Constructing Confidence Regions of Optimal Expected Size

Chad M. SCHAFER and Philip B. STARK

This article presents a Monte Carlo method for approximating the minimax expected size (MES) confidence set for a parameter known to lie in a compact set. The algorithm is motivated by problems in the physical sciences in which parameters are unknown physical constants related to the distribution of observable phenomena through complex numerical models. The method repeatedly draws parameters at random from the parameter space and simulates data as if each of those values were the true value of the parameter. Each set of simulated data is compared to the observed data using a likelihood ratio test. Inverting the likelihood ratio test minimizes the probability of including false values in the confidence region, which in turn minimizes the expected size of the confidence region. We prove that as the size of the simulations grows, this Monte Carlo confidence set estimator converges to the Γ-minimax procedure, where Γ is a polytope of priors. Fortran-90 implementations of the algorithm for both serial and parallel computers are available. We apply the method to an inference problem in cosmology.

KEY WORDS: Minimax procedure; Minimax regret; Monte Carlo method; Multivariate confidence sets; Physical science application; Restricted parameter.

1. INTRODUCTION

The relationship between hypothesis tests and confidence estimators can be exploited to construct confidence sets with desirable properties. For a fixed confidence level, it is natural to seek a confidence set that is as small as possible. Evans, Hansen, and Stark (2005) (hereafter, EHS) showed that the $1-\alpha$ confidence set with smallest maximum expected measure can be found by inverting a family of level α tests of simple null hypotheses against a common simple alternative hypothesis. This is the *minimax expected size* (MES) procedure. This article gives a computationally efficient algorithm for approximating MES and other optimal confidence sets, including the less conservative *minimax regret* (MR) procedure, when the parameter—which can be multidimensional—is known to lie in a compact set.

The method is well suited to scientific problems in which the parameter satisfies a priori bounds and the distribution of the observed data depends on the parameter in a complex way—e.g., through a numerical model. For example, there are theoretical and observational constraints on cosmological parameters such as Hubble's constant and the age of the Universe. Those parameters in turn affect the distribution of angular fluctuations in the cosmic microwave background radiation (CMB). The constraints can be combined with observations of the CMB to sharpen inferences about the power spectrum of angular fluctuations. In the following, we illustrate the method on a simpler, but similar, problem: estimating cosmological parameters using observations of Type Ia supernovae.

There have been several studies of loss functions for set estimators. Cohen and Strawderman (1973b) considered loss functions that are linear combinations of size of the region and an indicator of whether the region covers the truth. Aitchison (1966), Aitchison and Dunsmore (1968), and Winkler (1972) considered interval estimates of real-valued parameters using a loss function that combines distance from the truth to the lower endpoint of the interval, distance from the truth to the upper endpoint, and the length of the interval. Casella and Hwang (1991) and Casella, Hwang, and Robert (1994) studied confidence sets that are optimal with respect to such loss functions.

Here, we restrict attention to confidence sets with $1-\alpha$ coverage probability and use a loss function that depends only on size. EHS, Hwang and Casella (1982), and Joshi (1969) used the measure ν of the confidence set as loss. The expected ν-measure of the confidence set is the "expected size." The MES procedure minimizes the maximum expected size of the confidence set. Instead of using a single measure ν, Hooper (1982) and Cohen and Strawderman (1973a) allowed the measure to vary with the true value θ of the parameter. The theory presented here can be extended to that more general case; we will present applications of the generalization in a sequel.

Typically the MES procedure cannot be found analytically; we show here how to approximate it numerically. The approximation has several parts, including approximating the MES procedure by the Γ-minimax expected size (Γ-MES) confidence procedure, where Γ is a convex set of prior probability distributions supported on a finite subset of Θ; and approximating the Γ-MES procedure numerically by optimization and Monte Carlo simulation. The support points of Γ are spread throughout Θ so that the Γ-minimax risk is close to the minimax risk. Section 4.3 discusses how to use the results of the Monte Carlo step to select good support points for Γ.

Constructing the Γ-MES procedure amounts to finding the element of Γ for which the Bayes risk is maximal: the Γ-*least favorable alternative* (Γ-LFA). Finding the Γ-LFA is conceptually simple, but can be computationally intensive. Kempthorne (1987) and Nelson (1966) gave algorithms to approximate the least favorable prior distribution over compact parameter spaces for general risk functions. Those algorithms require calculating the Bayes risk for an arbitrary prior, which can be analytically intractable. To overcome that problem, we approximate the risk using a novel Monte Carlo algorithm. We show that the maximum expected size of the approximated confidence set con-

Chad M. Schafer is Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *cschafer@stat.cmu.edu*). Philip B. Stark is Professor, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: *stark@stat.berkeley.edu*). This work was supported by NSF Grants #9872979 and #0434343, and by the AX Division at the Lawrence Livermore National Laboratory through the Department of Energy under contract W-7405-Eng-48. The authors thank the referees for many helpful comments.

Journal of the American Statistical Association
September 2009, Vol. 104, No. 487, Theory and Methods
DOI: 10.1198/jasa.2009.tm07420

verges to that of the Γ-MES procedure as the size of the Monte Carlo simulations increases. The algorithm is implemented as a Fortran-90 subroutine designed to run efficiently on distributed computers.

This article is organized as follows. Section 2 gives notation, assumptions, and theory. Section 3 derives a consistent estimator for the Bayes risk. Section 4 shows how that estimator, along with techniques from convex game theory, can be used to approximate the Γ-LFA through Monte Carlo simulation. Section 5 shows that the new approach can construct confidence sets that minimize risk for a general class of loss functions involving the measure of the confidence set, including one that leads to the minimax regret procedure. Section 6 applies the method to an inference problem in cosmology. Section 7 summarizes the results, and proofs are in the Appendix.

2. PRELIMINARIES

We have a family of probability distributions indexed by θ:

$$\mathcal{P} \equiv \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

The probability distributions are all defined on the same σ-field $\mathcal{B}$ over the set $\mathcal{X}$; all are dominated by the measure μ. The density of $\mathbb{P}_\theta$ with respect to μ is f_θ. The set Θ is itself endowed with σ-field $\mathcal{A}$. Elements of $\mathcal{A}$ are possible confidence sets. We assume that $(\theta, x) \mapsto f_\theta(x)$ is product measurable. The random quantity X—which could be multivariate—has distribution $\mathbb{P}_{\theta_0}$ for some unknown $\theta_0 \in \Theta$. The confidence region will be based on one observation of X and an observation of $U \sim U[0,1]$, a uniform random variable independent of X. We have a set $\mathcal{D}$ of *decision functions*, measurable mappings from $\Theta \times \mathcal{X}$ into $[0,1]$. Decision functions let us use X and U to make random subsets of Θ:

$$\mathbf{C}_d(X,U) \equiv \{\eta \in \Theta : d(\eta, X) \geq U\}. \tag{1}$$

Such sets are candidate confidence sets for θ_0. The chance that $\mathbf{C}_d(X,U)$ covers the parameter value $\eta \in \Theta$ when in fact $X \sim \mathbb{P}_\theta$ is

$$\begin{aligned}\gamma_d(\theta,\eta) &\equiv \mathbb{P}_\theta[\mathbf{C}_d(X,U) \ni \eta] = \mathbb{P}_\theta[d(\eta,X) \geq U] \\ &= \int_{\mathcal{X}} d(\eta,x) f_\theta(x) \mu(dx). \end{aligned} \tag{2}$$

Decision rules that correspond to $1-\alpha$ confidence sets are elements of

$$\mathcal{D}_\alpha \equiv \{d \in \mathcal{D} : \gamma_d(\theta,\theta) \geq 1-\alpha \text{ a.e.}(\nu)\}. \tag{3}$$

Let ν be a measure on $(\Theta, \mathcal{A})$. We define the risk of a confidence set to be its expected ν-measure:

$$\mathbf{R}(\theta, d) \equiv \mathbb{E}_\theta\big[\nu(\mathbf{C}_d(X,U))\big]. \tag{4}$$

Pratt (1961) showed that the expected measure of a confidence set is the integral of its *false coverage probability*, the chance that it incorrectly includes the parameter value η when the true value is θ:

$$\mathbb{E}_\theta\big[\nu(\mathbf{C}_d(X,U))\big] = \int_\Theta \gamma_d(\theta,\eta)\nu(d\eta). \tag{5}$$

Let $\mathbf{R}_\Theta(d)$ denote the maximum risk of d over all $\theta \in \Theta$. Since $f_\theta(x)$ and $d(\eta,x)$ are $\mathcal{A}\times\mathcal{B}$-measurable,

$$\mathbf{R}_\Theta(d) \equiv \sup_{\theta\in\Theta} \mathbf{R}(\theta,d) = \sup_\pi \int_\Theta \mathbf{R}(\theta,d)\pi(d\theta), \tag{6}$$

where the supremum is over all probability distributions π on $(\Theta, \mathcal{A})$. We will find a numerical approximation to the decision rule $d_{\mathbf{R}}$ with minimax risk over a smaller class of distributions Γ:

$$\mathbf{R}_\Gamma(d_{\mathbf{R}}) = \inf_{d\in\mathcal{D}_\alpha} \sup_{\pi\in\Gamma} \int_\Theta \mathbf{R}(\theta,d)\pi(d\theta). \tag{7}$$

In applications, Γ might be the polytope of probability distributions on p parameter values $\{\theta_i\}_{i=1}^p$ spread evenly across Θ, or chosen randomly if Θ is high-dimensional. This is an ad hoc element in our approach, but in Section 4.3 we describe how to choose $\{\theta_i\}$ so that $\mathbf{R}_\Theta(d_{\mathbf{R}})$ is not much larger than

$$\inf_{d\in\mathcal{D}_\alpha} \mathbf{R}_\Theta(d). \tag{8}$$

Our numerical approximation produces a member of $\mathcal{D}_\alpha$, a $1-\alpha$ confidence procedure valid for all $\theta \in \Theta$, but its risk is approximately Γ-minimax, rather than exactly Γ-minimax. The algorithm estimates the critical values for the individual tests by simulation, so the confidence level is approximately $1-\alpha$ rather than exactly $1-\alpha$.

