Constructing Confidence Regions of Optimal Expected Size

Chad M. Schafer and Philip B. Stark*

June 18, 2007

Abstract

We present a Monte Carlo method for approximating the *minimax expected size* (MES) confidence set for a parameter known to belong to a compact set. Size refers to the measure of the confidence set; the measure can be indexed by the true parameter value, which allows the confidence procedure to be tailored for specific scientific goals. As the number of iterations increases, the Monte Carlo estimator converges to the Γ-minimax procedure, where Γ is a polytope of priors. The algorithm exploits Bayes/minimax duality by searching for the Γ-least favorable prior. A Fortran-90 implementation of the algorithm for both serial and parallel computers is available. We apply the method to estimate parameters of the primordial Universe from observations of the cosmic microwave background radiation.

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 et al. (2005) (hereafter, EHS) show that the $1 - \alpha$ confidence set with smallest maximum expected measure can potentially be found by inverting a family of level α tests of simple nulls versus a common simple alternative. This is the *minimax expected size* (MES) procedure. This paper gives a computationally efficient algorithm for computing MES and other optimal confidence sets, including the less conservative *minimax regret* (MR) procedure, when the parameter is known to lie in a compact set.

There have been several studies of loss functions for constructing 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) consider interval estimation of a real-valued parameter 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 et al. (1994) study confidence sets that are optimal with respect to such loss functions.

An alternative approach is to 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) show examples of using a measure as loss. Hooper (1982) and Cohen and Strawderman (1973a) allow the measure to depend on the true value of the parameter; we take the same approach. Here, "expected size" refers to the expected ν_θ-measure of the confidence set. The MES procedure minimizes the maximum of this risk function.

Exact determination of the MES procedure is not typically feasible. Our approximation algorithm uses the duality between Bayes and minimax procedures, established for confidence regions in some generality by EHS. The search for the Γ-minimax procedure becomes a search for the member of Γ with the largest average risk. This is called the *least favorable alternative* (LFA). Finding the LFA over a finite set is conceptually simple, but the risk calculations can be computationally intensive. Kempthorne (1987) and Nelson (1966) give algorithms to determine numerically the least favorable prior distribution over compact parameter spaces for general risk functions, but their algorithms assume the ability to calculate the Bayes risk for any given prior. In this work, risk is approximated using novel Monte Carlo simulations. We show that as the size of these Monte Carlo simulations increases, the maximum expected size of the confidence set converges to that of the unapproximated MES procedure. The algorithm is implemented as a Fortran-90 subroutine designed to run efficiently on distributed computers with little interprocessor communication.

The method is well suited for estimating parameters that satisfy *a priori* bounds, and when there is a complex model determining the distribution of the observed data as a function of these parameters. A common goal in the physical sciences is to estimate unknown physical constants accurately while efficiently utilizing the information provided by previous experimental and theoretical results. For example, there is a complex physical model for

the relationship between unknown cosmological parameters (e.g. Hubble's constant, the age of the Universe) and the angular distribution of fluctuations in the cosmic microwave background radiation (CMB). The model places useful restrictions on the class of power spectra. We seek to incorporate these bounds in the interest of "conserving" the rejection power of the associated hypothesis tests.

This paper is organized as follows. Section 2 gives notation, assumptions and theory. Section 3 describes the algorithm. Section 4 describes an implementation of the algorithm, including how results from the study of convex games are used to find the approximate LFA. Section 5 discusses selecting the vertices of Γ. Section 6 considers a general class of loss functions based on the measure of the confidence set, including one choice which leads to the minimax regret procedure. Section 7 shows results of application of the method to analysis of CMB data. Finally, Section 8 gives a summary, and Section 9 is supplemental material, including the technical proofs.

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 $\mathcal{B}$ on a 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 measurable, with σ-field $\mathcal{A}$. Let

$$\mathcal{V} \equiv \{\nu_\theta : \theta \in \Theta\}$$

be a family of positive measures on Θ, all dominated by the measure ν. The density of ν_θ with respect to ν is v_θ. The random variable X has distribution $\mathbb{P}_{\theta_0}$ for some unknown $\theta_0 \in \Theta$. We observe data X and $U \sim U[0,1]$, a uniform random variable independent of X. We have at our disposal a set $\mathcal{D}$ of *decision functions*, measurable mappings from $\Theta \times \mathcal{X}$ into $[0,1]$. The decision functions let us use X and U to make random subsets of Θ:

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

This set is a candidate confidence set for θ_0 from the data X and the auxiliary random variable U. The chance that $\mathbf{C}_d(X,U)$ covers the parameter value $\eta \in \Theta$ when in fact

