USING CONFIDENCE DISTRIBUTION SAMPLING TO VISUALIZE CONFIDENCE SETS

Daeyoung Kim and Bruce G. Lindsay

University of Massachusetts Amherst and The Pennsylvania State University

Abstract: This paper presents a new sampling-based methodology designed to facilitate the visual analysis of the confidence sets generated by an inference function such as the likelihood. This methodology generates a sample of parameters from a confidence distribution. This distribution is designed so that its probabilities on the parameter space are equal to the asymptotic coverage probabilities of the targeted confidence sets. Plotting these samples provides a picture of the inference function surface around the point estimator optimizing the inference function. Once the sample is created, one can also picture the profile inference function confidence sets for various functions of the parameters, all without further numerical optimization. The result is similar to a Bayesian analysis based on samples from the posterior. One distinction is that we can target the samples to obtain a clearer picture of the confidence set boundary. We illustrate the methodology with four different inference functions.

Although this methodology is related to Fisher's concept of fiducial inference, the fiducial-like confidence distribution we create here is chosen for its ability to recover the confidence sets generated by the inference function and for its ease in computation, nothing more. Unlike resampling methods such as the parametric bootstrap, our method uses only the original data set, as in Bayesian inference. We use illustrative examples to compare our sampling-based confidence sets with those based on numerical optimization, and to compare the confidence regions generated by different inference functions.

Key words and phrases: Confidence distribution, confidence set, empirical likelihood, likelihood ratio statistic, quadratic inference function, score statistic, wald statistic.

1. Introduction

Inference functions, such as the likelihood, are widely used throughout statistics. They have the virtue of providing, in a single package, methods for point estimation, testing hypotheses, and constructing confidence sets. A major challenge with the application of these methods to data is the display of the confidence sets that they generate, as the boundaries of these sets require numerical optimization. In order to visualize the targeted confidence sets for a set of parameters of interest, standard methodology would require a numerical optimization

for each parameter. In this paper we provide a unified method for describing these confidence sets. It is based on a simulation that samples points on the boundaries of the relevant sets. The resulting data set can be used to describe every confidence set of interest without further numerical optimization.

We first describe what we mean by an inference function. It is a function $H(\boldsymbol{\theta}; \mathbf{y})$ of data $\mathbf{y}$ and parameter $\boldsymbol{\theta}$ with the following four properties.

1. A point estimator $\hat{\boldsymbol{\theta}}$ can be constructed by minimizing $H(\boldsymbol{\theta}; \mathbf{y})$.

2. A test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ can be constructed using a test statistic $T_1 = H(\boldsymbol{\theta}_0; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$. We assume the limiting distribution of T_1 does not depend on $\boldsymbol{\theta}_0$. This test can then be inverted to form confidence sets.

3. Let $\hat{\boldsymbol{\theta}}_0^* = \arg\min_{\boldsymbol{\theta}} \{H(\boldsymbol{\theta}; \mathbf{y}) : g(\boldsymbol{\theta}) = \nu_0\}$. A test statistic for the composite null hypothesis $H_0 : g(\boldsymbol{\theta}) = \nu_0$ can be constructed using $T_2 = H(\hat{\boldsymbol{\theta}}_0^*; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$. We assume that the limiting distribution of T_2 does not depend on the true value of $\boldsymbol{\theta}$ in the null hypothesis. One can then invert this test to form confidence sets for $g(\boldsymbol{\theta})$. The function $H^*(\nu; \mathbf{y}) = H(\hat{\boldsymbol{\theta}}_0^*; \mathbf{y}) = \min \{H(\boldsymbol{\theta}; \mathbf{y}) : g(\boldsymbol{\theta}) = \nu\}$ is viewed as a function of ν and called the profile inference function for $g(\boldsymbol{\theta})$. Inversion of these tests yields the profile confidence sets for $g(\boldsymbol{\theta})$.

4. Asymptotic critical values are available for T_1 and T_2 with adjustments to these values for small sample sizes.

There are four commonly used inference functions in statistics that satisfy the four properties mentioned above namely, the likelihood function, the score test statistic (Rao (1948)), the quadratic inference function (Qu (1998); Lindsay and Qu (2003)), and the empirical likelihood (Owen (1988); Qin and Lawless (1994)). We briefly review these four inference functions and their asymptotic properties including asymptotic confidence regions in Section 2.

We are often interested in visualizing the profile confidence sets generated by $H(\boldsymbol{\theta}; \mathbf{y})$ when multiple functions of $\boldsymbol{\theta}$, $\nu = g(\boldsymbol{\theta})$, are of central interest. Finding the confidence sets for such parameters can be computationally demanding because this requires repeated computation of $T_2 = H(\hat{\boldsymbol{\theta}}_0^*; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$ over different constraint sets $g(\boldsymbol{\theta}) = \nu$, with ν varying throughout its range. In other cases we might be interested in describing the geometric features of the entire high-dimensional confidence set, as we demonstrate in a complex mixture example in Section 4.

One might compare these numerical challenges with the simplicity of a Bayesian data analysis in which one uses the samples for the parameters generated from Markov Chain Monte Carlo algorithm to visualize various marginal features of the full posterior density with respect to parameters of interest. In

this paper we create a similar sampling-based mechanism for exploring the inference function confidence sets. Just as in the Bayesian method, we need to be concerned about having an adequate density of sampling so that confidence limits are well described. Note, however, that our method uses independent sampling, so that we need not be concerned with burn-in or convergence.

There exists an alternative method for confidence construction that is sampling-based, the parametric bootstrap (Efron and Tibshirani (1993); Davison and Hinkley (1997); Hall (1992)). In this method, one would repeatedly simulate data from the fitted model, calculate the estimates for each replicate, and then use these samples to construct confidence sets. Beran (1988) compared rates of convergence of the coverage probabilities for parametric bootstrap confidence sets and asymptotic confidence sets, and showed that parametric bootstrap improved asymptotic coverage for given inference function as described above.

A number of authors have discussed the theoretical drawbacks of the bootstrap method as compared to confidence sets based on likelihood; see Owen (2001, p.5) and Lang (2008, p.5976). Even from the computational viewpoint, there are some drawbacks. First, the recomputation of parameter estimators from new data sets can be quite expensive in computing time. Secondly, there is no natural way to generate samples for the parameters describing the boundaries of the targeted confidence sets and so there is an inherent sparsity problem, as we demonstrate in Section 4.1.

We emphasize that the purpose of our methodology is to help users visualize the standard confidence sets that are generated by their particular choice of inference function. There are no new confidence procedures here, just an alternative way to represent standard ones. The advantage of our method is that with a single simulation, one can create profile confidence sets over a wide range of parameters of interest without any further optimization, and can do boundary sampling designed to describe clearer boundaries of the confidence sets. Furthermore, unlike resampling methods such as bootstrap method, our method uses only the original data set.

There are two major computational costs for our method. One is the one-time calculation of a parameter covariance matrix and its inverse. The other is that our simulation requires repeated one-dimensional root finding.

This paper is organized as follows. In Section 2 we describe our basic sampling paradigm that converts coverage probabilities into the sampling probabilities of a *confidence distribution.* In Section 3 we create two ways to sample from the confidence distribution, independence sampling and boundary sampling, and discuss them in terms of the coverage probability and computation.

In Section 4 we evaluate the performance of our method by using four inference functions in three different examples. We note that the development of

our visualization method was motivated by a difficult problem of mixture model analysis that is described in the third example. In Section 5 we carry out a small simulation study to exhibit the coverage properties of our sampling-based methodology for a nonlinear function of the parameters.

2. Basic Idea of Sampling Based Visualization

In this section we briefly review the properties of the four inference functions: the likelihood, score statistic, the quadratic inference function and the empirical likelihood. We also review how one constructs asymptotic (profile) confidence regions based on a given inference function, and then explain a basic idea of our sampling methodology.

2.1. Four inference functions and asymptotic confidence regions

The most used inference function is the log-likelihood of $\boldsymbol{\theta}$: for given each independent data $\mathbf{Y}_i$ with a density in the parametric family $\{p_i(\mathbf{y} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Omega}_{\boldsymbol{\theta}} \subset R^p\}$,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_i(\mathbf{y} \mid \boldsymbol{\theta}). \tag{2.1}$$

Under regularity $-2\ell(\boldsymbol{\theta})$ then satisfies our description of an inference function.

One can employ the score test statistic (Rao (1948)), $RS(\boldsymbol{\theta})$, as an inference function because it is asymptotically equivalent to the likelihood ratio statistic :

$$RS(\boldsymbol{\theta}) = U(\boldsymbol{\theta})'I(\boldsymbol{\theta})^{-1}U(\boldsymbol{\theta}), \tag{2.2}$$

where $U(\boldsymbol{\theta})$ is the vector of first-order derivatives of $\ell(\boldsymbol{\theta})$ and $I(\boldsymbol{\theta})$ is the Fisher information. The use of the observed empirical information in Rao's score test can help repair problems caused by model misspecification(Boos (1992)).

The likelihood ratio test statistic, the score test statistic, and the Wald test statistic are asymptotically pivotal and have the same limiting chi-squared distribution for a wide class of parametric models : for given $\hat{\boldsymbol{\theta}}$ maximizing $\ell(\boldsymbol{\theta})$,

$$LR(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\right) \xrightarrow{D} \chi_p^2, \tag{2.3}$$

$$RS(\boldsymbol{\theta}_0) = U(\boldsymbol{\theta}_0)'I(\boldsymbol{\theta}_0)^{-1}U(\boldsymbol{\theta}_0) \xrightarrow{D} \chi_p^2, \tag{2.4}$$

$$WD(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'I(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} \chi_p^2, \tag{2.5}$$

where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$ and χ_p^2 is the chi-squared distribution with p degrees of freedom. For multidimensional parameters, it is easy to visualize the corresponding confidence sets only for the Wald statistic. However, the other two

statistics have some theoretical and conceptual advantages, including the fact that they are invariant to reparametrization. As will be seen, our visualization method is based on taking the elliptical confidence sets generated by the Wald method and adjusting their shape for the change in the inference function.