2.1 Bayes-Minimax Duality

For any probability distribution π on $(\Theta, \mathcal{A})$, define

$$r_\pi(\eta, x) \equiv \frac{\int_\Theta f_\theta(x)\pi(d\theta)}{f_\eta(x)}. \tag{9}$$

This is the ratio of the likelihood of observing data x under the density mixed across values of θ according to the prior π to the likelihood under parameter value η.

The Bayes risk of d for prior π is

$$\begin{aligned}\mathbf{R}_\pi(d) &\equiv \int_\Theta \mathbf{R}(\theta,d)\pi(d\theta) \\ &= \int_\Theta \int_\Theta \gamma_d(\theta,\eta)\nu(d\eta)\pi(d\theta) \\ &= \int_\Theta \int_\Theta \int_{\mathcal{X}} d(\eta,x) f_\theta(x)\mu(dx)\nu(d\eta)\pi(d\theta) \\ &= \int_\Theta \int_{\mathcal{X}} d(\eta,x) f_\eta(x) r_\pi(\eta,x)\mu(dx)\nu(d\eta). \end{aligned} \tag{10}$$

The rule d is in $\mathcal{D}_\alpha$ if

$$\int_{\mathcal{X}} d(\eta,x) f_\eta(x)\mu(dx) \geq 1-\alpha \quad \text{a.e.}(\nu). \tag{11}$$

The optimal decision rule $d_\pi \in \mathcal{D}_\alpha$ for prior π minimizes (10) subject to (11). The optimal rule can be found using the construction in the Neyman–Pearson lemma:

Lemma 1.

$$\inf_{d\in\mathcal{D}_\alpha} \mathbf{R}_\pi(d) = \mathbf{R}_\pi(d_\pi), \tag{12}$$

where

$$d_\pi(\eta,x) = \begin{cases} 1, & r_\pi(\eta,x) < c_\eta \\ b_\eta, & r_\pi(\eta,x) = c_\eta \\ 0, & r_\pi(\eta,x) > c_\eta, \end{cases} \tag{13}$$

with the constants $b_\eta \in [0,1]$ and c_η chosen so that

$$\int_{\mathcal{X}} d_\pi(\eta,x) f_\eta(x)\mu(dx) = 1-\alpha. \tag{14}$$

If Γ is a collection of distributions on $(\Theta, \mathcal{A})$, then $\pi_0 \in \Gamma$ is a Γ-*least favorable alternative* if $\mathbf{R}_{\pi_0}(d_{\pi_0}) \geq \mathbf{R}_\pi(d_\pi)$ for all $\pi \in \Gamma$. The decision procedure d_0 is Γ-*minimax* if

$$\sup_{\pi\in\Gamma} \mathbf{R}_\pi(d_0) = \inf_{d\in\mathcal{D}_\alpha} \sup_{\pi\in\Gamma} \mathbf{R}_\pi(d) \equiv \mathbf{R}_\Gamma(d_{\mathbf{R}}). \tag{15}$$

Theorem 1 establishes the Bayes-minimax duality.

Theorem 1 (EHS, Corollary 1). If Γ is convex and π_0 is Γ-least favorable,

$$\inf_{d\in\mathcal{D}_\alpha} \sup_{\pi\in\Gamma} \mathbf{R}_\pi(d) = \mathbf{R}_{\pi_0}\big(d_{\pi_0}\big).$$

2.2 More Assumptions

Theorem 1 requires Γ to be convex. The following additional assumptions suffice for the Monte Carlo algorithm presented in Section 3 to converge to the correct value of the risk.

1. $\nu(\Theta) < \infty$.
2. If $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$, $\theta, \theta' \in \Theta$, there must be a measurable set $A \in \mathcal{A}$ for which $\theta \in A$, $\theta' \in A^c$, and $0 < \nu(A)/\nu(\Theta) < 1$.
3. The distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ all have the same support a.e.(ν).
4. The convex collection of priors Γ has a finite number of vertices.

The method is not practical unless:

1. For any fixed point $\theta \in \Theta$, it is computationally tractable to simulate from $\mathbb{P}_\theta$.
2. For each vertex δ_v of Γ, it is computationally tractable to calculate $r_{\delta_v}(\eta, x)$ for fixed η and x.

3. ESTIMATING THE BAYES RISK

A single set of simulations can be used to estimate d_π and $\mathbf{R}_\pi(d_\pi)$. We first show how to estimate $\mathbf{R}_\pi(d)$. Let $T \in \Theta$ be drawn at random from ν. Let $X \sim \mathbb{P}_\eta$ conditional on $T = \eta$. Recall from Lemma 1 that $r_\pi(\eta, X)$ is the test statistic for a test of the hypothesis $\theta_0 = \eta$. The test rejects the hypothesis for data x if $\mathbb{P}_\eta[r_\pi(\eta, X) \geq r_\pi(\eta, x)] \leq \alpha$. For any $d \in \mathcal{D}$,

$$\begin{aligned}
&\mathbb{E}[r_\pi(T, X)d(T, X)] \\
&\quad = \mathbb{E}\big[\mathbb{E}[r_\pi(T, X)d(T, X)|T]\big] \\
&\quad = \int_\Theta \left[\int_{\mathcal{X}} r_\pi(\eta, x)d(\eta, x) f_\eta(x)\mu(dx)\right]\nu(d\eta) \\
&\quad = \mathbf{R}_\pi(d).
\end{aligned} \tag{16}$$

Hence, for fixed π, the simulated distribution of $r_\pi(T, X)$ can be used to estimate the threshold for the Bayes decision rule and the Bayes risk of the Bayes decision.

We now show that the Monte Carlo estimate of the risk of the estimated optimal rule converges almost surely to $\mathbf{R}_\pi(d_\pi)$, uniformly in $\pi \in \Gamma$. Fix two positive integers n and q. These define the size of the Monte Carlo simulations; we consider later what happens as they increase. Let $T_1, T_2, \ldots, T_q$ be iid (ν) and let

$$\{X_{jk} : j = 1, 2, \ldots, q; k = 1, 2, \ldots, n\} \tag{17}$$

have distribution $\mathbb{P}_\eta$ conditional on $T_j = \eta$. Let $\{X_{jk}\}$ be independent, conditional on all of the T_j. Define a Monte Carlo estimate of $\mathbf{R}_\pi(d)$:

$$\widehat{\mathbf{R}}_\pi(d) \equiv \frac{1}{nq} \sum_j \sum_k r_\pi(T_j, X_{jk})d(T_j, X_{jk})K_j, \tag{18}$$

with r_π as defined in Equation (9). Here,

$$K_j \equiv \left[K \times \left(\frac{1}{n}\sum_{v=1}^{p}\sum_{k=1}^{n} r_{\delta_v}(T_j, X_{jk})\right)^{-1}\right] \wedge 1, \tag{19}$$

with $K > p$. The factor K_j makes $\widehat{\mathbf{R}}_\pi(d)$ uniformly bounded (in π), a technical requirement to prove convergence; it also limits the effect of simulation outliers. Although $\mathbb{E}(r_{\delta_v}(T_j, X_{jk})) = 1$, $r_{\delta_v}(T_j, X_{jk})$ can be large. But because $d(T_j, X_{jk}) = 0$ when $r_{\delta_v}(T_j, X_{jk})$ is large, such values do not affect the estimated risk. Hence, we recommend choosing K very large.