3

$X \sim \mathbb{P}_\theta$ is

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

The decision rules that correspond to $1 - \alpha$ confidence sets are

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

We are about to define a risk function on confidence sets. Think of $v_\theta(\eta)$ as the cost of including η in a confidence set for θ_0 when in fact $\theta_0 = \theta$. Pratt (1961) showed the risk of using the decision rule d to make a confidence set for θ_0 is the expected ν_θ-measure of the confidence set, which is the expected integrated cost:

$$\mathbf{R}(\theta, d) \equiv \mathbb{E}_\theta[\nu_\theta(\mathbf{C}_d(X, U))] = \int_\Theta \gamma_d(\theta, \eta) v_\theta(\eta) \nu(d\eta). \qquad (4)$$

With $v_\theta(\eta) = 1$, this risk is the expected ν-measure of the confidence set. Allowing the cost $v_\theta(\eta)$ to depend on θ and to vary with η lets us tailor confidence sets to specific scientific goals. Although generally we might prefer confidence sets that are as small as possible, there are situations where we might willingly sacrifice size. For example, a confidence procedure for some real-valued parameter might be more likely than the shortest interval to contain only positive values when the effect is positive, and more likely than the shortest interval to contain only negative values when the effect is negative. In this hypothetical, sacrificing length allows better inferences about the sign of the parameter, a cost we might happily pay. Such a preference could be incorporated by choosing

$$v_\theta(\eta) = \begin{cases} 1, & \text{sgn}(\theta) = \text{sgn}(\eta) \\ c, & \text{sgn}(\theta) = -\text{sgn}(\eta), \end{cases} \qquad (5)$$

for some real $c > 1$. In other problems, we might not care whether η is included in the confidence set when θ is the true value of θ_0, for example, if η and θ differ only with respect to nuisance parameters; then we might set $v_\theta(\eta) = 0$.

Let $\mathbf{R}_\Theta(d)$ denote the maximum risk of d over all $\theta \in \Theta$:

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

where the supremum is over all distributions π on Θ. The problem we address is to find a numerical approximation to the decision rule $d_\mathbf{R}$ that attains the 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 typical applications, Γ will be the polytope formed by taking the vertices as p parameter values $\theta_1, \theta_2, \ldots, \theta_p$, spread across Θ to ensure that $\mathbf{R}_\Theta(d_\mathbf{R})$ is not too much larger than

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

Our numerical approximation produces a valid $1 - \alpha$ confidence set—a member of $\mathcal{D}_\alpha$—but its risk is approximately Γ-minimax, rather than exactly Γ-minimax.

2.1 Bayes-Minimax Duality

For any probability distribution π on Θ, define

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

This is a weight function that combines the density of the observations f_θ, the density v_θ of the measure on Θ that determines the risk, mixed across values of θ using the prior π.

The Bayes risk of d for prior π is

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

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

$$\int_\mathcal{X} d(\eta, x) f_\eta(x)\mu(dx) = 1 - \alpha. \tag{11}$$

The optimal decision rule $d_\pi \in \mathcal{D}_\alpha$ for prior π is found by minimizing (10) subject to (11), which can be done in exactly the same way that the Neyman-Pearson Lemma is proved: for each η, d_π is 1 where $f_\eta(x)r_\pi(\eta, x)/f_\eta(x) = r_\pi(\eta, x)$ is below some threshold, 0 where $r_\pi(\eta, x)$ is above the threshold, and constant at the threshold, with the threshold and the constant chosen to make $d_\pi(\eta.x)$ correspond to a level-α test.

Lemma 1.

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

where

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

5

and the constants $0 \leq b_{\eta,\alpha} \leq 1$ and $c_{\eta,\alpha}$ satisfy

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

If Γ is a collection of distributions on Θ, 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. *If Γ is convex and π_0 is Γ-least favorable,*

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

Proof. See Section 9.2. This is an extension of Theorem 1 in EHS. $\square$

2.2 Underlying Assumptions

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

1. $\nu(\Theta) < \infty$.

2. If $P_\theta \neq P_{\theta'}$, $\theta, \theta' \in \Theta$, 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 ν-a.e.

4. The penalty function v_θ is nonnegative and bounded, i.e. $0 \leq v_\theta(\eta) \leq M$ for all $\theta, \eta \in \Theta$.