Alternatively, if one specifies a set of mean-zero estimating functions defining a semiparametric model, for statistical inference one might use the quadratic inference function (Qu (1998); Lindsay and Qu (2003)), referred to as the QIF, and the empirical likelihood (Owen (1988); Qin and Lawless (1994)) of $\boldsymbol{\theta}$: for given a q-dimensional vector of estimating functions $\mathbf{b}(\mathbf{Y}_i, \boldsymbol{\theta})$ with $p \leq q$, one can construct the QIF and the empirical likelihood of $\boldsymbol{\theta}$,

$$\mathbf{Q}(\boldsymbol{\theta}) = n\bar{\mathbf{b}}_{\boldsymbol{\theta}}' \hat{\mathbf{C}}_{\boldsymbol{\theta}}^{-1} \bar{\mathbf{b}}_{\boldsymbol{\theta}}, \tag{2.6}$$

$$L_{EL}(\boldsymbol{\theta}) = \sup \left\{ L_{EL}(\mathbf{p}) = \prod_{i=1}^{n} \mathbf{p}_i : \sum_{i=1}^{n} \mathbf{p}_i = 1, \sum_{i=1}^{n} \mathbf{p}_i \mathbf{b}(\mathbf{Y}_i, \boldsymbol{\theta}) = \mathbf{0} \right\}. \tag{2.7}$$

Here $\bar{\mathbf{b}}_{\boldsymbol{\theta}} = n^{-1} \sum_{i=1}^{n} \mathbf{b}(\mathbf{Y}_i, \boldsymbol{\theta})$ and $\hat{\mathbf{C}}_{\boldsymbol{\theta}}$ is a suitable estimator of the covariance matrix $Var_{\tau}(\bar{\mathbf{b}}_{\boldsymbol{\theta}})$. Note the τ in Var indicates that the true distribution τ is used to evaluate the variance. The QIF at (2.6) and the empirical likelihood at (2.7) can be used as inference functions in both parametric and semiparametric models.

In fact, the QIF and the empirical likelihood are invariant to reparametrization, and also asymptotically pivotal under the null hypothesis with a limiting chi-squared distribution (Owen (1988); Qin and Lawless (1994); Lindsay and Qu (2003)): for $\ell_{EL}(\boldsymbol{\theta}) = \log L_{EL}(\boldsymbol{\theta})$,

$$\mathbf{Q}(\boldsymbol{\theta}_0) - \mathbf{Q}(\hat{\boldsymbol{\theta}}) \xrightarrow{D} \chi_p^2, \tag{2.8}$$

$$2 \left(\ell_{EL}(\hat{\boldsymbol{\theta}}) - \ell_{EL}(\boldsymbol{\theta}_0) \right) \xrightarrow{D} \chi_p^2. \tag{2.9}$$

To apply our method, we also need a Wald statistic corresponding to the tests in (2.8) and (2.9). The estimators $\hat{\boldsymbol{\theta}}$ that optimize $\mathbf{Q}(\boldsymbol{\theta})$ and $L_{EL}(\boldsymbol{\theta})$ have the asymptotic property

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) \xrightarrow{D} N_p \left(\mathbf{0}, (D'(\boldsymbol{\theta}_0) \Sigma_{\boldsymbol{\theta}_0}^{-1} D(\boldsymbol{\theta}_0))^{-1} \right), \tag{2.10}$$

where $D(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} [\partial \mathbf{b}(X; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}]$, $\Sigma_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}} [\mathbf{b}(X; \boldsymbol{\theta}) \mathbf{b}(X; \boldsymbol{\theta})']$, and $N_p(a, A)$ is the p-variate normal distribution with mean vector a and covariance matrix A.

When the inference function $H(\boldsymbol{\theta}; \mathbf{y})$ is $-2\ell(\boldsymbol{\theta})$, $RS(\boldsymbol{\theta})$, $\mathbf{Q}(\boldsymbol{\theta})$, and $-2\ell_{EL}(\boldsymbol{\theta})$, the $100(1 - \alpha)\%$ confidence sets generated by $H(\boldsymbol{\theta}; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$ have the form

$$C_{H, 1-\alpha} \equiv \left\{ \boldsymbol{\theta} \mid H(\boldsymbol{\theta}; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) \leq c_{p, 1-\alpha} \right\}, \tag{2.11}$$

where $c_{p,1-\alpha}$ is the α upper quantile of the asymptotic distribution of $H(\boldsymbol{\theta}; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$.

2.2. Asymptotic confidence distribution

Our next goal is to create an *asymptotic confidence distribution* P_{CD} on $\boldsymbol{\theta}$ for given inference function $H(\boldsymbol{\theta}; \mathbf{y})$, by which we mean that the parameter space probability $P_{CD}(\boldsymbol{\theta} \in C_{H,1-\alpha} \mid \mathbf{y})$ matches the asymptotic sample space probability $P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in C_{H,1-\alpha}) = 1 - \alpha$ for every α. To consider the Wald confidence set, for example, a simple construction for the confidence distribution is to let the parameter have the distribution $\tilde{\boldsymbol{\theta}}_W \sim N_p(\hat{\boldsymbol{\theta}}, I(\hat{\boldsymbol{\theta}})^{-1})$. We call this the Wald confidence distribution. An additional goal for our construction is to make the confidence distribution easy to sample from, just like the Wald confidence distribution.

Given P_{CD}, one can generate a sample $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_B\}$ of $\boldsymbol{\theta}$ from it, and use empirical plots of these samples to visualize features of the confidence sets. One automatically knows the theoretical proportion of points lying in the 90% confidence set is 90%, and they are recognizable by having $H(\boldsymbol{\theta}; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$ larger than the asymptotic critical value. In addition, if one is interested in picturing a profile confidence set for $g(\boldsymbol{\theta})$, all one needs to do is to plot $g(\tilde{\boldsymbol{\theta}}_1), \ldots, g(\tilde{\boldsymbol{\theta}}_B)$ in a suitable way, using the adjusted critical value that corresponds to the degrees of freedom of the parameter $\nu = g(\boldsymbol{\theta})$.

While our method creates a parameter sample for the whole confidence set, we think it has greatest application when one has a considerable number of parameters of interest in one's investigation, and would like the confidence sets generated for each of these parameters while treating the remaining parameters as nuisance parameters.

The idea of creating a distribution on the parameter space that generates confidence sets is also the underpinning of Fisher's concept of fiducial inference (Fisher (1930)). The main goal of fiducial inference there was to find a distribution on the parameters(called a fiducial distribution of the parameters) which contained, ideally, all the information in the observed data about the parameters, so that one could have objective probability statements about the parameters on the basis of observed data, without any prior distribution.

We are proceeding rather differently. Instead of starting with a fiducial distribution, we start with a chosen inference function and its asymptotic confidence sets, and then create an asymptotic confidence distribution that will recreate those sets. In particular, for multidimensional parameters, there is an infinitude of such parameter space distributions. We choose one based on convenience of sampling and the fact that we want the boundaries of profile regions to be as clear as possible.

In passing we note that there have been several recent papers related to the idea of Fisher's fiducial inference. Their emphasis has been on the construction of approximate confidence sets called generalized confidence intervals, based on generalized pivotal quantities, or fiducial generalized confidence intervals, based on fiducial generalized pivotal quantities (Weerahandi (1993); Hannig, Lyer, and Patterson (2006)).

3. Implementation

In this section we create a confidence distribution on the parameter space, then describe two sampling methods. We also discuss the coverage properties and computation.

3.1. The confidence distribution based on Wald statistic rays

The algorithm for generating a single observation from this confidence distribution is as follows.

Algorithm for the Wald ray-based confidence distribution

Suppose one has already found the parameter estimator $\hat{\boldsymbol{\theta}}$ minimizing $H(\boldsymbol{\theta}; \mathbf{y})$ and the covariance estimator $V_{\hat{\boldsymbol{\theta}}}$. Suppose one also defines the Wald statistic ray generated by a vector $\mathbf{z} \in R^p$ to be $\boldsymbol{\theta}(\epsilon) = \hat{\boldsymbol{\theta}} + \epsilon V_{\hat{\boldsymbol{\theta}}}^{1/2} \mathbf{z}$ where $\epsilon \in R^+$.

Step 1. Generate a vector $\mathbf{z}$ from $N(0, \mathbf{I}_p)$ and find $\alpha(\mathbf{z}) = P(\chi_p^2 \geq \mathbf{z}'\mathbf{z})$.

Step 2. Find $\tilde{\epsilon} = \tilde{\epsilon}(\mathbf{z})$ such that $H(\boldsymbol{\theta}(\tilde{\epsilon}); \mathbf{y})$ - $H(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \mathbf{z}'\mathbf{z}$, where $\boldsymbol{\theta}(\epsilon)$ is the Wald statistic ray defined above. Let $\tilde{\boldsymbol{\theta}}(\mathbf{z}) = \boldsymbol{\theta}(\tilde{\epsilon})$ be the simulated value generated by $\mathbf{z}$.

Note that $\tilde{\boldsymbol{\theta}}(\mathbf{z})$ in Step 2 is a point from $\hat{\boldsymbol{\theta}}$ in direction $\mathbf{z}$ that belongs to the $100(1 - \alpha(\mathbf{z}))\%$ confidence set generated by $H(\boldsymbol{\theta}; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$ in the whole parameter space (if we were to allow ϵ to be negative, there would typically be two solutions $\tilde{\epsilon}$, one positive and one negative, but for now we consider only the positive one).