We next construct decision procedures supported on the simulated datasets $\{X_{jk}\}$. For each j, such a decision procedure is a vector of length n with entries in $[0, 1]$. Fix α and define $\mathcal{D}'_\alpha$ to be the class of decision procedures that satisfy

$$\sum_k d(T_j, X_{jk}) \geq n(1 - \alpha) \qquad \forall j. \tag{20}$$

Suppose $\widehat{d}_\pi$ minimizes $\widehat{\mathbf{R}}_\pi(d)$ among all $d \in \mathcal{D}'_\alpha$. Recall that d_π minimizes $\mathbf{R}_\pi(d)$ over all $d \in \mathcal{D}_\alpha$.

Theorem 2. As $n \to \infty$ and $q \to \infty$,

$$\widehat{\mathbf{R}}_\pi(\widehat{d}_\pi) \xrightarrow{\text{a.s.}} \mathbf{R}_\pi(d_\pi) \tag{21}$$

uniformly in $\pi \in \Gamma$.

Proof. See the Appendix.

Corollary 1. As $n \to \infty$ and $q \to \infty$,

$$\sup_{\pi\in\Gamma} \widehat{\mathbf{R}}_\pi(\widehat{d}_\pi) \xrightarrow{\text{a.s.}} \mathbf{R}_\Gamma(d_{\mathbf{R}}). \tag{22}$$

For a given set of simulations of the random quantities, a member of Γ that maximizes $\widehat{\mathbf{R}}_\pi(\widehat{d}_\pi)$ can be found numerically. Corollary 1 shows that when n and q are large enough, the Bayes risk of this supremal prior is close to the Bayes risk of the Γ-least favorable prior.

4. IMPLEMENTING THE ALGORITHM

We seek the (in simulations) Γ-least favorable prior: the $\pi \in \Gamma$ that maximizes $\widehat{\mathbf{R}}_\pi(\widehat{d}_\pi)$. [Recall that $\widehat{d}_\pi$ is the decision procedure $d \in \mathcal{D}'_\alpha$ that minimizes $\widehat{\mathbf{R}}_\pi(d)$.] Finding the Γ-least favorable prior amounts to finding the optimal strategy in a convex game, as we shall see. Theorem 2 shows that the value of this convex game is an arbitrarily good approximation to the Γ-minimax risk as the size of the simulations increases.

4.1 Matrix Games and Minimax Procedures

We cast Equation (18) in matrix form. Define the n by p matrix $\mathbf{A}_j$ with elements

$$A_{jkv} = r_{\delta_v}(T_j, X_{jk})K_j. \tag{23}$$

Let

$$\mathbf{A} \equiv \frac{1}{nq}[\,\mathbf{A}_1 \quad \mathbf{A}_2 \quad \cdots \quad \mathbf{A}_q\,]^T. \tag{24}$$

For a given decision rule d, let $\mathbf{d}_j$ be the n-vector whose kth entry is $d(T_j, X_{jk})$. Define the nq-vector

$$\mathbf{d} \equiv [\,\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_q\,]^T. \tag{25}$$

Any prior $\pi \in \Gamma$ can be written as a convex combination of the vertices of Γ:

$$\pi = \sum_{v=1}^{p} w_v \delta_v, \tag{26}$$

for some $\mathbf{w} = \{w_v\}_{v=1}^{p}$ with $w_v \geq 0$ and $\sum_v w_v = 1$. The matrix form of Equation (18) is

$$\widehat{\mathbf{R}}_\pi(d) = \frac{1}{nq} \sum_{\mathbf{j}=1}^{q} \mathbf{d}_j^T \mathbf{A}_j \mathbf{w} = \mathbf{d}^T \mathbf{A} \mathbf{w}. \tag{27}$$

4.1.1 Solving Matrix Games. A two-player *convex game* is a triple $(\mathbf{A}, \mathcal{S}_1, \mathcal{S}_2)$ where $\mathbf{A}$ is an a by b matrix, $\mathcal{S}_1$ is a convex, compact subset of $\mathbb{R}^a$ and $\mathcal{S}_2$ is a convex, compact subset of $\mathbb{R}^b$. Player 1 chooses a *strategy*, an element $\mathbf{s}_1$ of $\mathcal{S}_1$. Player 2 picks a strategy $\mathbf{s}_2$ from $\mathcal{S}_2$. Player 1 pays Player 2 the amount $\mathbf{s}_1^T \mathbf{A} \mathbf{s}_2$.

Theorem 3. There exists a pair of strategies $(\mathbf{s}_{1*}, \mathbf{s}_{2*}) \in \mathcal{S}_1 \times \mathcal{S}_2$ such that for any $(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{S}_1 \times \mathcal{S}_2$,

$$\mathbf{s}_{1*}^T \mathbf{A} \mathbf{s}_2 \leq \mathbf{s}_{1*}^T \mathbf{A} \mathbf{s}_{2*} \leq \mathbf{s}_1^T \mathbf{A} \mathbf{s}_{2*}. \tag{28}$$

Proof. This is a direct consequence of the classic von Neumann Minimax Theorem. See, for example, Berkovitz (2002, theorem 5.2).

The pair $(\mathbf{s}_{1*}, \mathbf{s}_{2*})$ has a special optimality: By picking $\mathbf{s}_{1*}$, Player 1 minimizes his maximum loss. By picking $\mathbf{s}_{2*}$, Player 2 maximizes his minimum gain. *Solving the game* is finding this saddle point. The Brown–Robinson fictitious play algorithm (Brown 1951; Robinson 1951) is a simple iterative approach to solving the game.

The Brown–Robinson Algorithm. Fix a tolerance $\epsilon > 0$ and initial plays for each player: $\mathbf{s}_{1,0} \in \mathcal{S}_1, \mathbf{s}_{2,0} \in \mathcal{S}_2$. Set $i = 1$. Then:

1. Player 1 finds the strategy $\mathbf{s}_1 \in \mathcal{S}_1$ that minimizes $v_{1,i} \equiv \mathbf{s}_1^T \mathbf{A} \mathbf{s}_{2,i-1}$.
2. Player 2 finds the strategy $\mathbf{s}_2 \in \mathcal{S}_2$ that maximizes $v_{2,i} \equiv \mathbf{s}_{1,i-1}^T \mathbf{A} \mathbf{s}_2$.
3. If $v_{2,i} - v_{1,i} \leq \epsilon$, we are done. Otherwise, go to Step 4.
4. Set

$$\mathbf{s}_{1,i} \equiv (\mathbf{s}_1 + (i-1)\mathbf{s}_{1,i-1})/i \tag{29}$$

and

$$\mathbf{s}_{2,i} \equiv (\mathbf{s}_2 + (i-1)\mathbf{s}_{2,i-1})/i. \tag{30}$$

5. Increment i and return to Step 1.

Theorem 4 (Robinson 1951). For each iteration i in the Brown–Robinson algorithm,

$$v_{1,i} \leq \mathbf{s}_{1*}^T \mathbf{A} \mathbf{s}_{2*} \leq v_{2,i} \tag{31}$$

and

$$\lim_{i \to \infty} (v_{2,i} - v_{1,i}) = 0. \tag{32}$$

Theorem 5. If Player 1 uses strategy $\mathbf{s}_{1,i}$, the amount Player 1 pays Player 2 is less than

$$\mathbf{s}_{1*}^T \mathbf{A} \mathbf{s}_{2*} + v_{2,i+1} - v_{1,i+1} \tag{33}$$

no matter what strategy Player 2 uses.

Proof. From Theorem 4, $\mathbf{s}_{1*}^T \mathbf{A} \mathbf{s}_{2*} - v_{1,i+1} \geq 0$, so

$$\mathbf{s}_{1,i}^T \mathbf{A} \mathbf{s} \leq v_{2,i+1} \leq \mathbf{s}_{1*}^T \mathbf{A} \mathbf{s}_{2*} + v_{2,i+1} - v_{1,i+1}, \tag{34}$$

where $\mathbf{s}$ is any strategy in $\mathcal{S}_2$.

Theorem 5 ensures that when the Brown–Robinson algorithm terminates, Player 1 has a strategy that limits his maximum loss to at most ϵ more than the loss at the saddle point. Although the maximum loss is close to optimal, the strategy $\mathbf{s}_{1i}$ need not be close to $\mathbf{s}_{1*}$ in the norm.

4.1.2 Finding the Approximate Γ-LFA by Solving a Matrix Game. We now show that the problem of finding the Γ-LFA can be written as a (large) convex game. Player 1 is the statistician. He or she chooses the $100(1-\alpha)\%$ confidence procedure d. Player 2 is the adversary ("Nature"). He or she chooses $\mathbf{w}$, specifying a distribution π over the possible values of θ_0. Player 1's set of possible strategies, $\mathcal{S}_1$, has a special form. All elements of $\mathbf{d}$ must be between zero and one. Each of the vectors $\mathbf{d}_j$ that comprise $\mathbf{d}$ must sum to $(1-\alpha)n$. These restrictions on $\mathbf{d}$ make $\mathcal{S}_1$ is convex. The set $\mathcal{S}_2$ is the p-dimensional simplex: all p-vectors $\mathbf{w}$ with $w_i \geq 0$ and $\sum_i w_i = 1$; this is also convex.