5. The convex collection of priors Γ has a finite number of vertices.

The method is not practical unless:

1. For any fixed point θ in the parameter space Θ, it is computationally tractable to simulate sampling from P_θ.

2. If π_i is a vertex of Γ, calculating $r_{\pi_i}(\eta, x)$ for fixed η and x is tractable.

3 The Monte Carlo Approach

Perhaps remarkably, a single set of simulations can estimate d_π and $\mathbf{R}_\pi(d_\pi)$. Let T be an element of Θ drawn at random according to the distribution ν. Conditional on $T = \eta$, the random variable X has distribution $\mathbb{P}_\eta$. Recall from Lemma 1 that $r_\pi(\eta, X)$ is the test statistic for a test of the hypothesis $\theta_0 = \eta$. Given data x, the test rejects the hypothesis if

$$\mathbb{P}_\eta[r_\pi(\eta, X) \geq r_\pi(\eta, x)] \leq \alpha. \tag{16}$$

For any $d \in \mathcal{D}$,

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

Hence, for fixed π, Monte Carlo simulation of the distribution of $r_\pi(T, X)$ can be used to estimate simultaneously the thresholds for the Bayes decision rule and the Bayes risk of the Bayes decision. This leads to defining $\widehat{\mathbf{R}}_{\pi,m}(d_{\pi,m})$, a function of simulated $r_\pi(T, X)$ variates, where m indexes the size of the Monte Carlo simulations; see Equation (21) below. We shall show that as m increases, $\widehat{\mathbf{R}}_{\pi,m}(d_{\pi,m})$ converges almost surely to $\mathbf{R}_\pi(d_\pi)$, uniformly in $\pi \in \Gamma$.

Fix $\{n_m\}_{m=1}^\infty$ and $\{q_m\}_{m=1}^\infty$, two strictly increasing sequences of integers. Let

$$\{T_{jm} : j = 1, 2, \ldots, q_m; \ m = 1, 2, \ldots\} \tag{17}$$

be iid (ν) and let

$$\{X_{jkm} : j = 1, 2, \ldots, q_m; \ k = 1, 2, \ldots, n_m; \ m = 1, 2, \ldots\} \tag{18}$$

have distribution $\mathbb{P}_\eta$ conditional on $T_{jm} = \eta$. All X_{jkm} are independent, conditional on all of the T_{jm}. Define

$$K_{jm} \equiv \left[K \times \left(\frac{1}{n_m} \sum_{i=1}^p \sum_{k=1}^{n_m} r_{\pi_i}(T_{jm}, X_{jkm}) \right)^{-1} \right] \bigwedge 1, \tag{19}$$

with $K > pM$. (Recall that $v_\theta(\eta) \leq M$.) Let

$$\widehat{\mathbf{R}}_m(\theta, d) \equiv \frac{1}{q_m} \sum_{j=1}^{q_m} K_{jm} \left[\frac{1}{n_m} \sum_{k=1}^{n_m} \frac{f_\theta(X_{jkm})}{f_{T_{jm}}(X_{jkm})} v_\theta(T_{jm}) d(T_{jm}, X_{jkm}) \right] \tag{20}$$

and

$$\widehat{\mathbf{R}}_\pi(d) \equiv \int_\Theta \widehat{\mathbf{R}}_m(\theta, d) \, \pi(d\theta) = \frac{1}{n_m q_m} \sum_j \sum_k r_\pi(T_{jm}, X_{jkm}) \, d(T_{jm}, X_{jkm}) \, K_{jm}, \qquad (21)$$

continuing the definition of r_π given in Equation (9). Multiplying by K_{jm} forces $\widehat{\mathbf{R}}_\pi(d)$ to be uniformly bounded while maintaining a linear relationship between $\widehat{\mathbf{R}}_\pi(d)$ and the bracketed function on the right hand side of Equation (20).

Fix α and define $\mathcal{D}'_\alpha$ to be the class of decision procedures that for all j satisfy

$$\frac{1}{n_m} \sum_k d(T_{jm}, X_{jkm}) \geq 1 - \alpha. \qquad (22)$$

Let $d_{\pi, m}$ be the decision procedure that minimizes $\widehat{\mathbf{R}}_\pi(d)$ among all $d \in \mathcal{D}'_{\alpha_m}$. Recall that d_π is the decision procedure that minimizes $\mathbf{R}_\pi(d)$ over all $d \in \mathcal{D}_\alpha$.

Theorem 2. *As $m \to \infty$,*

$$\widehat{\mathbf{R}}_\pi(d_{\pi, m}) \xrightarrow{a.s.} \mathbf{R}_\pi(d_\pi) \qquad (23)$$

uniformly in $\pi \in \Gamma$.