We use the expression '*independent sampling*' to refer to repeated sampling from the algorithm for the Wald ray-based confidence distribution. To ensure the validity of this algorithm, we assume that $H(\boldsymbol{\theta}(\epsilon); \mathbf{y})$ is continuous and strictly increasing in ϵ. In this case $\tilde{\epsilon}$ is uniquely determined for every $\mathbf{z}$. This assumption implies that the confidence sets are *star-shaped*. This means that every point in the confidence set can be reached by a ray from $\hat{\boldsymbol{\theta}}$ that stays entirely in the confidence set. We examine this assumption later. Note that for the Wald statistic $\tilde{\epsilon} = 1$, so we are in effect assessing the deviation of the inference function region from the Wald region along the ray. The choice of $V_{\boldsymbol{\theta}}$ estimator in the Wald ray is discussed in Section 3.4.

Note that the independent sampling does not recreate the Wald confidence region, nor even depend on it very much. The main purpose of the Wald-type construction is to ensure that we are sampling the boundaries of our targeted confidence regions in a manner consistent with the limiting distribution. We note that there are many possible ways to construct a confidence distribution. The method described here is designed so that the generation of sample values is quite straightforward, requiring only the solution of a simple one-dimensional optimization in Step 2.

The next theorem says that the distribution generated by the algorithm matches the asymptotic coverage probabilities for the confidence sets, and so is a true confidence distribution.

Theorem 3.1. *Given the observed data* $\mathbf{y}$ *and the parameter* $\boldsymbol{\theta} \in R^p$ *in an assumed statistical model, suppose that* $H(\boldsymbol{\theta}; \mathbf{y})$ *is an inference function with the four properties as in Section* 1. *Assume further that* $H(\boldsymbol{\theta}(\epsilon); \mathbf{y})$ *is monotone increasing in* ϵ *for all* $\mathbf{z}$. *Let* $\tilde{\boldsymbol{\theta}}(\mathbf{z})$ *be generated from the confidence distribution. Then the following hold.*

(i) $P[\tilde{\boldsymbol{\theta}}(\mathbf{Z}) \in C_H(\mathbf{y})] = 1 - \alpha$, *where* $\mathbf{y}$ *is treated as fixed and* $\mathbf{Z}$ *is treated as random.*

(ii) *If* $\nu = g(\boldsymbol{\theta}) \in R^r$ *is a linear function of* $\boldsymbol{\theta}$, *then* $P[\tilde{\nu}(\mathbf{Z}) \in PC_H(\mathbf{y})] = 1 - \alpha$ *for* $\tilde{\boldsymbol{\theta}}(\mathbf{z}) \in \{\boldsymbol{\theta} \mid H(\boldsymbol{\theta}; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) \leq \chi^2_{r,1-\alpha}\}$ *and* $\tilde{\nu}(\mathbf{z}) = g(\tilde{\boldsymbol{\theta}}(\mathbf{z}))$, *where* $\chi^2_{r,1-\alpha}$ *is the* α *upper quantile of* χ^2_r.

Proof. (i) We can find the probability that $\tilde{\boldsymbol{\theta}}(\mathbf{z})$ lies in the set $C_H(\mathbf{y}) = \{\boldsymbol{\theta} : H(\boldsymbol{\theta}; \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) \leq \chi^2_{p,1-\alpha}\}$ by calculating this probability conditional on the direction $\mathbf{z}/\|\mathbf{z}\|$. The independence of this direction from $\|\mathbf{z}\|$ implies that conditionally, given this direction, $\|\mathbf{z}\|^2$ is chi-squared distributed with p degrees of freedom. We can then calculate $P\left[\tilde{\boldsymbol{\theta}}(\mathbf{z}) \in C_H(\mathbf{y}) \big| \mathbf{z}/\|\mathbf{z}\|\right] = P\left[\chi^2_p \leq \chi^2_{p,1-\alpha}\right] = 1 - \alpha$. Note that this calculation does not depend on the direction; also recall that $\tilde{\boldsymbol{\theta}}(\mathbf{z})$ was constructed so that $H(\boldsymbol{\theta}(\tilde{\epsilon}); \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \mathbf{z}'\mathbf{z}$.

ii) Let $g(\boldsymbol{\theta}) = A'\boldsymbol{\theta}$ be an $r \times 1$ linear combination of $\boldsymbol{\theta}$, where A is a $p \times r$ constant matrix, $r \leq p$, and $\partial g(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = A$. If we want to find the inferred distribution for $g(\boldsymbol{\theta})$ under the Wald confidence distribution, we can find it from $\tilde{\boldsymbol{\theta}}_W = \hat{\boldsymbol{\theta}} + V_{\hat{\boldsymbol{\theta}}}^{1/2} \mathbf{z}$:

$$\tilde{g}_W(\boldsymbol{\theta}) \equiv g(\tilde{\boldsymbol{\theta}}_W) = A'\tilde{\boldsymbol{\theta}}_W = g(\hat{\boldsymbol{\theta}}) + A'V_{\hat{\boldsymbol{\theta}}}^{1/2}\mathbf{z} = g(\hat{\boldsymbol{\theta}}) + \left(A'V_{\hat{\boldsymbol{\theta}}}A\right)^{1/2} \mathbf{z}^{\star},$$

where $A'V_{\hat{\boldsymbol{\theta}}}A$ is the asymptotic covariance of $g(\hat{\boldsymbol{\theta}})$ and $\mathbf{z}^{\star} \sim N(0, I_r)$. Hence the Wald confidence distribution gives the correct profile coverage probabilities to

all linear combinations of $\boldsymbol{\theta}$. It follows that for other confidence distributions, adjusting the length of the rays is also an asymptotically size-correct method for all linear combinations of $\boldsymbol{\theta}$ (this follows because the rays are themselves linear combinations of the parameters).

The coverage properties of the marginal confidence distribution of a general nonlinear function of $\boldsymbol{\theta}$, $g(\boldsymbol{\theta})$, are more complex. However, we do know through the delta method that appropriate coverage properties for smooth nonlinear functions $g(\boldsymbol{\theta})$ can be generated by linearization. Since linear functions of $\boldsymbol{\theta}$ have appropriate coverage by the Theorem 3.1, this property holds approximately for smooth nonlinear functions. We might compare this situation with sampling from a Bayesian posterior; in that case, a marginal sample of $g(\boldsymbol{\theta})$-values is indeed a sample from the correct marginal posterior, and so no appeal to asymptotics is needed.

3.2. A boundary sampling method

One advantage of confidence sampling over posterior sampling is that we know the critical values corresponding to described confidence set boundaries, both for the full confidence set and for the parameters of interest. Moreover, there is a nice geometric relationship between the full confidence set and the profile confidence sets: the parameter value $g(\boldsymbol{\theta}) = \nu$ is in the profile confidence set at a given critical value $c_{1-\alpha}$ if and only if there exists one or more $\boldsymbol{\theta}$ in the full confidence set based on $c_{1-\alpha}$ that satisfy $g(\boldsymbol{\theta}) = \nu$. Hence one can use one simulation for the full parameter confidence set generated by the appropriate $c_{1-\alpha}$ in order to determine the confidence set for any and all $g(\boldsymbol{\theta})$.

It is important to our method that we sample $\boldsymbol{\theta}$ with enough dispersion and density that we can recover the shape of the profile function near the critical values. The density of sampling is sufficient if there are enough sampled $\boldsymbol{\theta}$ points near the curve in $\boldsymbol{\theta}$-space defined by the profile estimators $\hat{\boldsymbol{\theta}}_{\nu}^{*}$. This ensures that a plot of $(\nu(\tilde{\boldsymbol{\theta}}), H(\tilde{\boldsymbol{\theta}}; \mathbf{y}))$ reveals the shape of the profile function $H(\hat{\boldsymbol{\theta}}_{\nu}^{*}; \mathbf{y})$ at the right critical values (see the example in Section 4.2).

We recommend the following modification of independent sampling from the confidence distribution to sharpen the boundaries of the confidence sets. We independently sample $\mathbf{z}$-values as before, but now for each ray $\mathbf{z}/||\mathbf{z}||$ we calculate in a deterministic fashion the $\boldsymbol{\theta}$-values found along the ray $\hat{\boldsymbol{\theta}} + \epsilon V_{\hat{\boldsymbol{\theta}}}^{1/2} \mathbf{z}$ that correspond to the confidence boundary points at a fixed grid of interesting α's. That is, for each generated $\mathbf{z}$, obtain $\boldsymbol{\theta}(\tilde{\epsilon}_1)$ for α_1, $\boldsymbol{\theta}(\tilde{\epsilon}_2)$ for α_2, etc. Moreover, we now calculate the boundary points along the opposite direction of the ray, so that we obtain two $\boldsymbol{\theta}$'s for each α.

In this method one is sampling the directions $\mathbf{z}/||\mathbf{z}||$ uniformly on the sphere, but the lengths $||\mathbf{z}||$ are no longer random. We call this methodology '*boundary*
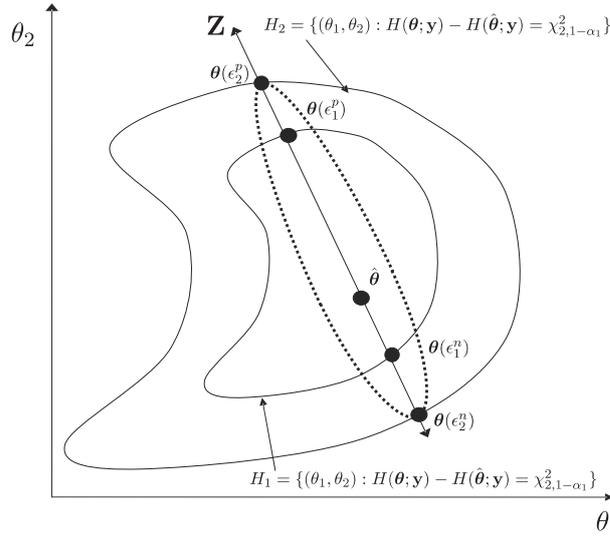
Figure 1. Geometric concept of the boundary sampling from the confidence distribution.

sampling' from the confidence distribution. In this case, the collection of $\boldsymbol{\theta}$-points corresponding to a single α each lie on the boundary of a targeted confidence set. They correspond to points uniformly chosen on the sphere $||\mathbf{z}|| = 1$. However, the fraction of points found above a particular critical value will no longer correspond to a confidence probability.