The statistician and Nature play the convex game $(\mathbf{A}, \mathcal{S}_1, \mathcal{S}_2)$. The Brown–Robinson algorithm is well-suited to this problem, because for any fixed strategy $\mathbf{s}_{2,i-1}$ Nature picks, it is straightforward to find the strategy in $\mathcal{S}_1$ that is best for the statistician. Other algorithms for solving games (e.g., by linear programming) might take fewer iterations, but are difficult to implement when $\mathcal{S}_1$ is complex. Recent work by Bryan, McMahan, Schafer, and Schneider (2007) shows how to exploit sparsity of the payoff matrix to solve this convex game more efficiently.

4.2 Algorithmic Implementation and Parallelization

The approach parallelizes naturally: different processors can simulate independent samples of parameter values $\{T_j\}$ and data $\{X_{jk}\}$. Interprocessor communication is required only to calculate the outer sum in Equation (18), which involves $\{\widehat{\mathbf{R}}_{\delta_v}(d_\pi)\}_{v=1}^{p}$.

A Fortran-90 implementation of the algorithm with documentation is available at *http://www.stat.cmu.edu/~cschafer/LFA_Search*. The implementation is parallel and uses dynamic memory allocation.

Table 1 shows the largest storage requirements. The algorithm requires fast access to $n \times q \times p$ values, the simulated realizations of

$$\left\{ \left\{ \{r_{\delta_v}(T_j, X_{jk})\}_{j=1}^{q} \right\}_{k=1}^{n} \right\}_{v=1}^{p}. \tag{35}$$

Table 1. The primary storage requirements for the algorithm. The dimension of the parameter space Θ is b. The number of randomly chosen parameter points on each processor is q. The number of datasets generated from each random parameter is n

Data	Size	Precision
Random likelihood ratios	$n \times q \times p$	single
Random parameter points	$q \times b$	single
Thresholds	$q \times 2$	double
Confidence region	q	single

One might instead store the simulated data; however, these would be a $[n \times q \times (\text{dimension of } \mathcal{X})]$ array, and then the quantities $\{r_{\delta_v}\}_{v=1}^{p}$ would need to be calculated repeatedly. The operation count for calculating $\widehat{\mathbf{R}}_\pi(d_\pi)$ is $O(q \times n^2 \times p)$, not including calculating the likelihood $f_\eta(x)$ (the number of operations required to calculate the likelihood depends on details of the problem).

4.3 Choosing the Vertices of Γ

Whatever the true value of the parameter $\theta \in \Theta$, the coverage probability of the procedure is approximately $1 - \alpha$, but $\mathbf{R}(\theta, d_\mathbf{R})$ is guaranteed to be less than or equal to $\mathbf{R}_\Gamma(d_\mathbf{R})$, the Γ-minimax expected size, only if Γ includes a point mass at θ. The following result (proved in Section A.2) can help select the vertices $\{\delta_v\}$.

Theorem 6. For $\theta \in \Theta$, define

$$\mathcal{Z}(\theta) = \inf_{\mathbf{w}\in\mathcal{W}} \sup_{x\in\mathcal{X}} \left[\frac{f_\theta(x)}{\sum_{v=1}^{p} w_v f_{\delta_v}(x)} \right],$$

where $\mathcal{W}$ denotes the p-dimensional regular simplex. Then $\mathbf{R}(\theta, d) \le \mathcal{Z}(\theta)\mathbf{R}_\Gamma(d)$.

The theorem is useful in practice because the Monte Carlo simulations give approximations of $\mathcal{Z}(\theta)$ for each of the q randomly chosen values of θ:

$$\widehat{\mathcal{Z}}(\eta_j) = \left(\max_{\mathbf{w}\in\mathcal{W}} \min \mathbf{A}_j \mathbf{w} \right)^{-1}, \tag{36}$$

where the minimum is over the entries of the vector. Typically $q \gg p$, so the simulations approximate $\mathcal{Z}(\theta)$ for many values of θ. The estimates can be smoothed to reduce random variability from the simulations. Points in Θ for which $\mathcal{Z}(\theta)$ is large can be added to the vertices of Γ.

5. GENERAL LOSS FUNCTIONS

The theory developed above also applies to loss functions of the form $\nu(\mathbf{C}_d(x, u)) - \ell(\theta)$, where ℓ is any uniformly bounded function on Θ. A particularly interesting choice of ℓ is

$$\ell_r(\theta) \equiv \inf_{d\in\mathcal{D}_\alpha} \mathbb{E}_\theta(\mathbf{C}_d(X, U)).$$

The $d \in \mathcal{D}$ that minimizes the maximum expectation of this loss is the *minimax regret* (MR) procedure. The *regret at θ for using the decision function d* (DeGroot 1988) is the difference between the risk at θ of d and the infimal risk at θ over all decision functions. In the present problem, the regret at θ of the confidence procedure d is the difference between the expected size of the confidence set using procedure d when the true parameter value is θ, and the expected size of the confidence set that has smallest expected size when the true parameter value is θ. In some inference problems, parameter values θ for which $\ell_r(\theta)$ is relatively large can have a strong influence on the MES procedure: The least favorable alternative will place a lot of weight on such θ, increasing the expected size under other parameter values. Using MR can reduce this tradeoff.

Consider the following example (Schafer and Stark 2003; EHS): Suppose $X \sim N(\theta, 1)$ with $\theta \in [-3, 3]$. The LFA for the minimax expected length 95% confidence interval assigns probability one to $\theta = 0$. MES minimizes the expected length for $\theta = 0$, effectively ignoring other values of θ. The MR procedure provides a different tradeoff, as shown in Figure 1. The solid line is $\ell_r(\theta)$. The dashed–dotted line is the expected length of the MES interval. They are equal at $\theta = 0$. The dashed line is the expected length of the MR interval. The expected length of MR is about 20% larger than that of MES near $\theta = 0$, but about 33% smaller when θ is far from zero. The dotted line is the expected length of the truncated standard interval, $[X - 1.96, X + 1.96] \cap [-3, 3]$.

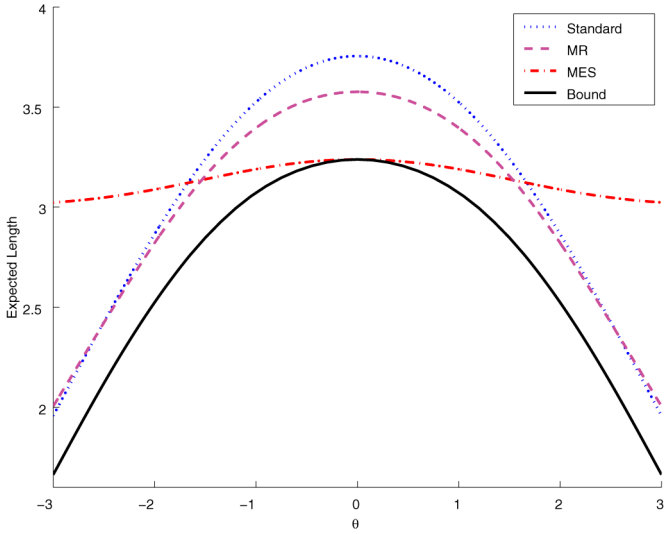

Figure 1. Expected lengths of 95% confidence intervals for a bounded normal mean $\theta \in [-3, 3]$ from the datum $X \sim N(\theta, 1)$, as a function of θ.

The minimum risk at θ, $\ell_r(\theta)$, is a complicated function. For fixed θ, $\ell_r(\theta)$ can be calculated using the Neyman–Pearson Lemma. If the vertices of Γ are point masses, the algorithm described in Section 4 can approximate $\ell_r(\theta)$ by taking the prior to be a point mass at θ. The subroutine LFA_Search mentioned in Section 4.2 can approximate the minimax regret procedure.

6. EXAMPLE: EXPANSION OF THE UNIVERSE

MES and MR were developed to solve scientific problems: find precise confidence sets for constrained physical parameters using theory that relates the parameters to a probability distribution on data. In many interesting problems, there are relatively few parameters (5–15); the constraints are nonlinear; and the model is not given in closed form, but rather as a complex computer simulation—a "black box" from the user's perspective. As a result, traditional methods for constructing confidence regions can be inaccurate, inapplicable, or computationally infeasible.

In this section, we use observations of Type Ia supernovae to compute MES and MR confidence sets for $\theta = (\Omega_m, H_0)$, where Ω_m is the amount of matter in the Universe relative to the "critical density" of matter required for the Universe to be spatially flat, and H_0 is the Hubble parameter, the current rate of expansion of the Universe. (See, for instance, Riess et al. 2007; Wood-Vasey et al. 2007; Wright 2007.) The stochastic model in this example is simple, which makes it possible to compare MES and MR confidence regions with some standard approaches; in more complicated problems, touchstone methods are rare.