Proof. See Section 9.2. □

Corollary 1. *As $m \to \infty$,*

$$\sup_{\pi \in \Gamma} \widehat{\mathbf{R}}_\pi(d_{\pi, m}) \xrightarrow{a.s.} \mathbf{R}_\Gamma(d_{\mathbf{R}}). \qquad (24)$$

With simulated realizations of the random quantities (17) and (18), a member of Γ that maximizes $\widehat{\mathbf{R}}_\pi(d_{\pi, m})$ can be found numerically. Corollary 1 shows that for large enough m this supremal prior has Bayes risk close to the Bayes risk of the Γ-least favorable prior.

4   Implementation of the Algorithm

At the jth stage, $1 \leq j \leq q$, draw a parameter value T at random according to ν, independently from all previous stages. Let η_j be the observed value of T. Then draw n data i.i.d. from $\mathbb{P}_{\eta_j}$. Let the observed data values be $\{x_{jk}\}_{k=1}^n$. Calculate the constant

$$\widetilde{K}_j \equiv \left[K \times \left(\frac{1}{n} \sum_{i=1}^p \sum_{k=1}^n r_{\pi_i}(\eta_j, x_{jk}) \right)^{-1} \right] \bigwedge 1. \qquad (25)$$

Find the n by p matrix $\mathbf{A_j}$ with elements

$$\mathbf{A_j}_{\,ki} = r_{\pi_i}(\eta_j, x_{jk}) \widetilde{K}_j. \qquad (26)$$

A test of the hypothesis that $\theta_0 = \eta_j$ can be represented by an n-vector $\mathbf{d_j}$. The kth component of $\mathbf{d_j}$ is the probability of not rejecting the hypothesis $\theta_0 = \eta_j$ if the observed datum is $X = x_{jk}$. The collection $\{d_j\}_{j=1}^q$ comprise an approximate decision rule that can be used to make an approximate confidence set for θ_0.

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

$$\pi = \sum_{i=1}^p w_i \pi_i, \tag{27}$$

for some $\mathbf{w} = (w_i)_{i=1}^p$ with $w_i \geq 0$ and $\sum_i w_i = 1$. The quantity

$$\widetilde{\mathbf{R}}_\pi(d) \equiv \frac{1}{nq} \sum_{\mathbf{j}=1}^q \mathbf{d_j}^{\mathbf{T}} \cdot \mathbf{A_j} \cdot \mathbf{w} \tag{28}$$

is an empirical approximation to the risk of the decision function represented by $\{\mathbf{d_j}\}_{j=1}^q$ for prior π.

For fixed $\pi \in \Gamma$ let $\widetilde{d}_\pi$ be the collection $\{\mathbf{d_j}\}_{j=1}^q$ that minimizes $\widetilde{\mathbf{R}}_\pi(d)$. Our goal is to find the (empirically) Γ-least favorable prior, the $\pi \in \Gamma$ that maximizes $\widetilde{\mathbf{R}}_\pi(\widetilde{d}_\pi)$. We shall see that this problem can be couched as a convex game. Theorem 2 shows that solving this convex game gives an arbitrarily good approximation to the theoretical Γ-minimax problem as the size of the simulations increases.

4.1 Matrix Games and Minimax Procedures

4.1.1 Solving Matrix Games

A *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 one chooses a *strategy*, an element $\mathbf{s_1}$ from $\mathcal{S}_1$. Player 2 picks a strategy $\mathbf{s_2}$ from $\mathcal{S}_2$. Player one pays player two $\mathbf{s_1}^{\mathbf{T}} \mathbf{A} \mathbf{s_2}$.

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* $(\mathbf{s_1}, \mathbf{s_2}) \in \mathcal{S}_1 \times \mathcal{S}_2$,

$$\mathbf{s_{1*}}^{\mathbf{T}} \mathbf{A} \mathbf{s_2} \leq \mathbf{s_{1*}}^{\mathbf{T}} \mathbf{A} \mathbf{s_{2*}} \leq \mathbf{s_1}^{\mathbf{T}} \mathbf{A} \mathbf{s_{2*}}. \tag{29}$$

Proof. This is a direct consequence of the classic Von Neumann Minimax Theorem. See, for example, Theorem 5.2 in Berkovitz (2002). $\qquad\square$

The pair $(\mathbf{s_{1*}}, \mathbf{s_{2*}})$ has a special optimality: By picking $\mathbf{s_{1*}}$, Player one minimizes his maximum loss. By picking $\mathbf{s_{2*}}$, Player two maximizes his minimum gain. *Solving the game*

is finding this saddle point. The *Brown-Robinson fictitious play algorithm* ((Robinson, 1951; Brown, 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 one finds the strategy $\mathbf{s_1} \in \mathcal{S}_1$ that minimizes $v_{1,i} \equiv \mathbf{s_1}^{\mathbf{T}}\mathbf{As_{2,i-1}}$.