Figure 1 graphically illustrates how the boundary sampling method works for obtaining simulated values of a two dimensional $\boldsymbol{\theta} = (\theta_1, \theta_2)$ where there are two target levels of α, α_1 and α_2. The confidence sets are denoted by H_1 and H_2, with $\alpha_1 > \alpha_2$. Given a generated $\mathbf{z} \sim N(0, \mathbf{I}_2)$ and the $\hat{\boldsymbol{\theta}}$ that minimizes $H(\boldsymbol{\theta}; \mathbf{y})$, one defines the Wald statistic ray $\boldsymbol{\theta}(\epsilon) = \hat{\boldsymbol{\theta}} + \epsilon V_{\hat{\boldsymbol{\theta}}}^{1/2} \mathbf{z}$. Let $\chi_{2,1-\alpha_1}^2$ denote the upper α_1 quantile of χ_2^2. Given this quantile, the first goal is to find the smallest positive ϵ such that the ray just touches the boundary of the targeted confidence set. This positive ϵ, ϵ_1^p, is such that $\boldsymbol{\theta}(\epsilon_1^p)$ satisfies $H(\boldsymbol{\theta}(\epsilon_1^p); \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \chi_{2,1-\alpha_1}^2$. Next, one repeats this operation, but now for negative values of ϵ, ϵ_1^n such that $\boldsymbol{\theta}(\epsilon_1^n)$ is on the opposite boundary from $\boldsymbol{\theta}(\epsilon_1^p)$. Note that ϵ_1^n is generally not equal to -ϵ_1^p. Repeat this procedure for α_2 to get two boundary points of H_2.

Algorithm for boundary confidence sampling

Suppose one has found the estimator $\hat{\boldsymbol{\theta}}$ minimizing $H(\boldsymbol{\theta}; \mathbf{y})$ and the covariance estimator $V_{\hat{\boldsymbol{\theta}}}$. Suppose one also defines the Wald statistic ray generated by a vector $\mathbf{z} \in R^p$ to be $\boldsymbol{\theta}(\epsilon) = \hat{\boldsymbol{\theta}} + \epsilon V_{\hat{\boldsymbol{\theta}}}^{1/2} \mathbf{z}$ where $\epsilon \in R$.

Step 1. Generate $\mathbf{z} \sim N(0, \mathbf{I}_p)$ and find the α upper quantile $\chi_{p,1-\alpha}^2$ of χ_p^2.

Step 2. Determine ϵ satisfying $H(\boldsymbol{\theta}(\epsilon); \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \chi^2_{p,1-\alpha}$. Denote the computed ϵ as $\tilde{\epsilon} = \tilde{\epsilon}(\alpha, \mathbf{z})$. Let $\tilde{\boldsymbol{\theta}}(\alpha, \mathbf{z}) = \boldsymbol{\theta}(\tilde{\epsilon})$ be the simulated value generated by $(\alpha, \mathbf{z})$. Compute $\tilde{\boldsymbol{\theta}}(\alpha, \mathbf{z})$ for both the positive and negative ϵ solutions.

This algorithm requires the determination of the one-dimensional ϵ many times for each α.

3.3. The Star-shaped assumption

In a star-shaped confidence set, the inference function is monotonically increasing along every ray from $\hat{\boldsymbol{\theta}}$, so that one finds exactly one solution to $H(\boldsymbol{\theta}(\epsilon)) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \chi^2_{p,1-\alpha}$ along the positive and negative rays, unless the inference function never increases enough in that direction, in which case there is no solution and an infinite distance to the confidence boundary. However, if this assumption is violated, the number of solutions can exceed two. In order to study these failures of the star-shaped assumption when using the boundary sampling algorithm we search for extraneous solutions, and then categorize $(\alpha, \mathbf{z})$ into two groups, *acceptable* and *unacceptable* as follows.

When there is more than one solution of ϵ on either side of $\epsilon = 0$ for a given $(\alpha, \mathbf{z})$, that $(\alpha, \mathbf{z})$ is called *unacceptable*, as this indicates a violation of the assumption of star-shaped confidence set. Thus a search for unacceptable $(\alpha, \mathbf{z})$ is a check of the star-shaped assumption at various levels of α, and along various rays.

On the other hand, the following three types of $(\alpha, \mathbf{z})$ are *acceptable*: $(\alpha, \mathbf{z})$ for which there is only one solution on each side of $\epsilon = 0$, so that one has the two-sided interval for ϵ; where there is only one solution of ϵ on only one side of $\epsilon = 0$ for a given $(\alpha, \mathbf{z})$, in which case we say that the solution on the opposite side is $+/-\infty$; where there is no solution of ϵ for a given $(\alpha, \mathbf{z})$, in which case we say the two endpoints are $-\infty$ and $+\infty$. Note that one could check the star-shaped assumption during independent sampling, but it is less efficient to do so there.

3.4. Computational comments

We now give some comments on the implementation of the algorithm starting with the dispersion matrix $V_{\boldsymbol{\theta}}$ in the Wald statistic ray. The form of $V_{\boldsymbol{\theta}}$ depends on the inference function. For the likelihood and score statistic the Fisher information or its asymptotic equivalents can be used, and for the QIF and empirical likelihood, $D^T(\boldsymbol{\theta})\Sigma^{-1}_{\boldsymbol{\theta}}D(\boldsymbol{\theta})$ at (2.10) is available. Next, if one has numerical challenges due to constrained parameter sets or non-star shaped confidence sets,

suitable reparametrization might need to be considered in the construction of the Wald statistic rays.

Our methodology requires the specification of a critical value. If one is concerned that the critical value based on asymptotic theory might not provide suitable coverage, there are various adjustments to this value, such as Bartlett type corrections (Barndorff-Nielsen and Cox (1994)) or the bootstrap method at an estimated parameter value (Efron and Tibshirani (1993); Davison and Hinkley (1997); Hall (1992)). See also Chen, Mulayath and Bovas (2008) for a new correction method for empirical likelihood intervals.

Finally, we consider methods for determining the ϵ solutions in the Wald statistic ray. Since ϵ is one-dimensional, only simple algorithms are needed for computation. We used the MATLAB function "fzero" in the examples.

We suggest two ways to generate starting values for ϵ. First, available only for the boundary sampling method, use the fact that for the same $\mathbf{z}$, the intervals of ϵ become wider as the confidence levels increase: given $\mathbf{z}$ and the smallest confidence level $(1 - \alpha_{\max})$, find positive and negative ϵ's satisfying $H(\boldsymbol{\theta}(\epsilon); \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \chi^2_{p,1-\alpha_{\max}}$, denoted as $(\epsilon^n_{(1-\alpha_{\max},\mathbf{z})}, \epsilon^p_{(1-\alpha_{\max},\mathbf{z})})$; then use these as starting values of ϵ for the same $\mathbf{z}$ and the next larger confidence level.

A second algorithm (and the one used in our examples) employs the interval of ϵ associated with $\alpha_{\min}$, the smallest α grid point. First, find the most extreme values of ϵ, say ϵ^n negative and ϵ^p positive, that solve $H(\boldsymbol{\theta}(\epsilon); \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \chi^2_{p,1-\alpha_{\min}}$. Then for $\epsilon \in (\epsilon^n, \epsilon^p)$, evaluate $H(\boldsymbol{\theta}(\epsilon); \mathbf{y}) - H(\hat{\boldsymbol{\theta}}; \mathbf{y})$ on a grid of ϵ-values. For other α select the grid values closest to the target $\chi^2_{p,1-\alpha}$ value as the initial value for ϵ(one could use interpolation instead). This algorithm helps find unacceptable $(\alpha, \mathbf{z})$ values better than the first.

3.5. Formal random boundary of the confidence set

To this point, we have described how to create a sample of points from the confidence sets, and thereby to create plots that represent their basic shape. However, a more formal definition of the boundaries of the sets would be needed if one wanted to use them to test hypotheses, or to assess their coverage properties. This is easily done for a one-dimensional parameter. Assuming the confidence interval at the given critical value is starshaped, one would use the leftmost and rightmost sampled values as the two endpoints of the interval. Of course, that interval is somewhat shorter than the true confidence interval, so there is some chance of undercoverage relative to the nominal level.

The issue of sparsity of sampling grows as one goes into two dimensions. Although one could define the confidence set to be the convex hull of the sampled values, that would not allow a banana shaped confidence set that would be allowed under the star-shaped assumption. One possibility is to construct

a kernel density estimator $K(\boldsymbol{\theta})$ from the simulated samples, and to let $c = \min\{K(\boldsymbol{\theta}_s), s = 1, \ldots, S\}$, the smallest density value of any sampled $\boldsymbol{\theta}$. A simple construction of the confidence set C would be $\{\boldsymbol{\theta} : K(\boldsymbol{\theta}) \geq c\}$. This set would include every sampled $\boldsymbol{\theta}$, and all $\boldsymbol{\theta}$ in regions of reasonable density. It would also be easy numerically. A more complex definition would be to use $C^\star$, the smallest star-shaped set containing C. A point would be in this set if it lay on a ray from $\hat{\boldsymbol{\theta}}$ to a point in the boundary of C (that is, one with $K(\boldsymbol{\theta}) = c$).

4. Data Analysis

In this section we provide three examples to illustrate the performance of the proposed confidence distribution. We also compare the samples from the boundary sampling method with those from independent sampling method. Note that an uncentered empirical variance estimator, $\hat{\mathbf{C}}_{\boldsymbol{\theta}} = n^{-1} \sum_{i=1}^{n} \mathbf{b}(\mathbf{Y}_i, \boldsymbol{\theta})\mathbf{b}(\mathbf{Y}_i, \boldsymbol{\theta})'$, is used in the QIF (we denote the QIF based on $\hat{\mathbf{C}}_{\boldsymbol{\theta}}$ by $\mathbf{Q}_u(\boldsymbol{\theta})$). The confidence distribution values $\tilde{\boldsymbol{\theta}}$ forming the likelihood region, the score region, the QIF_u region, and the empirical likelihood region are denoted by $\tilde{\boldsymbol{\theta}}_L$, $\tilde{\boldsymbol{\theta}}_S$, $\tilde{\boldsymbol{\theta}}_{Q_u}$, and $\tilde{\boldsymbol{\theta}}_{EL}$, respectively.