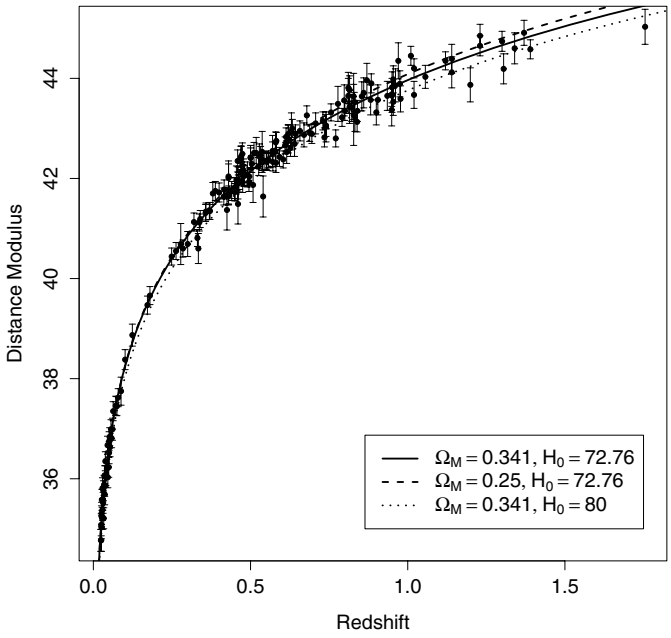

Figure 2. Supernovae data. The error bars represent $\pm 1\sigma$.

Type Ia supernovae are *standard candles*: two Type Ia supernovae at the same distance from the observer have the same apparent brightness. The difference between the apparent brightness and the brightness at the source is the *distance modulus*. The *redshift* of a supernova is the difference in wavelength of light emitted by the supernova in the reference frame of the supernova and in the reference frame of the observer. Figure 2 shows observations of redshift and distance modulus for 182 Type Ia supernovae, as reported by Riess et al. (2007). The error bars represent uncertainty in the distance modulus.

A standard theory relates redshift to distance modulus through a function of $\theta = (\Omega_m, H_0)$. Define

$$\mu(z \mid \theta) = 5 \log_{10}\left(\frac{c(1+z)}{H_0} \int_0^z \frac{du}{\sqrt{\Omega_m(1+u)^3 + (1-\Omega_m)}}\right) + 25, \tag{37}$$

where c is the speed of light. According to the theory, the observed pairs (z_i, Y_i) are realizations of $Y_i = \mu(z_i \mid \theta) + \sigma_i \epsilon_i$, where the ϵ_i are iid standard normal. The standard deviations σ_i are assumed to be known; in practice they are estimated from properties of the observing instrument.

6.1 Other Methods

There are several standard approaches to constructing confidence sets for θ in this problem. The confidence sets are derived from pivots that have approximately or exactly chi-squared distributions.

The CSQ (chi-squared) confidence set is based on the fact that

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu(z_i \mid \theta)}{\sigma_i}\right)^2 \tag{38}$$

has the chi-squared distribution with n degrees of freedom if θ is the true value of (Ω_m, H_0).

The MLE confidence set is based on the asymptotic distribution of the maximum likelihood estimator: If $\widehat{\theta}$ is the maximum likelihood estimator of θ and $\mathcal{I}(\theta)$ is the information matrix when θ is the truth, then

$$(\widehat{\theta} - \theta)^T \mathcal{I}(\theta)(\widehat{\theta} - \theta) \tag{39}$$

is approximately chi-squared distributed with two degrees of freedom.

The score test (SCR) confidence set is based on the asymptotic distribution of Rao's score test statistic (Lehmann and Romano 2005): Define

$$\mathcal{S}_j = \frac{\partial}{\partial \theta_j} \log f(\theta) \tag{40}$$

and $\mathcal{S} = [\mathcal{S}_1 \ \mathcal{S}_2]^T$. Then

$$\mathcal{S}^T \mathcal{I}^{-1}(\theta) \mathcal{S} \tag{41}$$

is approximately chi-squared distributed with two degrees of freedom.

6.2 Results

Figure 3 shows confidence sets for θ based on the data in Figure 2 for the five methods (CSQ, MLE, SCR, MES, and MR). The parameter vector θ was restricted to the compact set Θ with $60 \le H_0 \le 90$ and $500 \le \Omega_m H_0^2 \le 2500$, which is displayed in Figure 3 as the white area outlined in gray. [The quantity $\Omega_m H_0^2$ is constrained well by measurements of the cosmic microwave background radiation: the WMAP experiment (Spergel et al. 2007) found $\Omega_m H_0^2$ to be 1277, with a standard error of 80.0.] The MES region is the smallest. The SCR, MLE, and MR are very similar. The CSQ region is much larger. The areas of the sets are 2.86, 0.41, 0.40, 0.30, and 0.36 for CSQ, MLE, SCR,

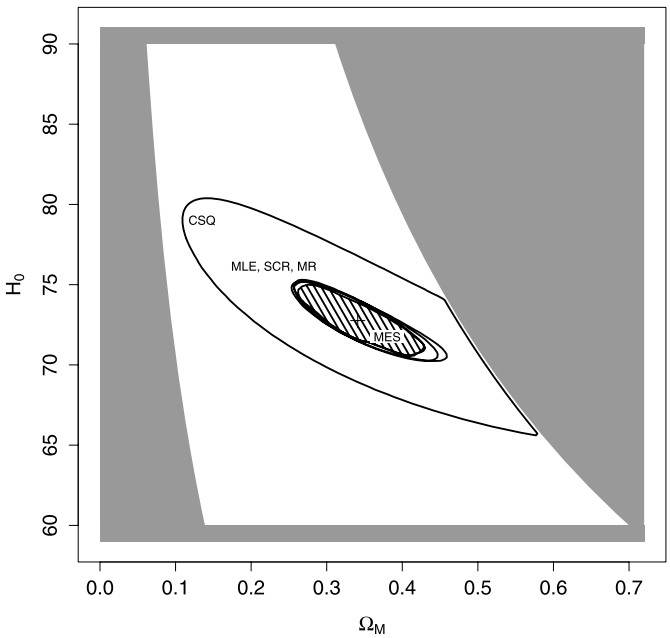

Figure 3. Confidence regions given by five approaches applied to the data shown in Figure 2. The smallest region (the hashed ellipse) is MES. The MLE, SCR, and MR regions are nearly identical. The larger truncated ellipse is the CSQ region. See text for descriptions of the methods. The plus sign marks the maximum likelihood estimate.

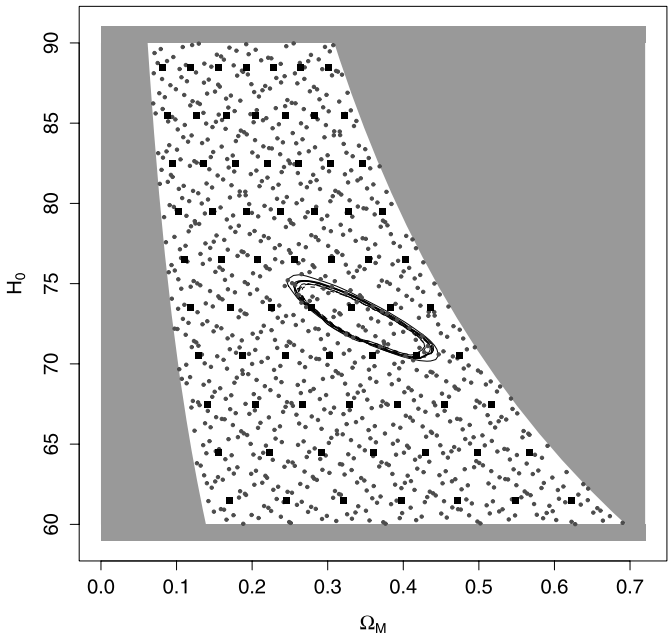

Figure 4. The $q = 1000$ hypothesized values of θ (circular dots) and the $p = 70$ alternative values (squares). The ellipses are confidence regions from five replications of the algorithm applied to the same data to show the sampling variability from the Monte Carlo steps. The five MES regions are the nearly overlapping dashed ellipses; the five MR regions are the slightly larger solid ellipses.

MES, and MR, respectively. For these data, the MES confidence set is smaller than standard confidence sets. Simulation results given below show that in this problem the expected sizes of MES, MR, MLE, and SCR sets are comparable, and CSQ is substantially larger. This suggests that MES and MR will be valuable in applications where the physical theory is complex, because then MLE and SCR are not generally feasible—CSQ is the only standard method available.