2. Player two finds the strategy $\mathbf{s_2} \in \mathcal{S}_2$ that maximizes $v_{2,i} \equiv \mathbf{s_{1,i-1}}^{\mathbf{T}}\mathbf{As_2}$.

3. If $v_{2,i} - v_{1,i} \leq \epsilon$, we are done. Otherwise, go to step four.

4. Set

$$\mathbf{s_{1,i}} \equiv \left(\mathbf{s_1} + (i-1)\mathbf{s_{1,i-1}}\right)/i \tag{30}$$

   and

$$\mathbf{s_{2,i}} \equiv \left(\mathbf{s_2} + (i-1)\mathbf{s_{2,i-1}}\right)/i. \tag{31}$$

5. Increment i and return to step one.

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

$$v_{1i} \leq \mathbf{s_{1*}}^{\mathbf{T}}\mathbf{As_{2*}} \leq v_{2i} \tag{32}$$

and

$$\lim_{i \to \infty} v_{2i} - v_{1i} = 0. \tag{33}$$

Theorem 5. *If player one uses strategy $\mathbf{s_{1,i}}$, the amount player one pays player two is less than*

$$\mathbf{s_{1*}}^{\mathbf{T}}\mathbf{As_{2*}} + v_{2,i+1} - v_{1,i+1} \tag{34}$$

no matter what strategy player two uses.

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

$$\mathbf{s_{1,i}}^{\mathbf{T}}\mathbf{As} \leq v_{2,i+1} \leq \mathbf{s_{1*}}^{\mathbf{T}}\mathbf{As_{2*}} + v_{2,i+1} - v_{1,i+1}, \tag{35}$$

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

Theorem 5 ensures that when the Brown-Robinson algorithm terminates, player one has a strategy that limits his maximum loss to at most ϵ more than the loss at the saddle point. While the maximum loss is close to optimal, the theorem does not show that the strategy $\mathbf{s_{1i}}$ is close to $\mathbf{s_{1*}}$ in the norm.

4.1.2 Finding the Approximate LFA by Solving a Matrix Game

We now show that the problem of finding the LFA can be written as a (large) convex game. Define the nq by p matrix

$$\mathbf{A} \equiv \frac{1}{nq} \begin{bmatrix} \mathbf{A_1} \\ \mathbf{A_2} \\ \dots \\ \mathbf{A_q} \end{bmatrix}. \tag{36}$$

Define the nq-vector

$$\mathbf{d} \equiv \begin{bmatrix} \mathbf{d_1} \\ \mathbf{d_2} \\ \dots \\ \mathbf{d_q} \end{bmatrix}. \tag{37}$$

Equation (21) can be written

$$\widetilde{\mathbf{R}}_\pi(d) = \mathbf{d^T A w}. \tag{38}$$

Player one is the statistician. He chooses the $100(1 - \alpha)\%$ confidence procedure d. Player two is an intelligent adversary ("Nature"). She chooses $\mathbf{w}$, corresponding to a distribution π over the possible values of θ_0. Player one's set of possible strategies $\mathcal{S}_1$ takes a special form in this case. 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 $\mathcal{S}_2$ consists of all p-vectors $\mathbf{w}$ with $w_i \geq 0$ and $\sum_i w_i = 1$; this is also a convex set.

The statistician and Nature play the convex game $(\mathbf{A_m}, \mathcal{S}_1, \mathcal{S}_2)$. 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., solving the game by linear programming) might take fewer iterations, but are difficult to implement when $\mathcal{S}_1$ is complex. Recent work by Bryan et al. (2007) shows

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

Table 1: *The major storage demands of 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 data generated from each random parameter is n.*

how the sparsity of the payoff matrix in this case can be exploited to find solutions to this convex game in significantly less time.

4.2 Implementation

The approach parallelizes naturally: different processors can simulate independent samples of parameter values $\{\eta_j\}$ and data $\{x_{jk}\}$. Interprocessor communication is required only to calculate the outer sum in Equation (21), which involves $\{\widetilde{\mathbf{R}}_{\pi_i}(d_\pi)\}_{i=1}^p$.

A Fortran-90 implementation of the algorithm with documentation is available at:

http://www.stat.cmu.edu/$\sim$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

$$\{\{\{r_{\pi_i}(T_j, X_{jk})\}_{j=1}^q\}_{k=1}^n\}_{i=1}^p. \tag{39}$$

One might instead store the randomly simulated data; but this would be a $[n \times q \times (\text{dimension of } \mathcal{X})]$ array, and then the quantities $\{r_{\pi_i}\}_{i=1}^p$ would need to be calculated repeatedly. The operation count of the algorithm is $O(q \times n^2 \times p)$, from calculating $\widetilde{\mathbf{R}}_\pi(d_\pi)$. This neglects the number of operations involved in calculating the likelihoods $f_\eta(x)$.