We used parametric models in all three examples. In the QIF and the empirical likelihood we defined the estimating functions through the parametric log likelihood function. Since the number of estimating functions and the number of parameters are equal, the estimating functions defined above are optimal (Godambe and Heyde (1987)) and, therefore, the maximum likelihood estimator is an estimator optimizing all candidates for $H(\boldsymbol{\theta}; \mathbf{y})$ considered in three examples. In this case QIF can be thought of as a score statistic that uses an empirical covariance estimator.

4.1. Multiple linear regression model

The first data, introduced in Moore and McCabe (1989), are the taste of matured cheese and concentrations of several chemicals in 30 samples of mature cheddar cheese. The goal is to investigate the relationships between the taste of matured cheese (response variable, Taste) and three chemicals deciding the taste of cheese: acetic acid, hydrogen sulfide, and lactic acid (predictors, Acetic, H2S, and Lactic, respectively). Note that the first two predictors are log transformed.

Since there was a linear relationship between a response variable and the three predictors, we fitted a multiple linear regression model with independent normal error terms to the data. Based on residual plots and normal probability plot of the residuals, we took the regression of Taste on H2S and Lactic as the best regression model, Taste $= \beta_0 + \beta_1$ H2S $+ \beta_2$ Lactic $+ \tau$, where $\tau \sim N(0, \sigma^2)$. In this model we were interested in the likelihood surface of β_1, β_2, and σ^2. Thus β_1, β_2, and σ^2 were the parameters of interest, and β_0 was a

Table 1. Percentage of acceptable and unacceptable $(\alpha, \mathbf{z})$ among 6,300 $(\alpha, \mathbf{z})$'s.

Interval of ϵ	Acceptable $(\alpha, \mathbf{z})$			Unacceptable $(\alpha, \mathbf{z})$
	Two sided	Doubly infinite	Half infinite	
Likelihood region				
All α's	100	0	0	0
Score region				
All α's	100	0	0	0
QIF_u region				
All α's > 0.1	100	0	0	0
$\alpha = 0.1$	99	0	0	1
$\alpha = 0.05$	96.714	0.286	1	2
Empirical likelihood region				
All α's > 0.1	100	0	0	0
$\alpha = 0.1$	99.286	0	0.714	0
$\alpha = 0.05$	98.857	0	1.143	0

nuisance parameter. The maximum likelihood estimate for $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma^2)'$ was $\hat{\boldsymbol{\theta}} = (-27.5918, 3.94627, 19.8872, 88.9655)'$.

For the application of our independent sampling visualization to the data, we generated 700 $\mathbf{z}$'s from $N(0, \mathbf{I}_4)$. For the boundary sampling method we used the same $\mathbf{z}$'s and nine α values $(1-\alpha=[0.05, 0.1, 0.1534, 0.3, 0.5, 0.7, 0.8002, 0.9, 0.95])$. In all there were 700 $\alpha(\mathbf{z})$'s in the independent sample and 6,300 $(\alpha, \mathbf{z})$'s in the boundary sample. Note that the log transformation for σ^2 was used to eliminate the non-negativity constraint, and the expected Fisher information was used in the Wald statistic ray and the score test statistic.

We use this example to compare the geometric features of various confidence sets. We are interested in two features: the degree to which the star shaped assumption holds, and how the confidence set shapes vary among inference functions.

We first examine the star-shaped assumption. Table 1 gives the percentage of acceptable and unacceptable $(\alpha, \mathbf{z})$ in four confidence regions for boundary sampling method. Table 1 shows that all 6,300 $(\alpha, \mathbf{z})$'s in the boundary sample were acceptable for the likelihood, score, and empirical likelihood region, as there was either a two sided interval or a half-infinite interval of $\tilde{\epsilon}$ for every $(\alpha, \mathbf{z})$. However, the QIF_u region did have a small percentage of unacceptable $\mathbf{z}$'s for large $(1-\alpha)$s.

We can illustrate the various types of solutions by selecting several $(\alpha, \mathbf{z})$'s in the boundary sample and plotting QIF_u as a function of ϵ along the rays generated by values of ϵ. Figure 2 shows the plot of $\mathrm{QIF}_u - c_{4,(1-\alpha,\mathbf{z})}$ vs. ϵ for four $(\alpha, \mathbf{z})$'s. The $(\alpha, \mathbf{z})$'s in the first, second and third plots were acceptable,
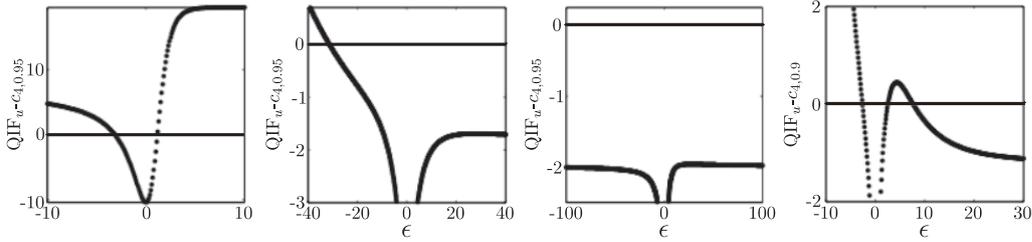
Figure 2. QIF_u-$c_{4,(1-\alpha,\mathbf{z})}$ vs. ϵ for four (α, z)'s where the black line represents a reference line with $\mathrm{QIF}_u - c_{4,(1-\alpha,\mathbf{z})} = 0$ and $\mathbf{z}_i$ means $\mathbf{z}$ with the ith smallest norm - $(0.05, \mathbf{z}_{498})$ (1st), $(0.05, \mathbf{z}_{194})$ (2nd), $(0.05, \mathbf{z}_{357})$ (3rd) and $(0.1, \mathbf{z}_{668})$ (4th).
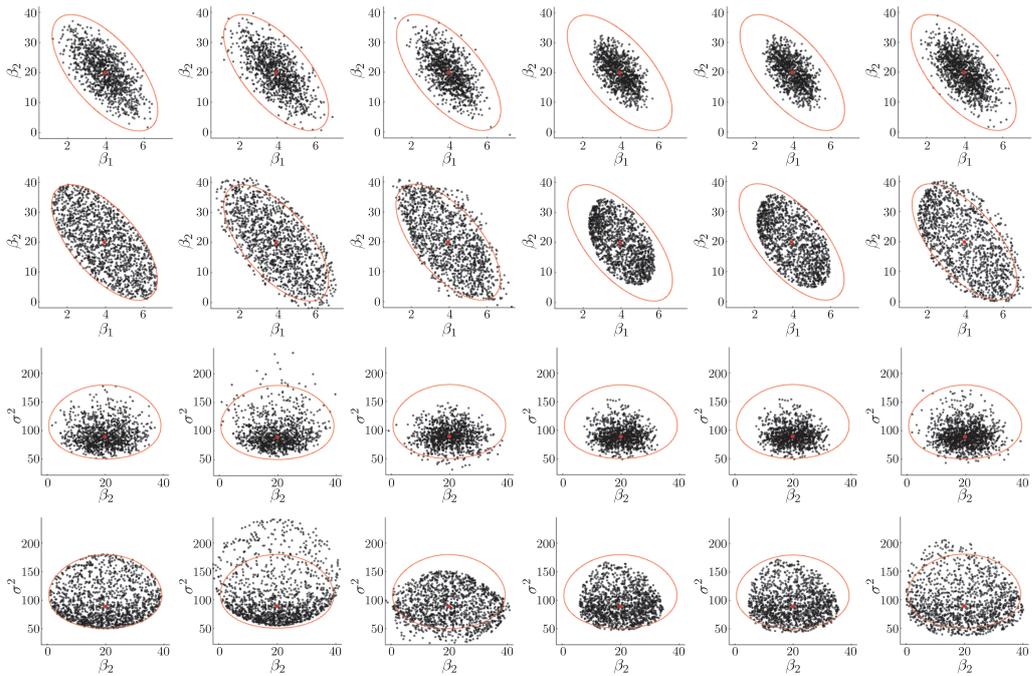

Figure 3. 80% joint likelihood confidence regions (red) overlaid on joint simulated confidence regions(black dots) using independent samples(Rows 1 and 3) and boundary samples(Rows 2 and 4) for (β_1, β_2) and (β_2, σ^2). Note that the first and second rows are the joint confidence regions for (β_1, β_2), and third and fourth rows are the joint confidence regions for (β_2, σ^2). Columns $1-6$ give profile simulated regions of $\tilde{\boldsymbol{\theta}}_L$, $\tilde{\boldsymbol{\theta}}_S$, $\tilde{\boldsymbol{\theta}}_{Q_u}$, $\tilde{\boldsymbol{\theta}}_{EL}$, $\tilde{\boldsymbol{\theta}}_{ADEL}$ and $\tilde{\boldsymbol{\theta}}_{ELBOOT}$, respectively.

as the number of solutions of ϵ were two(two-sided interval), one(half-infinite interval) and zero(doubly infinite interval), respectively. However, $(\alpha, \mathbf{z})$ in the fourth plot of Figure 2 was unacceptable because there were two solutions on the

right side of $\epsilon = 0$.