The MES and MR confidence regions were constructed five times independently to assess the variability due to Monte Carlo sampling. In each case, $p = 70$ alternatives were chosen from a regular grid, $q = 1000$ values of θ were chosen via a quasi-Monte Carlo scheme, and $n = 200$ datasets were simulated for each value of θ. Figure 4 plots the null values of θ, the alternative values, and the five MES and five MR regions that resulted. The variation across simulations is small.

On a desktop computer (3.80 GHz Pentium 4), the median time for the five runs was 25.00 minutes to calculate the MES region and 17.81 minutes to calculate the MR region.

Table 3. Fraction of "wins" of each of the five methods in simulations from five models. A method "wins" for a particular realization if its confidence region is the smallest among those that cover the true value of the parameter. Each row represents 5,000 replications

Truth		Proportion "won"				
Ω_m	H_0	CSQ	MLE	SCR	MES	MR
0.150	86.000	0.044	0.013	0.164	0.224	**0.554**
0.200	70.000	0.044	0.170	**0.534**	0.031	0.219
0.300	62.000	0.042	0.165	**0.349**	0.348	0.094
0.350	75.000	0.040	0.134	0.024	**0.533**	0.268
0.450	67.000	0.042	0.064	0.067	**0.780**	0.046

The size and coverage of the five methods were compared using simulation. We simulated 5,000 error vectors $\{\epsilon_i\}_{i=1}^{182}$ and added each to the predictions of five models. Using the same error vectors for five models helps isolate the effect of varying the model from the variability due to noise. The five methods were applied to each of the resulting 25,000 datasets. Table 2 lists the average size of the regions for each method, along with the empirical coverage of the true value of θ. Column "BND" gives the theoretical minimum average size of a $1 - \alpha$ confidence set for each model. Table 3 shows the "winning percentage" for each method. For each simulated data set, a method "wins" if its region is the smallest among those that cover the true value of θ.

The results are qualitatively similar to the performance in the bounded normal mean problem of Section 5: towards the center of Θ, where the lower bound on expected size is largest, MES performs best. But where the bound is smallest (when $\Omega_m H_0^2$ is small), MR has smaller expected size. In this example, comparing "wins" is about the same as comparing average size: controlling the expected size controlled the size in individual realizations. The coverage of the MES and MR procedures is close to the nominal confidence level (since these results are based on 5,000 realizations, the standard error of the coverage estimates is approximately 0.003). The SCR method performs well, which is not surprising given that it is asymptotically optimal under certain conditions (see, e.g., Lehmann and Romano 2005, theorem 13.5.5).

7. CONCLUSION

Minimax expected size and minimax regret procedures are theoretically attractive because they can exploit structural constraints. We show how to approximate minimax expected size and minimax regret confidence sets numerically for real, complex applications using Monte Carlo simulation. We establish

Table 2. Average sizes of confidence sets and their coverage in simulations from five models. Five thousand sets of 182 data were simulated from each model. The column "BND" shows the lowest possible expected size for the corresponding parameter value

Truth		Average size						Coverage proportion				
Ω_m	H_0	BND	CSQ	MLE	SCR	MES	MR	CSQ	MLE	SCR	MES	MR
0.150	86.000	0.194	1.500	0.328	0.318	0.314	**0.296**	0.948	0.930	0.952	0.958	0.948
0.200	70.000	0.181	1.559	0.304	**0.297**	0.376	0.322	0.948	0.942	0.952	0.955	0.959
0.300	62.000	0.192	1.192	0.299	**0.293**	0.300	0.353	0.948	0.929	0.952	0.952	0.978
0.350	75.000	0.268	1.745	0.406	0.408	**0.371**	0.384	0.948	0.940	0.952	0.945	0.958
0.450	67.000	0.272	1.827	0.424	0.425	**0.365**	0.396	0.948	0.923	0.952	0.950	0.952

that the maximum risk of the numerical procedure converges almost surely to the Γ-minimax risk as the size of the simulations grows. In a two-dimensional application in cosmology, the minimax expected size and minimax regret confidence procedures give results comparable to classical confidence sets based on the score test, and are much smaller than chi-squared confidence regions. This suggests that MES and MR will be especially valuable in applications where the theory that links parameters and data is complex: in such problems, only chi-squared regions have generally been considered to be computationally tractable.

A parallel Fortran-90 implementation of the algorithm is available at *http://www.stat.cmu.edu/~cschafer/LFA_Search*.

APPENDIX: PROOFS

A.1 Proof of Theorem 2

In this appendix, m indexes the Monte Carlo simulations: the number of simulated null values of θ at stage m is q_m and the number of datasets simulated from each θ is n_m. We assume that n_m and q_m increase with m; in fact, we take $n_m = m$. We allow the level of the test to depend on m. At stage m, the level is α_m. We require $\alpha_m \to \alpha$.

Lemma 2 (van Zwet 1980). Suppose that $J, J_1, J_2, \ldots$ are uniformly bounded Lebesgue measurable functions from $[0,1]$ into $\mathbb{R}$, such that for all $t \in (0,1)$,

$$\lim_{m\to\infty} \int_0^t J_m(u)\,du = \int_0^t J(u)\,du.$$

Let $U_1, U_2, \ldots$ be a sequence of independent $U[0,1]$ random variables.

Define $U_{1:m}, U_{2:m}, \ldots, U_{m:m}$ to be $U_1, U_2, \ldots, U_m$ in increasing order. Let $g:[0,1]\to\mathbb{R}$ be a Borel measurable, integrable function and define

$$g_m(t) \equiv g\bigl(U_{\lfloor mt\rfloor+1:m}\bigr).$$

Then,

$$\int_0^1 J_m(u)g_m(u)\,du \xrightarrow{\text{a.s.}} \int_0^1 J(u)g(u)\,du.$$

Lemma 3. Fix $\eta\in\Theta$ and π. Define

$$\bar{K} \equiv \left[K \times \left(\frac{1}{m}\sum_{v=1}^{p}\sum_{k} r_{\delta_v}(\eta, X_k)\right)^{-1}\right] \wedge 1. \tag{A.1}$$

Then

$$\begin{aligned} Z_{m,\pi}(\eta) \;&\equiv\; \inf_{d\in\mathcal{D}'_{\alpha_m}} \frac{1}{m}\sum_{k=1}^{m} r_\pi(\eta, X_k)d(\eta, X_k)\bar{K} \\ &\xrightarrow{\text{a.s.}} \inf_{d\in\mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta,\eta)\pi(d\theta). \end{aligned}$$

Proof. We will apply Lemma 2 with $J_m(u)$ equal to one for $u \le 1-\alpha_m$ and zero otherwise; $J(u)$ is equal to one for $u \le 1-\alpha$ and zero otherwise. Let R denote the cdf of $r_\pi(\eta, X)$ when $X\sim\mathbb{P}_\eta$, that is, $R(t) = \mathbb{P}_\eta(r_\pi(\eta, X)\le t)$. The function $g(\cdot)$ of Lemma 2 is $g(u) = \inf\{t : R(t)\ge u\}$. Thus, if $U\sim U[0,1]$, $g(U)$ is a random variable with cdf $R(\cdot)$. We know $g(\cdot)$ is integrable since

$$\int_0^1 |g(u)|\,du = \mathbb{E}\bigl(|g(U)|\bigr) = \mathbb{E}_\eta(r_\pi(\eta, X)) = 1.$$

Define $u' = \inf\{u : g(u) = g(1-\alpha)\}$, $a = g(1-\alpha)$, and

$$c = \begin{cases} \dfrac{1-\alpha-u'}{\mathbb{P}_\eta(r_\pi(\eta, X) = a)}, & \mathbb{P}_\eta(r_\pi(\eta, X)=a) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} \int_0^1 J(u)g(u)\,du &= \int_0^{u'} g(u)\,du + \int_{u'}^{1-\alpha} g(u)\,du \\ &= \mathbb{E}\bigl(g(U)\mathbf{1}_{\{U<u'\}}\bigr) + \mathbb{E}\bigl(g(U)\mathbf{1}_{\{u'\le U\le 1-\alpha\}}\bigr) \\ &= \mathbb{E}\bigl(g(U)\mathbf{1}_{\{g(U)<g(u')\}}\bigr) + a(1-\alpha-u') \qquad \text{(A.2)} \\ &= \mathbb{E}\bigl(g(U)\mathbf{1}_{\{g(U)<a\}}\bigr) + a(1-\alpha-u') \\ &= \mathbb{E}_\eta\bigl(r_\pi(\eta, X)\mathbf{1}_{\{r_\pi(\eta,X)<a\}}\bigr) \\ &\quad + c\mathbb{E}_\eta\bigl(r_\pi(\eta, X)\mathbf{1}_{\{r_\pi(\eta,X)=a\}}\bigr) \\ &= \int_{\mathcal{X}} r_\pi(\eta, x)d^*(\eta, x)\mathbb{P}_\eta(dx) \\ &= \inf_{d\in\mathcal{D}_\alpha}\int_{\mathcal{X}} r_\pi(\eta, x)d(\eta, x)\mathbb{P}_\eta(dx) \qquad \text{(A.3)} \\ &= \inf_{d\in\mathcal{D}_\alpha}\int_\Theta \gamma_d(\theta,\eta)\pi(d\theta), \qquad \text{(A.4)} \end{aligned}$$

where

$$d^*(\eta, x) = \begin{cases} 1, & r_\pi(\eta, x) < a \\ c, & r_\pi(\eta, x) = a \\ 0, & \text{otherwise.} \end{cases}$$

Equation (A.2) holds because $g(U) < g(u')$ if and only if $U < u'$; Equation (A.3) holds because $d^*\in\mathcal{D}_\alpha$.

