $\varrho = 100$. The critical values in both cases

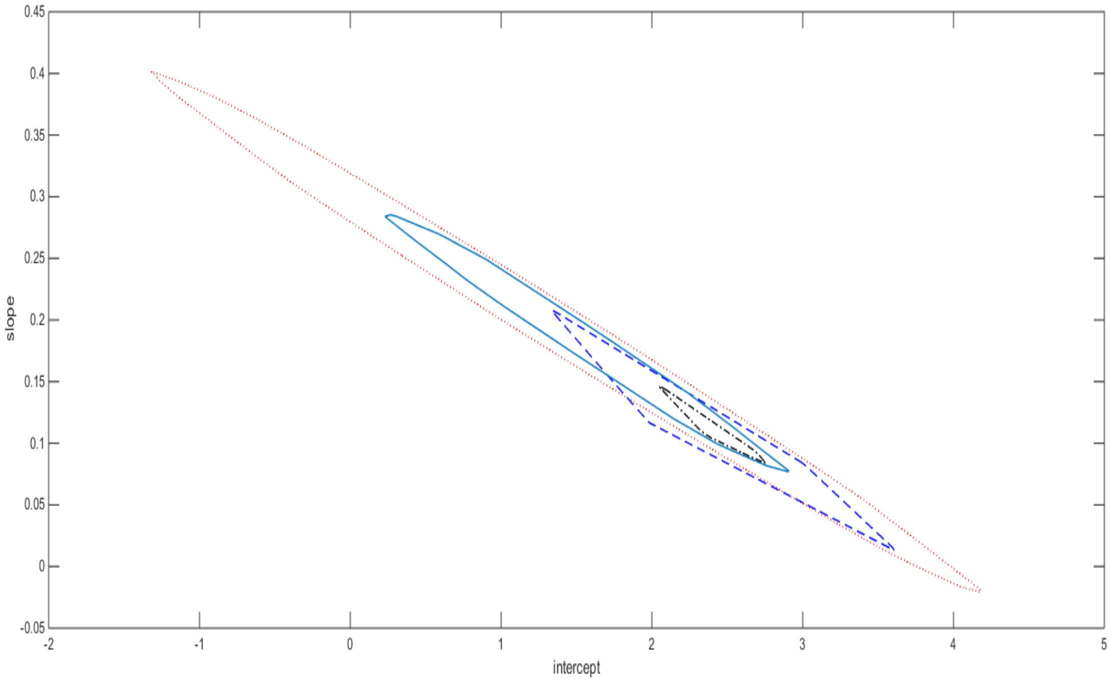


Figure 1. The population identification regions for regression with interval outcomes $\Xi$ (dash-dotted line) and for the best linear prediction $\Upsilon$ (solid line) as well as the corresponding 95% confidence regions via CKM (dashed line) and the marked empirical likelihood statistic (dotted line). The sample size is $n = 1000$.

were obtained using bootstrap with 999 repetitions. Unsurprisingly, the CKM confidence region is smaller than that obtained by the marked empirical likelihood. This is due to the fact that the region $\Xi$ is considerably smaller than $\Upsilon$ as can be seen from Figure 1. From Figure 1, we can thus infer the following: If it is possible to impose additional assumptions to satisfy the conditional moment restriction $E[y|x] = x'\theta$, then characterizing the set using moment inequalities leads to a much smaller confidence region. At the same time, the best linear predictor is more robust to possible misspecification and thus, is applicable more generally, albeit at the expense of a larger confidence set.

## Appendix D. Simulation results for Section 3.5

We consider the problem of testing the shape of a set based on noisy measurements of the support function, as discussed in Section 3.5. We employ the simulation design of

Fisher *et al.* (1997), where the underlying set is an ellipse relative to the origin with the support function taking the form $s(\Theta, p) = (\theta_1^2 \cos^2 p + \theta_2^2 \sin^2 p)^{1/2}$ for $p \in [-\pi, \pi]$. Noisy measurements $\{s_i, p_i\}_{i=1}^n$ of the support function are generated using $s_i = s(\Theta, p_i) + \epsilon_i$ with $p_i \sim \text{Uniform}[-\pi, \pi]$ and $\epsilon_i \sim N(0, 0.16)$.