Next we examine the shapes of the various joint confidence sets with 80% confidence level in the whole parameter space (this corresponds to a confidence level of 95% for two dimensional profiles). Figure 3 contains scatter plots of the samples generated by our methods for all four inference functions (Columns 5 and 6 will be discussed later). Rows 1 and 2 are marginal plots for (β_1, β_2), with Row 1 being independent sampling and Row 2 boundary sampling. Rows 3 and 4 are marginal plots for (β_2, σ^2), with Row 3 being independent sampling and Row 4 boundary sampling. For the boundary sampling all sample points below the 80% critical value are deleted, so the shapes defined by the scatter plots are representative of the corresponding confidence sets. For comparison, we have overlaid on these the boundary of the joint likelihood region, computed numerically. We use this set because it is often considered the gold standard for parametric inference. Notice that the boundary sampling method(Row 2 and 4) captured the boundaries of confidence sets better than the independent sampling method(Row 1 and 3). This shows that independent sampling is inefficient for describing the confidence sets. Our second observation is that the joint boundary sampling plot of $\tilde{\boldsymbol{\theta}}_L$ was very successful at representing the shape of the numerically calculated joint likelihood confidence regions.

As plots to compare inference functions, we can see that the joint score-based region was much larger than the likelihood region, and the empirical likelihood region $\tilde{\boldsymbol{\theta}}_{EL}$ was smaller than the others. This is surprising, and a bit disappointing, given that there should be more uncertainty about the parameter when the likelihood is unknown than when it is known. The smaller joint confidence region based on $\tilde{\boldsymbol{\theta}}_{EL}$ might reflect a known undercoverage issue for smaller sample sizes(Owen (1988); Hall and La Scala (1990); Qin and Lawless (1994)). The shape of the QIF$_u$ region for (β_2, σ^2) was shifted more toward small σ^2 than those of the likelihood, the score statistic, and the empirical likelihood. Notice that the boundary samples, especially of the score statistic and the QIF$_u$, appear to be comparatively sparse near the set boundary for large confidence levels, suggesting the need to do more sampling in those cases.

4.1.1. Improving coverage properties

Confidence regions based on using an asymptotic critical value inevitably have errors in coverage at small sample sizes. The methods described here can be used with adjusted critical values. For illustration, we applied our visualization method to two types of adjustment designed to improve the coverage of the empirical likelihood. We first considered using the bootstrap method to find the critical value of the empirical likelihood. For this method we assumed that normality in the error terms of the linear regression model was known so that we

could use parametric bootstrap adjustments. We first generated 5,000 parametric bootstrap samples of size 30, and calculated 5,000 bootstrap parameter values of $\boldsymbol{\theta}$. We then obtained the bootstrap empirical likelihood which provided an empirical critical value for the 80% confidence level in the full parameter space. These critical values were larger than the asymptotic ones, so the volumes of the sets increased.

The second method we considered was an adjustment to the empirical likelihood proposed by Chen, Mulayath and Bovas (2008). The adjusted empirical likelihood was designed to be well-defined for all parameter values and to retain all the optimal properties of the empirical likelihood, while improving coverage probabilities, all without appealing to bootstrap simulation or Bartlett correction.

In order to evaluate the effect of these alternative critical values on the confidence set shapes of the empirical likelihood, we reanalyzed the cheese data set using the same 700 $\mathbf{z}$'s and the same nine α values employed for the other inference functions. We denote simulated confidence distribution values describing the adjusted empirical likelihood region using the asymptotic critical value and the empirical likelihood region using the bootstrap critical value as $\tilde{\boldsymbol{\theta}}_{ADEL}$ and $\tilde{\boldsymbol{\theta}}_{ELBOOT}$, respectively.

The last three columns of Figure 3 provide visual evidence for two types of adjustment. For boundary samples, the sampling plot of $\tilde{\boldsymbol{\theta}}_{ADEL}$(Column 5) was indeed larger than that of $\tilde{\boldsymbol{\theta}}_{EL}$(Column 4). This agrees with what Chen, Mulayath and Bovas (2008) showed: the adjusted empirical likelihood region using the critical value of the chi-square distribution was generally larger than the unadjusted empirical likelihood using the same critical value. We can also see that the joint confidence regions of $\tilde{\boldsymbol{\theta}}_{ELBOOT}$ (Column 6) contained the joint confidence regions of $\tilde{\boldsymbol{\theta}}_{EL}$ (Column 4) and $\tilde{\boldsymbol{\theta}}_{ADEL}$(Column 5), suggesting it had the greatest coverage. See Owen (2001, pp.33) for further discussion.

4.1.2. Comparison with bootstrap confidence sets

In Section 1 we explained why one might prefer to use inference-function based confidence sets to parametric-bootstrap confidence sets. We now use our linear regression example to compare these two methods.

Assuming that the error terms in the multiple linear regression model to be independent and normally distributed, we generated 5,000 parametric bootstrap samples of size 30, and calculated 5,000 bootstrap parameter values. We then obtained the bootstrap likelihood ratio statistic and calculated an empirical critical value for the 80% confidence level in the full parameter space. Figure 4 shows the 80% joint sampling plots using boundary samples and parametric bootstrap
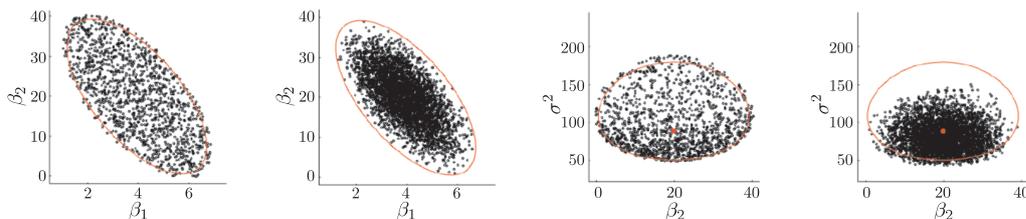
Figure 4. 80% joint likelihood confidence regions (red) superimposed on joint simulated confidence regions(black dots) using boundary samples(first and third) and bootstrap samples(second and fourth) for (β_1, β_2) and (β_2, σ^2). Note that the first two plots are the joint confidence regions for (β_1, β_2), and the last two plots are the joint confidence regions for (β_2, σ^2).

Table 2. Case-control study - relationship between smoking and myocardial infarction.

		Y			
		0	1-24	>25	
X	Control	25	25	12	62
	Myocardial infarction	0	1	3	4

samples overlaid on the numerical joint likelihood confidence sets for (β_1, β_2) and (β_2, σ^2).

The bootstrap method generated confidence sets of similar shapes as our boundary method. However, the samples from the bootstrap method became sparser as one approached the possible boundaries of the confidence sets, making it harder to identify the true boundary. It also seems plausible that the bootstrap method had an undercoverage problem in the (β_2, σ^2) plot. By comparing Figure 4 with Figure 3, we can also see that the empirical critical value increased the size of the confidence sets used for boundary sampling.

4.2. Independent multinomial model

Table 2 shows the data collected from a case-control study relating the disease-status, X(control or myocardial infarction), to the level of smoking, Y(number of cigarettes per day)(Agresti (2002, p.98)). Given the marginal counts, $n_1 = 62$ and $n_2 = 4$, a control and a myocardial infarction sample were independently obtained. Then the probability function for these counts in the table is the product of two independent multinomial functions, $\mathbf{n}_i = (n_{i1}, n_{i2}, n_{i3}) \sim multinomial(n_i, p_{1|i}, p_{2|i}, p_{3|i})$ where $i = 1, 2$ and $p_{j|i}$ is the probability that a subject in the ith disease-status has the jth level of smoking.

To examine our methodology, we consider a highly nonlinear parameter. One measure of association between smoking level and myocardial infarction is

a variant of the gamma coefficient introduced in Lang (2008) :

$$\gamma(\boldsymbol{\theta}) \equiv P(Y_2 > Y_1 \mid Y_1 \neq Y_2) = \frac{p_{1|1}(p_{2|2} + p_{3|2}) + p_{2|1}p_{3|2}}{1 - \sum_{j=1}^{3} p_{j|2}p_{j|1}}. \qquad (4.1)$$

Here Y_1 and Y_2 are the smoking levels for a control subject and a myocardial infarction subject, respectively, and $\boldsymbol{\theta} = (p_{1|1}, p_{2|1}, p_{1|2}, p_{2|2})'$. This gamma variant $\gamma(\boldsymbol{\theta})$ in (4.1) is the probability that a myocardial infarction subject smokes more than a control subject if smoking levels of subjects in each category of X are different. A value of $\gamma(\boldsymbol{\theta}) = 0.5$ means no difference.

Since the maximum likelihood estimate of $\boldsymbol{\theta}$ was $\hat{\boldsymbol{\theta}} = (0.403, 0.403, 0, 0.25)'$, $\gamma(\hat{\boldsymbol{\theta}}) = 0.9358$, association between smoking level and myocardial infarction appeared plausible. We investigated the confidence regions for $\gamma(\boldsymbol{\theta})$ by using a numerically calculated region based on the likelihood ratio statistic and samples from four inference functions, $\tilde{\boldsymbol{\theta}}_L$, $\tilde{\boldsymbol{\theta}}_S$, $\tilde{\boldsymbol{\theta}}_{Q_u}$, and $\tilde{\boldsymbol{\theta}}_{EL}$.

For the numerically calculated profile likelihood region, we used an approach proposed by Lang (2008). Lang (2008) developed a numerical method to compute score and profile likelihood confidence intervals for a function of parameters $g(\boldsymbol{\theta})$ in a contingency table where the counts had a product multinomial distribution. His approach consists of two parts, computation of the restricted maximum likelihood estimator for $g(\boldsymbol{\theta})$ and, given the restricted maximum likelihood estimator, calculation of the confidence interval for $g(\boldsymbol{\theta})$ using a simple and efficient algorithm(a so called 'sliding quadratic' algorithm). Note that $\gamma(\boldsymbol{\theta})$ in (4.1) was one of four examples considered by Lang (2008). Two R-functions(mph.fit and ci.table) for the algorithm are available from Professor J. B. Lang (e-mail: jblang@stat.uiowa.edu).