Consider the function $U:\mathcal{X}\times[0,1]\to[0,1]$ defined by

$$U(x,w) = \mathbb{P}_\eta(r_\pi(\eta, X) < r_\pi(\eta, x)) + w\mathbb{P}_\eta(r_\pi(\eta, X) = r_\pi(\eta, x)),$$

where $X\sim\mathbb{P}_\eta$. If $\{W_j\}_{j=1}^\infty$ are independent $U[0,1]$ random variables, $\{X_j\}_{j=1}^\infty$ are independent random variables distributed as $\mathbb{P}_\eta$, and $\{W_j\}$ and $\{X_j\}$ are independent, then $U_1 \equiv U(X_1, W_1)$, $U_2 \equiv U(X_2, W_2), \ldots$ are independent $U[0,1]$ random variables. Moreover,

$$\begin{aligned} g(U_i) &= \inf\{x : R(x)\ge U_i\} \\ &= \inf\{x : R(x)\ge U(X_i, W_i)\} \\ &= \inf\bigl\{x : \mathbb{P}_\eta(r_\pi(\eta, X)\le x)\ge U(X_i, W_i)\bigr\} \\ &= r_\pi(\eta, X_i). \end{aligned}$$

Let $X_{1:m}, X_{2:m}, \ldots, X_{m:m}$ denote $X_1, X_2, \ldots, X_m$ ordered by the (increasing) value of $r_\pi(\eta, X_i)$, with ties broken arbitrarily. Likewise, let $U_{1:m}, U_{2:m}, \ldots, U_{m:m}$ denote $U_1, U_2, \ldots, U_m$ in increasing order. Note that $U(x_1, w_1) < U(x_2, w_2)$ if and only if either $r_\pi(\eta, x_1) < r_\pi(\eta, x_2)$ or $r_\pi(\eta, x_1) = r_\pi(\eta, x_2)$ and $w_1 < w_2$. So, $g(U_{i:m}) = r_\pi(\eta, X_{i:m})$.

Thus,

$$\begin{aligned} \int_0^1 J_m(u)g_m(u)\,du &= \int_0^{1-\alpha_m} g_m(u)\,du \\ &= \frac{1}{m}\sum_{k=1}^{m} g(U_{k:m})d^*(\eta, k) \\ &= \frac{1}{m}\sum_{k=1}^{m} r_\pi(\eta, X_{k:m})d^*(\eta, k) \\ &= \inf_{d\in\mathcal{D}'_{\alpha_m}} \frac{1}{m}\sum_{k=1}^{m} r_\pi(\eta, X_k)d(\eta, X_k), \qquad \text{(A.5)} \end{aligned}$$

where

$$d^*(\eta, k) = \begin{cases} 1, & k < k' \\ (1-\alpha_m)m - k' + 1, & k = k' \\ 0, & k > k', \end{cases}$$

and $k' = \lceil (1-\alpha_m)m \rceil$.

Lemma 2 together with Equations (A.4) and (A.5) show that

$$\inf_{d \in \mathcal{D}'_{\alpha_m}} \frac{1}{m} \sum_{k=1}^{m} r_\pi(\eta, X_k) d(\eta, X_k) \xrightarrow{\text{a.s.}} \inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta) \pi(d\theta).$$

By the law of large numbers, $\bar{K} \to 1$ almost surely since

$$\mathbb{E}\left[\sum_{v=1}^{p} r_{\delta_v}(\eta, X_k) \right] = p < K. \tag{A.6}$$

Hence,

$$\begin{aligned} Z_{m,\pi}(\eta) &\equiv \inf_{d \in \mathcal{D}'_{\alpha_m}} \frac{1}{m} \sum_{k=1}^{m} r_\pi(\eta, X_k) d(\eta, X_k) \bar{K} \\ &\xrightarrow{\text{a.s.}} \inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta) \pi(d\theta). \end{aligned} \tag{A.7}$$

Lemma 4. As $m \to \infty$,

$$\mathbb{E}[Z_{m,\pi}(T_{jm})] \longrightarrow \mathbf{R}_\pi(d_\pi). \tag{A.8}$$

Proof. Apply the bounded convergence theorem twice to show that for fixed $\eta \in \Theta$

$$\mathbb{E}[Z_{m,\pi}(\eta)] \longrightarrow \inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta) \pi(d\theta) \tag{A.9}$$

and that

$$\int_\Theta \mathbb{E}[Z_{m,\pi}(\eta)] \nu(d\eta) \longrightarrow \int_\Theta \left[\inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta) \pi(d\theta) \right] \nu(d\eta). \tag{A.10}$$

But

$$\int_\Theta \mathbb{E}[Z_{m,\pi}(\eta)] \nu(d\eta) = \mathbb{E}[Z_{m,\pi}(T_{jm})] \tag{A.11}$$

and

$$\begin{aligned} &\int_\Theta \left[\inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta) \pi(d\theta) \right] \nu(d\eta) \\ &\quad = \inf_{d \in \mathcal{D}_\alpha} \int_\Theta \int_\Theta \gamma_d(\theta, \eta) \pi(d\theta) \nu(d\eta) \\ &\quad = \mathbf{R}_\pi(d_\pi). \end{aligned}$$

The infimum and integral can be switched because, as established in Lemma 1, the infimal d minimizes at each η.

Lemma 5. Suppose that $\{U_m\}_{m=1}^{\infty}$ is a sequence of random variables such that

$$U_m = \frac{1}{q_m} \sum_{j=1}^{q_m} V_{jm}, \tag{A.12}$$

where

1. $\{V_{jm}\}_{j=1}^{q_m}$ are iid for each m and independent across m;
2. $\mathbb{E}[V_{jm}] \equiv \mu_m \to \mu$;
3. $\{V_{jm}\}_{j=1}^{q_m}$ are nonnegative and uniformly bounded for all m; and
4. the sequence $\{q_m\}_{m=1}^{\infty}$ is strictly increasing.

Then $U_m \xrightarrow{\text{a.s.}} \mu$.

Proof. Fix $\epsilon > 0$. For m large enough that $|\mu_m - \mu| < \epsilon/2$,

$$\begin{aligned} \mathbb{P}[|U_m - \mu| > \epsilon] &\le \mathbb{P}[|U_m - \mu_m| > \epsilon/2] \\ &\le \left(\frac{16}{\epsilon^4} \right) \mathbb{E}[(U_m - \mu_m)^4], \end{aligned} \tag{A.13}$$

by Markov's inequality. Set $W_{jm} \equiv V_{jm} - \mu_m$.

$$\begin{aligned} \mathbb{E}[(U_m - \mu_m)^4] &= q_m^{-4} \mathbb{E}\left[\left(\sum_{j=1}^{q_m} W_{jm} \right)^4 \right] \\ &= q_m^{-4} \left(q_m \mathbb{E}[W_{1m}^4] + 3q_m(q_m - 1)\mathbb{E}[W_{1m}^2]^2 \right) \\ &\le c q_m^{-2} \le c m^{-2}, \end{aligned}$$

where the constant c does not depend on m. See the proof of theorem 6.1 in Billingsley (1995). Hence, by Borel–Cantelli,

$$\mathbb{P}[|U_m - \mu| > \epsilon \text{ i.o.}] = 0. \tag{A.14}$$

This implies that $U_m \to \mu$ almost surely.

These results in combination imply that as $m \to \infty$,

$$\widehat{\mathbf{R}}_\pi(d_{\pi,m}) = \frac{1}{q_m} \sum_{j=1}^{q_m} Z_{m,\pi}(T_{jm}) \xrightarrow{\text{a.s.}} \mathbf{R}_\pi(d_\pi) \tag{A.15}$$

for any probability distribution π on $(\Theta, \mathcal{A})$.