5 Choosing the Vertices of Γ

The choice of Γ (the choice of the vertices $\{\pi_i\}_{i=1}^p$ of Γ) is Bayesian in flavor. In fact, the idea of Γ-minimax comes from research in robust Bayesian methods; see Vidakovic (2000),

for example. Chamberlain (2000) gives examples of applications of Γ-minimax estimators in econometrics. This procedure presents an interesting mixture of Bayesian and frequentist ideas: The $1 - \alpha$ coverage probability requirement holds regardless of the true value of the parameter, but it is optimized to control the Bayes risk for priors that are in Γ. The user decides how broad a collection of possible truths Γ will cover. With MES, $\mathbf{R}(\theta, d_{\mathbf{R}})$ is guaranteed to be less than or equal to $\mathbf{R}_{\Gamma}(d_{\mathbf{R}})$, the Γ-minimax expected size, only if the distribution P_{θ} is a member of Γ.

5.1 Nelson's Approach

One option would be to use the provided minimax algorithm in conjunction with the following larger iterative algorithm to determine the least favorable alternative, as proposed by Nelson (1966). In this procedure, the number of vertices in Γ increases over a series of iterations. Ideally, at each step the new vertices would be in places where the risk was largest when calculated using the minimax procedure of the previous iteration. When it is no longer possible to find a candidate vertex that has larger risk than the current Bayes risk, then one knows that the least favorable alternative has been found. Nelson proved that with exact risk calculations this procedure converges to the least favorable of all possible priors, and hence leads to the global minimax decision procedure. The feasibility of the algorithm depends on being able to find parameter values (vertices) where the risk is maximal.

5.2 Incorporating Known Restrictions on Γ

In some cases it is clear that the LFA must belong to a subclass of priors. If this subclass is convex, each vertex of Γ should also be a member. Consider the example of estimating the mean of a normal random variable when that mean is known to lie in $[-\tau, \tau]$. The LFA is clearly symmetric about zero; note that the class of priors with that property is convex. Each vertex of Γ should be a prior that puts equal weight at $-b$ and b, where $b \leq \tau$.

6   General Loss Functions

The theory developed here applies equally to a loss function of the form $\nu_\theta(\mathbf{C}_d(x,u)) - \ell(\theta)$, where ℓ is any uniformly bounded function on Θ. A choice of particular interest is

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

Finding the $d \in \mathcal{D}$ which minimizes the maximum expectation of this loss will be the *minimax regret* (MR) procedure. The regret (DeGroot, 1988) from using procedure d is the difference between the expected ν-measure of the confidence set while using d and minimum possible expected size over all $d \in \mathcal{D}_\alpha$. In some inference problems, parameter values θ for which $\ell_r(\theta)$ are relatively large can have a significant effect on the MES procedure: The least favorable alternative will place a large amount of weight on these θ at the expense of increasing the expected size under other parameter values.

Consider the following extreme example; more details can be found in Schafer and Stark (2003) and EHS. Let the random variable X have the normal distribution with mean θ and variance one; assume it is known that $-3 \leq \theta \leq 3$. The minimax expected length 95% confidence interval is defined by the LFA which places probability one at $\theta = 0$. Thus, the expected length when $\theta = 0$ is made as small as possible (it equals $\ell_r(0)$) while ignoring the expected length of the interval for all other values of θ. The MR procedure provides a less conservative alternative. See Figure 1. The solid line is $\ell_r(\theta)$, the dashed-dotted line is the expected length using the MES procedure; note that they are equal at $\theta = 0$. The expected length when using the MR procedure is given by the dashed line. Average length increases near zero, but significant gains are made away from zero. Finally, the dotted line is the expected length if one were to utilize the standard interval $(X - 1.96, X + 1.96)$ intersected with $[-3, 3]$.