Here we apply our visualization method to picture the inference function of $\gamma(\boldsymbol{\theta})$, and see if our method can recover the numerically calculated profile likelihood function. Note that our proposed methodology does not require computing the restricted maximum likelihood estimator over $\gamma(\boldsymbol{\theta}) = \nu$ for a variety of ν values.

For the likelihood, score statistic, QIF_u, and empirical likelihood regions, we generated 1,000 $\mathbf{z}$'s from $N(0, I_4)$ and chose six α values designated to target the 20%, 40%, 60%, 80%, 90%, and 95% confidence intervals for $\gamma(\boldsymbol{\theta})$. In all there were 1,000 $\alpha(\mathbf{z})$'s in the independent sample and 6,000 $(\alpha, \mathbf{z})$'s in the boundary sample. The logistic transformation for $\boldsymbol{\theta}$ was considered, and the expected Fisher information for $\boldsymbol{\theta}$ was used in the Wald statistic ray and the score test statistic.

We first investigated the star-shaped assumption. We found that all 1,000 $\alpha(\mathbf{z})$'s and 6,000 $(\alpha, \mathbf{z})$'s were acceptable for the likelihood, score, and empirical likelihood regions, as the number of $\tilde{\epsilon}$ for every $(\alpha, \mathbf{z})$ was two. However, the
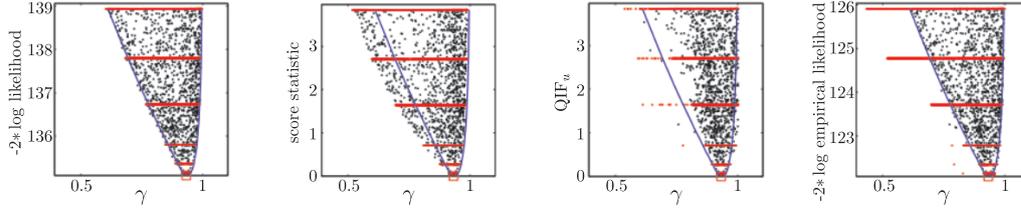
Figure 5. Numerically calculated twice negative profile log likelihood ratio(blue line) superimposed on profile sampling plots for $\gamma(\boldsymbol{\theta})$ based on $\tilde{\boldsymbol{\theta}}_L$, $\tilde{\boldsymbol{\theta}}_S$, $\tilde{\boldsymbol{\theta}}_{Q_u}$, and $\tilde{\boldsymbol{\theta}}_{EL}$ (red square : $\hat{\boldsymbol{\theta}}$). Note the black dots represent independent samples; the boundary samples appear as red bands.

QIF_u region had a small percentage of unacceptable $(\alpha, \mathbf{z})$'s with more than one solution on the either side of $\epsilon = 0$: 0.4%, 2.4%, 4.7%, 4.8%, 5.4% and 4.4% $\mathbf{z}$'s in the boundary sample were unacceptable for the 20%, 40%, 60%, 80%, 90%, and 95% confidence levels, respectively. The only difference between the score statistic and QIF_u in this example is that the score statistic used the expected Fisher information matrix, but the QIF_u used an empirical covariance matrix(equivalent to the observed information matrix in this model).

Figure 5 shows the numerically calculated twice negative log profile likelihood ratio for $\gamma(\boldsymbol{\theta})$ overlaid on the various sampling plots. For likelihood and score regions the independent samples and boundary samples produced same shape of profile functions. However, for QIF_u and empirical likelihood the independent samples showed serious sparsity near the confidence interval boundaries for $\gamma(\boldsymbol{\theta})$, with boundary sampling being somewhat better.

Figure 5 also shows that the boundary sampling succeeded in recovering the shape of the profile likelihood for $\gamma(\boldsymbol{\theta})$. We can also note that the confidence intervals based on the other methods were much larger than that of the profile likelihood, especially for moderate and large confidence levels.

4.3. Parametric finite mixture model : confidence and mixture labels

In this section we introduce an interesting, but challenging, problem from finite mixture analysis. It was the original motivation for our sampling based visualization.

Figure 6(a) is the histogram of log concentrations of antibodies against mumps measured from blood samples of 385 14-year old children with no vaccination(Flury (1997)). In order to see if diseases such as mumps are successfully controlled or not, one needs to know how immunity to specific disease is widespread in a particular age group by estimating the proportion of immune subjects in the same group. Flury (1997) pointed out the existence of two different groups in this data, one group whose members are immune children (near

three) and the other group whose children are susceptible to mumps (near zero), but these categories are not directly observable from this data.

Flury (1997, p.663) used maximum likelihood estimation to fit a two-component normal mixture model with equal variances

$$g(\mathbf{Y}; \boldsymbol{\theta}) = \pi_1 N(\mathbf{Y}; \mu_1, \sigma^2) + \pi_2 N(\mathbf{Y}; \mu_2, \sigma^2), \tag{4.2}$$

where $\pi_1 + \pi_2 = 1$ and $\boldsymbol{\theta} = ((\pi_1, \mu_1, \sigma^2)', (\pi_2, \mu_2, \sigma^2)')$, each column corresponding to a component. The component with the larger mean in (4.2) is generally believed to be the immune group, so the corresponding mixing proportion could be used as an estimate of the proportion of children immune to mumps at this age. The maximum likelihood estimate of the parameter vector in (4.2) was $\hat{\boldsymbol{\theta}} = ((0.256, -0.319, 0.935)', (0.744, 2.893, 0.935)')$.

For further statistical inference one might want to construct confidence sets for the parameters and estimate the standard errors of parameter estimates. However defining the likelihood confidence regions or the standard errors of the estimators is not easy. There is a label nonidentifiability intrinsic to the mixture model: in a mixture problem where the number of components is fixed, label nonidentifiability occurs because the parameter is only identifiable up to a column permutation of $\boldsymbol{\theta}$. For instance, $\hat{\boldsymbol{\theta}}$ for (4.2) has the same density and likelihood as $\hat{\boldsymbol{\theta}}^{cp} = ((0.744, 2.893, 0.935)', (0.256, -0.319, 0.935)')$, where $\boldsymbol{\theta}^{cp}$ is the column permutation of $\boldsymbol{\theta}$, and so, for example $\hat{\mu}_1$ could be defined as either 2.893 or -.319. In repeated sampling from this model, the estimator labels can be switched arbitrarily and so there is really no meaning to a finite sample standard error for $\hat{\mu}_1$.

The good news is that there is a form of asymptotic identifiability (Redner and Walker (1984)) that guarantees the existence of a consistent way to choose a permutation of the estimator as the sample size grows large. In practice, this means that the sample size must be large relative to the separation of the components. Kim (2008) pointed out that for the proper use of finite mixture models, one should measure how confident one is in using the asymptotic identifiability to label the parameters, and so to construct confidence sets.

The reason that asymptotic identifiability occurs is that the parameters are *locally identifiable*. That is, even though other parameters (such as $\boldsymbol{\theta}^{cp}$) generate the same probability distribution as $\boldsymbol{\theta}$, one can still find an open neighborhood of $\boldsymbol{\theta}$ such that every parameter in the same neighborhood generates a unique distribution. If we can be relatively confident that the estimators lie in such a locally identifiable region, one might reasonably appeal to asymptotic identifiability.

This is how Kim (2008) used likelihood to quantify identifiability. In the full parameter space of (4.2) there are really two maximum likelihood estimators,

corresponding to $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{cp}$, each creating a mode of equal height on the like-lihood surface. Similarly, in a confidence distribution simulation based on the mode $\hat{\boldsymbol{\theta}}$, for every $\boldsymbol{\theta}$ generated there would be a corresponding permuted version $\boldsymbol{\theta}^{cp}$ generated from the second. If the $100(1-\alpha)\%$ likelihood region for $\boldsymbol{\theta}$ in (4.2) decomposes into two distinct regions, one around each mode, and these two regions are locally identifiable, then we say the problem has a *labeling confidence level* of $100(1-\alpha)\%$. The reason is that either region could be used to describe profile confidence sets for the labeled parameters within the region. If the confidence level at which the two modal regions are well separated is large (say, 95%), we argue that one can be comfortable appealing to asymptotic identifiability. Otherwise, one should be cautious in the use of asymptotic identifiability as well as Fisher information.

We consider how one might visualize the likelihood region in a manner that clarifies the separation between the two modal regions. Unfortunately, the like-lihood region can be very complex geometrically. Moreover, producing multiple exploratory plots using numerical profiles is a daunting and time-consuming task.

To create exploratory plots we propose to use boundary sampling based on the confidence distribution associated with one of the two modes. We generated 2,000 $\mathbf{z}$'s from $N(0, I_4)$ and considered 12 confidence levels: $1 - \alpha = [0.01, 0.1, \ldots,$ 0.9, 0.95, 0.99]. Note that all 24,000 $(\alpha, \mathbf{z})$'s were acceptable. However, now we can plot both $\tilde{\boldsymbol{\theta}}(\epsilon, \mathbf{z})$ and its relabeled version, thereby creating two points, one in each modal region. One can now answer the question of the separation of the modal regions by determining the minimal separation of the points from the two modal regions at the given confidence level. The next step, sometimes a challenging one, is to find the profile functions that give the sharpest picture of the separation of confidence sets at various α-levels.