Lemma 6. Let $\pi, \pi' \in \Gamma$ such that $\pi = \sum_v w_v \delta_v$ and $\pi' = \sum_v w'_v \delta_v$. Then, for all m,

$$|\widehat{\mathbf{R}}_\pi(d_{\pi,m}) - \widehat{\mathbf{R}}_{\pi'}(d_{\pi',m})| \le K \|w - w'\|_1. \tag{A.16}$$

Proof. For fixed indices j and m, let d' be the decision procedure $d \in \mathcal{D}'_{\alpha_m}$ that minimizes the smaller of

$$\sum_k r_\pi(T_{jm}, X_{jkm}) d(T_{jm}, X_{jkm}) \tag{A.17}$$

and

$$\sum_k r_{\pi'}(T_{jm}, X_{jkm}) d(T_{jm}, X_{jkm}). \tag{A.18}$$

Then d' is either $d_{\pi,m}$ or $d_{\pi',m}$. Thus,

$$\begin{aligned} &|Z_{m,\pi}(T_{jm}) - Z_{m,\pi'}(T_{jm})| \\ &\quad \le \sum_v |w_v - w'_v| \left(\frac{1}{m} \sum_k r_{\delta_v}(T_{jm}, X_{jkm}) d'(T_{jm}, X_{jkm}) K_{jm} \right) \\ &\quad \le K \sum_{v=1}^{p} |w_v - w'_v| \\ &\quad = K \|w - w'\|_1. \end{aligned}$$

Since

$$\widehat{\mathbf{R}}_\pi(d_{\pi,m}) = \frac{1}{q_m} \sum_{j=1}^{q_m} Z_{m,\pi}(T_{jm}), \tag{A.19}$$

we have the desired result.

Lemma 6 implies that $\{\widehat{\mathbf{R}}_\pi(d_{\pi,m})\}_{m=1}^{\infty}$ is an equicontinuous family of functions of the weight vector w associated with π. The space of possible weights is compact, so the pointwise convergence for fixed π yields uniform convergence in π. (See Royden 1988, page 168, lemma 39.) This completes the proof of Theorem 2.

A.2 Proof of Theorem 6

It follows from the definition of $\mathcal{Z}(\theta)$ that there exists some $\mathbf{w} \in \mathcal{W}$ such that for any $\epsilon > 0$, $f_\theta(x) \le (\mathcal{Z}(\theta) + \epsilon)(\sum_{v=1}^{p} w_v f_{\delta_v}(x))$ for all $x \in \mathcal{X}$. Hence,

$$\begin{aligned}
\mathbf{R}(\theta, d) &= \int_\Theta \int_{\mathcal{X}} d(\eta, x) f_\theta(x) \mu(dx) \nu(d\eta) \\
&\le \int_\Theta \int_{\mathcal{X}} d(\eta, x)(\mathcal{Z}(\theta) + \epsilon) \left[\sum_{v=1}^{p} w_v f_{\delta_v}(x) \right] \mu(dx) \nu(d\eta) \\
&= (\mathcal{Z}(\theta) + \epsilon) \sum_{v=1}^{p} w_v \mathbf{R}_{\delta_v}(d) \\
&\le (\mathcal{Z}(\theta) + \epsilon) \mathbf{R}_\Gamma(d).
\end{aligned}$$

Since this is true for all ϵ, it follows that $\mathbf{R}(\theta, d) \le \mathcal{Z}(\theta) \mathbf{R}_\Gamma(d)$.

[Received August 2007. Revised September 2008.]

REFERENCES

Aitchison, J. (1966), "Expected-Cover and Linear-Utility Tolerance Intervals," *Journal of the Royal Statistical Society*, Ser. B, 28, 57–62.

Aitchison, J., and Dunsmore, I. (1968), "Linear-Loss Interval Estimation of Location and Scale Parameters," *Biometrika*, 55, 141–148.

Berkovitz, L. (2002), *Convexity and Optimization in R^n*, New York: Wiley.

Billingsley, P. (1995), *Probability and Measure*, New York: Wiley.

Brown, G. (1951), "Iterative Solution of Games by Fictitous Play," in *Activity Analysis of Production and Allocation*, ed. T. Koopmans, New York: Wiley, Chap. 24.

Bryan, B., McMahan, H., Schafer, C., and Schneider, J. (2007), "Efficiently Computing Minimax Expected Size Confidence Regions," in *Proceedings of the 24th International Conference on Machine Learning*.

Casella, G., and Hwang, J. (1991), "Evaluating Confidence Sets Using Loss Functions," *Statistica Sinica*, 1, 159–173.

Casella, G., Hwang, J., and Robert, C. (1994), "Loss Functions for Set Estimation," in *Statistical Decision Theory and Related Topics V*, eds. S. Gupta and J. Berger, New York: Springer-Verlag, pp. 237–251.

Cohen, A., and Strawderman, W. (1973a), "Admissibility Implications for Different Criteria in Confidence Estimation," *Annals of Statistics*, 1, 363–366.

——— (1973b), "Admissible Confidence Interval and Point Estimation for Translation or Scale Parameters," *Annals of Statistics*, 1, 545–550.

DeGroot, M. (1988), "Regret," in *Encyclopedia of Statistical Science*, Vol. 8, eds. S. Kotz, N. Johnson, and C. Read, New York: Wiley, pp. 3–4.

Evans, S., Hansen, B., and Stark, P. (2005), "Minimax Expected Measure Confidence Sets for Restricted Location Parameters," *Bernoulli*, 11, 571–590.

Hooper, P. (1982), "Invariant Confidence Sets With Smallest Expected Measure," *Annals of Statistics*, 10, 1283–1294.

Hwang, J., and Casella, G. (1982), "Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution," *Annals of Statistics*, 10, 868–881.

Joshi, V. (1969), "Admissibility of the Usual Confidence Sets for the Mean of a Univariate or Bivariate Normal Population," *Annals of Mathematical Statistics*, 40, 1042–1067.

Kempthorne, P. (1987), "Numerical Specification of Discrete Least Favorable Prior Distributions," *SIAM Journal on Scientific and Statistical Computing*, 8, 171–184.

Lehmann, E., and Romano, J. (2005), *Testing Statistical Hypotheses* (3rd ed.), New York: Springer.

Nelson, W. (1966), "Minimax Solution of Statistical Decision Problems by Iteration," *Annals of Mathematical Statistics*, 37, 1643–1657.

Pratt, J. (1961), "Length of Confidence Intervals," *Journal of the American Statistical Association*, 56, 549–567.

Riess, A. G., Strolger, L.-G., Casertano, S., Ferguson, H. C., Mobasher, B., Gold, B., Challis, P. J., Filippenko, A. V., Jha, S., Li, W., Tonry, J., Foley, R., Kirshner, R. P., Dickinson, M., MacDonald, E., Eisenstein, D., Livio, M., Younger, J., Xu, C., Dahlén, T., and Stern, D. (2007), "New Hubble Space Telescope Discoveries of Type Ia Supernovae at $z \ge 1$: Narrowing Constraints on the Early Behavior of Dark Energy," *Astrophysical Journal*, 659, 98–121.

Robinson, J. (1951), "An Iterative Method for Solving a Game," *Annals of Mathematics*, 54, 296–301.

Royden, H. (1988), *Real Analysis*, New York: Macmillan.

Schafer, C., and Stark, P. (2003), "Using What We Know: Inference With Physical Constraints," in *PHYSTAT2003: Statistical Problems in Particle Physics, Astrophysics and Cosmology*, eds. L. Lyons, R. Mount, and R. Reitmeyer, Stanford, CA: SLAC.

Spergel, D., Bean, R., Dore, O., and Nolta, M. (2007), "Three-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Implications for Cosmology," *Astrophysical Journal Supplement Series*, 170, 377–408.

van Zwet, W. (1980), "A Strong Law for Linear Functions of Order Statistics," *Annals of Probability*, 8, 986–990.

Winkler, R. (1972), "A Decision-Theoretic Approach to Interval Estimation," *Journal of the American Statistical Association*, 67, 187–191.

Wood-Vasey, W. M., Miknaitis, G., Stubbs, C. W., Jha, S., Riess, A. G., Garnavich, P. M., Kirshner, R. P., Aguilera, C., Becker, A. C., Blackman, J. W., Blondin, S., Challis, P., Clocchiatti, A., Conley, A., Covarrubias, R., Davis, T. M., Filippenko, A. V., Foley, R. J., Garg, A., Hicken, M., Krisciunas, K., Leibundgut, B., Li, W., Matheson, T., Miceli, A., Narayan, G., Pignata, G., Prieto, J. L., Rest, A., Salvo, M. E., Schmidt, B. P., Smith, R. C., Sollerman, J., Spyromilio, J., Tonry, J. L., Suntzeff, N. B., and Zenteno, A. (2007), "Observational Constraints on the Nature of Dark Energy: First Cosmological Results From the ESSENCE Supernova Survey," *Astrophysical Journal*, 666, 694–715.

Wright, E. (2007), "Constraints on Dark Energy From Supernovae, Gamma-Ray Bursts, Acoustic Oscillations, Nucleosynthesis, Large-Scale Structure, and the Hubble Constant," *Astrophysical Journal*, 664, 633–639.