Calculating $\ell_r(\theta)$ for fixed θ is theoretically simple; it requires another application of the Neyman-Pearson Lemma. In practice, however, $\ell_r(\theta)$ will be a complicated function of θ. The algorithm described in Section 4 allows one to approximate $\ell_r(\theta)$ by using the prior which places all of the weight on θ. Note that in order for this to work, each vertex π_i must place all of its weight on one point in the parameter space. The provided subroutine can also find the minimax regret procedure.
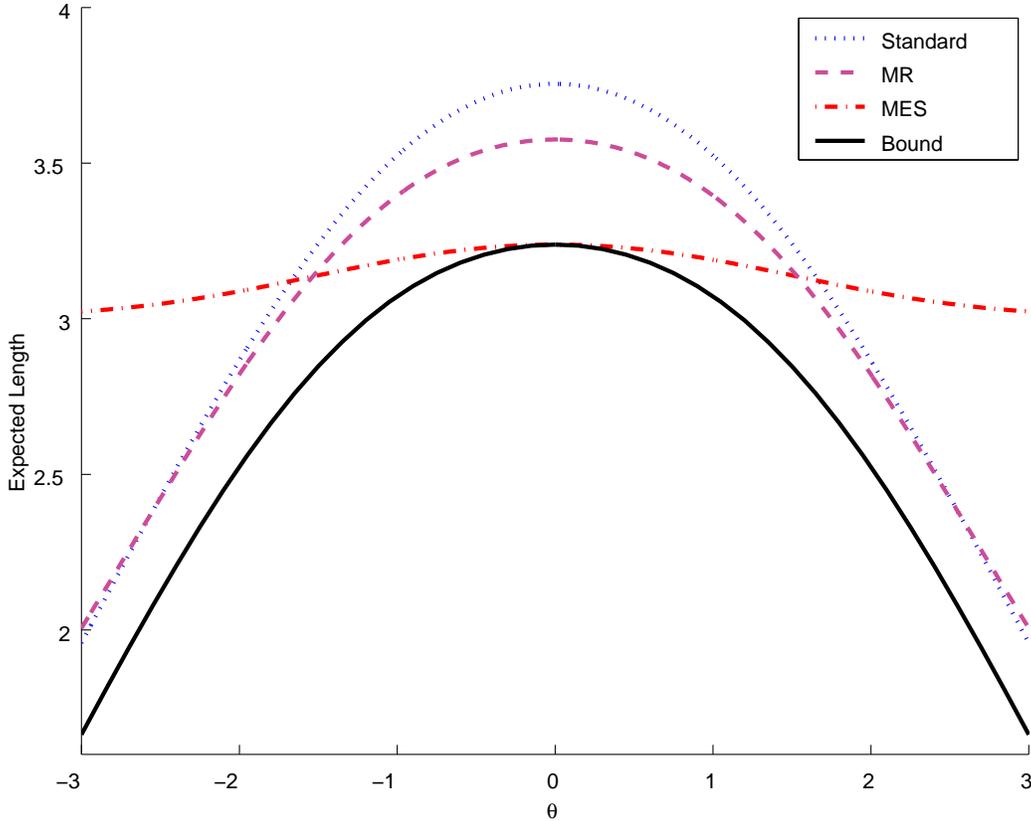
Figure 1: A comparison of 95% confidence intervals for the unknown mean θ when it is assumed that $-3 \le \theta \le 3$.

7   Example: Parameter Inference with CMB Data

Right after the Big Bang, the temperature of the Universe was too high for atoms to form: most matter existed as charged ions. Photons interact with charged particles, so the Universe was an opaque soup in which matter and photons exchanged energy freely. In about 400,000 years—at the *time of last scattering*—the Universe had cooled enough that (electrically neutral) atoms could form. Photons and matter interacted much less: the Universe became largely transparent. The photons that were freed at the time of last scattering have an imprint of the structure of the primordial Universe: according to theory, the chance that a photon liberated then has interacted with matter since is extremely small. (Some theorists have proposed that the Universe re-ionized at some later time.)

The photons that were freed the time of last scattering have cooled to about $2.7K$ as a result of the expansion of the Universe. They are observable as an extremely faint, nearly

isotropic microwave signal, the *cosmic microwave background radiation* (CMB).

Our night sky shows heterogeneity at many scales, including stars, galaxies and clusters of galaxies. That heterogeneity had seeds in the primordial Universe. Those seeds are evident in fluctuations in the CMB of about $200\mu K$. The observed anisotropy of the CMB can help to test theories of the origin and evolution of the Universe. For example, cosmological theories relate CMB anisotropy to physical parameters such as Hubble's constant (H_0), the density of baryonic matter density relative to the critical density (Ω_b), and the optical depth (τ).

The accuracy, resolution, and quantity of measurements of the CMB have increased enormously over the last 15 years or so. The COBE satellite (Smoot et al., 1990), the first experiment to detect CMB anisotropy, made about six thousand measurements at seven-degree angular resolution. More recently, WMAP (Bennett et al., 2003) made about two million measurements at ten arcminute angular resolution. The theoretical relationship between cosmological models and CMB observations is complicated. The data are noisy. The data sets are large. And there are physical constraints that might reduce the uncertainty. This is a situation where sophisticated statistical analysis might be particularly helpful.