In this case, we consider the plot formed by the sample values of (π_1, μ_1). See Figure 6(b). If we plot these values from the original sample, we get a tight cluster around $(\hat{\pi}, \hat{\mu}) = (0.256, -.319)$. If we plot them from the permuted samples, which correspond to the second mode, we get a second tight plot around $(0.744, 2.893)$. In Figure 6(b) we provide a joint (π, μ) plot along with an appropriate 95% contour. In this (π, μ) profile one can see that the two modal regions are clearly separated and are elliptically shaped. The implication is that one can, with high labeling confidence, use order constraint labeling on either μ or π to construct confidence sets and one can be relatively comfortable using Fisher information for the standard error of the estimators. For example, the resulting π_1 confidence set would be based on a single profile plot using those π-values that were associated with μ_1 in the confidence distribution simulation(See Figure 6(c)).
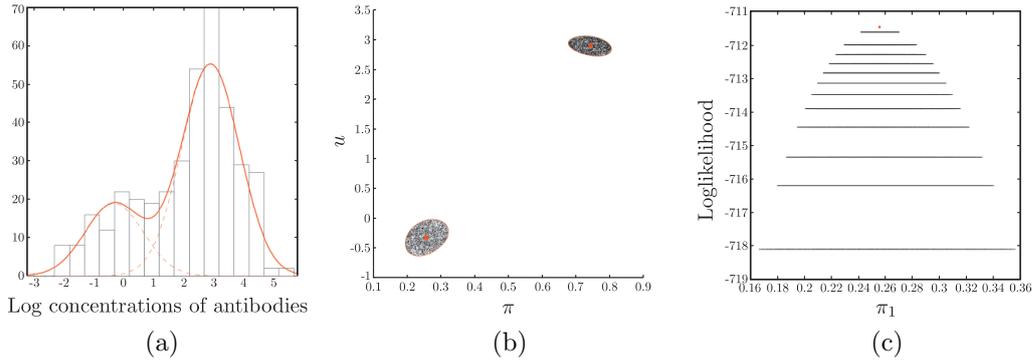
Figure 6. (a) - a histogram and a fitted mixture density for log concentrations of antibodies against mumps from 385 unvaccinated children (a solid red is an estimated mixture density and a broken red is a fitted component density); (b) - an approximate 95% profile of (π, μ) (the red square is $\hat{\boldsymbol{\theta}}$, the black dots represent boundary samples and a numerically calculated 95% likelihood confidence set appears as red contour); (c) - a π_1-profile boundary sampling plot.

The labeling problem was virtually nonexistent in this data set; in fact, any number of different profile plots would have revealed how the two modal likelihood regions were clearly separated. However, the problem becomes considerably more difficult when data and model are multivariate and when the mixture components are not so well separated. More methodology for this problem can be found in Yao and Lindsay (2009) for Bayesian sampling, and in Kim (2008) for likelihood. One can find therein profile functions that work quite well at separating the two modes to the maximal extent, and so help in determining the labeling confidence level.

5. Simulation Study

In this section we report on a small-scale simulation to examine the coverage probabilities of our proposed method for a nonlinear function of the parameters. We also compare the coverage probabilities of our procedure with the numerical profile procedure.

Consider the simulation model discussed in Lang (2008), the two independent multinomial models used in Section 4.2, and follow the simulation set-up there (simulation model, parameters, sample sizes, number of simulations and use of asymptotic critical value in defining a confidence set). We first generated 10,000 samples of size (n_1, n_2) from two independent multinomial models with the cell probabilities $\mathbf{p} = ((p_{1|1}, p_{2|1}, p_{3|1}), (p_{1|2}, p_{2|2}, p_{3|2}))$. Then we estimated the coverage probabilities for the nominal 95% confidence interval for $\gamma(\boldsymbol{\theta})$ in (4.1). Here we considered the likelihood function and score statistic,

Table 3. Estimated coverage probabilities.

(n_1, n_2)	$(\mathbf{p}_1, \mathbf{p}_2)$	True γ	Estimated coverage probabilities(estimated simulation standard errors)					
			Likelihood			Score		
			Independent	Boundary	Lang	Independent	Boundary	Lang
(62,62)	$\mathbf{p}^{(1)}$	0.903	0.9216	0.9444	0.9448	0.9285	0.9507	0.9492
			(0.0027)	(0.0023)	(0.0023)	(0.0026)	(0.0022)	(0.0022)
(62,62)	$\mathbf{p}^{(2)}$	0.778	0.9243	0.9459	0.9497	0.9260	0.9482	0.9506
			(0.0026)	(0.0023)	(0.0022)	(0.0026)	(0.0022)	(0.0022)
(62,62)	$\mathbf{p}^{(3)}$	0.500	0.9246	0.9465	0.9464	0.9263	0.9478	0.9472
			(0.0026)	(0.0023)	(0.0023)	(0.0026)	(0.0022)	(0.0022)
(62,4)	$\mathbf{p}^{(1)}$	0.903	0.9630	0.9723	0.9742	0.9456	0.9611	0.9606
			(0.0019)	(0.0016)	(0.0016)	(0.0023)	(0.0019)	(0.0019)
(62,4)	$\mathbf{p}^{(2)}$	0.778	0.9589	0.9710	0.9713	0.9515	0.9657	0.9668
			(0.0020)	(0.0017)	(0.0017)	(0.0021)	(0.0018)	(0.0018)
(62,4)	$\mathbf{p}^{(3)}$	0.500	0.9092	0.9271	0.9185	0.9725	0.9820	0.9813
			(0.0029)	(0.0026)	(0.0027)	(0.0016)	(0.0013)	(0.0014)

and used the cell probabilities : $\mathbf{p}^{(1)} = ((0.4, 0.4, 0.2), (0.04, 0.24, 0.72))$, $\mathbf{p}^{(2)} = ((0.4, 0.4, 0.2), (0.2, 0.2, 0.6))$ and $\mathbf{p}^{(3)} = ((0.4, 0.4, 0.2), (0.4, 0.4, 0.2))$. When we applied our method to each data set generated from each combination of (n_1, n_2) and $\mathbf{p}$, we used 500 $\mathbf{z}$'s from $N(0, \mathbf{I}_4)$.

Table 3 gives the estimated coverage probabilities and their estimated simulation standard errors. Here we cited the simulation results of the numerical profile procedure in Lang (2008) (labeled as "Lang"). Table 3 shows that the coverage probabilities of the boundary sampling agreed with those of the numerical profile procedure, the differences being insignificant relative to the magnitude of the simulation errors. Furthermore, the procedure based on boundary samples had coverage probabilities close to the nominal level 0.95 for the confidence interval of $\gamma(\boldsymbol{\theta})$, even when the true value of $\gamma(\boldsymbol{\theta})$ was close to the boundary of the parameter space and the sample size was small. Note that the theoretical confidence sets we were simulating were formed from inverting tests.

In a similar fashion, we could view the sampled confidence set, with formal boundaries defined in Section 3.5, as providing tests for hypothesized parameter values by the rule, reject $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ if it fails to be in the confidence set. From this interpretation, one minus coverage is the probability of Type I error. As shown in Table 3, there was a satisfying agreement between the error levels of our procedure and the errors based on numerical profile calculations.

6. Conclusion

In this paper we proposed a new sampling based methodology to visualize the confidence sets generated by a given inference function. Our methodology can generate samples to picture the boundaries of (profile) confidence sets over a

wide range of parameters of interest in a single simulation. The proposed method only requires the generation of samples from the standard multivariate normal distribution, the one-time calculation of the dispersion matrix at the estimator, and repeated one-dimensional root finding.

In applications to three examples and in a small simulation, we observed that our boundary sampling variant was superior to independent sampling in identifying the boundaries of the confidence regions and having comparable coverage probabilities. We also found, in the case of likelihood, that the boundary sampling method was very successful at visualizing numerically calculated profile likelihood boundaries. Generally, our sampling method provides a simple way to compare the confidence regions generated by a variety of inference functions on multiple functions of parameters.

We used parametric models in all our examples. A valuable extension of this research would be to investigate the performance of our visualization methodology for QIF and the empirical likelihood when there is a semiparametric model with parameters defined by estimating functions and the number of estimating functions is larger than the number of parameters.

Finally, our proposed method provides a random boundary to the confidence set. If one wants to use our sampling methods to carry out a hypothesis test, one needs to consider the issues of sparsity of sampling as the dimension of the parameter increases. Future work will consider accuracy of random boundary and stopping rules of (independent/boundary) sampling.

Acknowledgements

References

Agresti, A. (2002). *Categorical Data Analysis*. 2nd edited. Wiley, New York.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*, Chapman & Hall, London.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83**, 687-697.

Boos, D. D. (1992). On generalized score tests. *Amer. Statist.* **46**, 327-333.

Chen, J., Mulayath, A. and Bovas, A. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*. **17**, 426-443.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge Univ. Press.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society.* **26**, 528-535.

Flury, B. D. (1997). *A First Course in Multivariate Statistics.* Springer, New York.

Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* **55**, 231-244.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Hall, P. and La Scala, B. (1990). Methodology and algorithm of empirical likelihood. *Internat. Statist. Rev.* **58**, 109-127.

Hannig, J., Iyer, H. K. and Patterson, P. (2006). Fiducial generalized confidence intervals. *J. Amer. Statist. Assoc.* **101**, 254-269.

Kim, D. (2008). Mixture inference at the edge of identifiability. PhD thesis, Department of Statistics, The Pennsylvania State University, University Park, PA.

Lang, J. B (2008). Score and profile likelihood confidence intervals for contingency table parameters. *Statist. Medicine Sci.* **27**, 5975-5990.

Lindsay, B. G. and Qu, A. (2003). Inference functions and quadratic score tests. *Statist. Sci.* **18**, 394-410

Moore, D. S. and McCabe, G. P. (1989). *Introduction to the Practice of Statistics.* W.H. Freeman, New York.

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

Owen, A. (2001). *Empirical likelihood.* Chapman & Hall, New York.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300-325

Qu, P. (1998). Adaptive generalized estimating equations. Ph.D. thesis, Department of Statistics, The Pennsylvania State University, University Park, PA.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society.* **44**, 50-57.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195-239.

Weerahandi, S. (1993). Generalized confidence intervals. *J. Amer. Statist. Assoc.* **88**, 899-905.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *J. Amer. Statist. Assoc.*, **104**. 758-767.

Department of Mathematics and Statistics, Lederle Graduate Research Tower 1434, University of Massachusetts, Amherst, MA 01003-9305, USA.

E-mail: kdystat@math.umass.edu

Department of Statistics, Penn State University, 326 Thomas Building, University Park, PA 16802, USA.

E-mail: bgl@psu.edu