We consider two types of testing problems here. First, we test whether the set $\Theta$ takes a particular shape. In the first four columns of Table 1, we report the rejection frequencies of the marked empirical likelihood test based on eq. (12) of the paper for the null hypotheses $\mathrm{H}_0^a$ :'$\Theta$ is a circle with $(\theta_1, \theta_2) = (1, 1)$' and $\mathrm{H}_0^b$ :'$\Theta$ is an ellipse with $(\theta_1, \theta_2) = (1, 2)$'. To compute the test statistic we follow Fisher *et al.* (1997) in employing the von Mises density function $K_b(z) = e^{b \cos z} / \int_{-\pi}^{\pi} e^{b \cos z} dz$ on the circle as the kernel and set the smoothing parameter to be $b = 8$ (which corresponds to the inverse of the square of the bandwidth for the conventional kernel density estimator). In the last two rows of Table 1 we present the results for different values of the bandwidth by setting $b = 4$ and 16 when $n = 200$. The critical value of the test is computed using the wild bootstrap based on Härdle and Mammen (1993). We consider sample sizes of $n$ =100, 200, and 500. The number of Monte Carlo replications is 1000 for all cases. The first and third columns of Table 1 indicate that the marked empirical likelihood test based on eq. (12) of the paper has reasonable size properties for both null hypotheses and over all sample sizes. The second and fourth columns evaluate power properties of the test against the alternatives $\mathrm{H}_1^a : (\theta_1, \theta_2) = (1.1, 1)$ and $\mathrm{H}_1^b : (\theta_1, \theta_2) = (1.1, 2)$, respectively. In both cases, the power of the empirical likelihood test increases with the sample size at a reasonably fast rate.

Second, we conduct a goodness-of-fit test for the null $\mathrm{H}_0^c$ :'$\Theta$ is a ellipse with $s(\Theta, p) = (\theta_1^2 \cos^2 p + \theta_2^2 \sin^2 p)^{1/2}$ for some $(\theta_1, \theta_2)$'. For this testing problem, $(\theta_1, \theta_2)$ are nuisance parameters to be estimated. The marked empirical likelihood statistic is modified by

TABLE 1. Rejection frequencies of the marked empirical likelihood test at the nominal 5% level

| $n, b$ | $\mathrm{H}_0^a$ :circle | $\mathrm{H}_1^a$ | $\mathrm{H}_0^b$ :ellipse | $\mathrm{H}_1^b$ | $\mathrm{H}_0^c$ | $\mathrm{H}_1^c$ |
|---|---|---|---|---|---|---|
| 100, 8 | 0.022 | 0.425 | 0.079 | 0.413 | 0.049 | 0.182 |
| 200, 8 | 0.026 | 0.851 | 0.029 | 0.608 | 0.020 | 0.409 |
| 500, 8 | 0.039 | 0.999 | 0.025 | 0.958 | 0.013 | 0.991 |
| 200, 4 | 0.036 | 0.854 | 0.025 | 0.557 | 0.016 | 0.332 |
| 200, 16 | 0.014 | 0.774 | 0.037 | 0.668 | 0.012 | 0.443 |

replacing $\{s_i - s(\Theta_0, p)\}$ in eq. (12) of the paper with its estimated counterpart $\{s_i - (\hat{\theta}_1^2 \cos^2 p + \hat{\theta}_2^2 \sin^2 p)^{1/2}\}$, where $(\hat{\theta}_1, \hat{\theta}_2)$ is the nonlinear least squares estimator. Under the null $\mathrm{H}_0^c$, the measurements on the support function are generated by $(\theta_1, \theta_2) = (1, 2)$. Under the alternative $\mathrm{H}_1^c$, the data are generated by $s(\Theta, p) = (\cos^2 p + \cos p \sin p + 4 \sin^2 p)^{1/2}$. The critical value is again computed using the wild bootstrap. The last two columns of Table 1 report the rejection frequencies of this test. Although the test is slightly undersized, it shows good size and power performance.

Finally, the last two rows of Table 1 show that the rejection frequencies are not very sensitive to the choice of the smoothing parameter $b$ under the null and alternative hypotheses.

## REFERENCES


[1] Andrews, D. W. K. and X. Shi (2015) Inference based on conditional moment inequalities, *Econometrica*, 81, 609-666.

[2] Chernozhukov, V., Kocatulum, E. and K. Menzel (2015) Inference on sets in finance, *Quantitative Economics*, 6, 309-358.

[3] de Jong, R. M. and H. J. Bierens (1994) On the limit behavior of a chi-square type test if the number of conditional moments tested approaches infinity, *Econometric Theory*, 9, 70-90.

[4] Fisher, N. I., Hall, P., Turlach, B. A. and G. S. Watson (1997) On the estimation of a convex set from noisy data on its support function, *Journal of the American Statistical Association*, 92, 84-91.

[5] Härdle, W. and E. Mammen (1993) Comparing nonparametric versus parametric regression fits, *Annals of Statistics*, 21, 1926-1947.

[6] Hjort, N. L., McKeague, I. W. and I. van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079-1111.

[7] Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, 75, 237-249.