According to many cosmological theories, CMB anisotropy is a realization of an isotropic, Gaussian process on the sphere. Such a process is fully characterized by its spherical harmonic power spectrum $\{C_\ell(\theta)\}_{\ell=1}^{\infty}$. Theory connects the power spectrum to a set of cosmological parameters $\theta = (H_0, \Omega_b, \tau, \ldots)$.

We observe only one realization of the process–our CMB. That limits the accuracy with which we can estimate C_ℓ. If we could measure the CMB perfectly over the entire sky, the maximum likelihood estimate of C_ℓ would be

$$\frac{1}{2\ell+1}\left(\sum_{m=-\ell}^{\ell}|a_{\ell m}|^2\right), \tag{40}$$

where $a_{\ell m}$ is the (ℓ, m) coefficient of the empirical spherical harmonic transform of the CMB. Noise, censoring, foreground contamination, instrumental smoothing, binning, etc., make the problem harder.

The physical model that links the cosmological parameters θ to the anisotropy spectrum $\{C_\ell\}$ constrains the class of possible spectra. The constraints can be exploited to improve estimates of the spectrum, and of the values of the cosmological parameters. Bayesian

Symbol	Description
Ω_b	The density of baryonic matter relative to the critical density
Ω_m	The total density of matter relative to the critical density
Ω_Λ	The density of dark energy relative to the critical density
H_0	The *Hubble Constant*, the current rate of expansion of the Universe
τ	The optical depth
A	The amplitude of the primordial density perturbations at frequency $0.05\mathrm{Mpc}^{-1}$
n_s	The scalar spectral index for the primordial density perturbations

Table 2: *The seven parameters in the ΛCDM model. See text for explanation.*

methods are appealing in problems with physical constraints, because the constraints can be imposed through the prior distribution for the parameters. The WMAP team analyzed their CMB data using Bayesian methods to incorporate bounds on the parameters (Verde et al., 2003). Their inferences depend on the prior they used, and their "confidence levels" need not relate to frequentist coverage probability.

Frequentist methods for dealing with constraints are not widely known. MES is one way to incorporate physical constraints from a frequentist perspective. MES avoids the possibility that the estimate is sensitive to the ad hoc choice of a prior to capture the constraints. MES confidence sets have the right frequentist level and are optimally precise if the physical model and constraints are correct.

7.1 The ΛCDM Cosmological Model

Following Spergel et al. (2003), we base our analysis on the *Power Law, Flat, ΛCDM Model*, or briefly, the ΛCDM model. In the ΛCDM model, the CMB spectrum $\{C_\ell\}$ depends on seven parameters, listed in Table 2. The density of the Universe is denoted by Ω. It has contributions from matter, Ω_m, and from "dark energy," Ω_Λ, which is related to Einstein's cosmological constant Λ. The contribution Ω_m of matter is the sum of a contribution Ω_b from ordinary baryonic matter, and—in the ΛCDM model—a contribution from cold dark matter.

The topology of the Universe is determined by Ω. At the critical density, $\Omega = 1$, the Universe is topologically flat (in four-dimensional space). If $\Omega > 1$, the Universe is *closed*. It will eventually contract, ending in a "big crunch." If $\Omega < 1$, the Universe is *open*, and will

continue to expand forever. Observations suggest that the Universe is nearly flat; *inflation* explains this as a consequence of very rapid expansion just after the Big Bang (Guth, 1981). The ΛCDM model assumes that the Universe is exactly flat: $\Omega_m + \Omega_\Lambda = 1$.

The Hubble Constant H_0 measures the rate of the expansion of the Universe; it has units $\text{kms}^{-1}\text{Mpc}^{-1}$. The dimensionless quantity $h \equiv H_0/(100\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1})$ appears frequently.

A photon released at the time of last scattering might have interacted with matter since then. The ΛCDM model represents this possibility using the optical depth parameter τ. The probability a photon has traveled since the time of last scattering without scattering is $\exp(-\tau)$.

The density of the primordial Universe varied spatially; that is why the modern Universe has structure. Because matter and radiation were tightly coupled until the time of last scattering, the modern CMB is the evolution of an imprint of the density fluctuations at that time. In the ΛCDM model, the spatial spectrum of primordial density fluctuations follows a power law: as a function of spatial frequency k, the spectrum is

$$A(k) = A\left(\frac{k}{k_0}\right)^{n_s-1}, \tag{41}$$

where $k_0 = 0.05\text{Mpc}^{-1}$ is a reference scale.

The angular spectrum of the CMB is related to the parameters of the ΛCDM model through

$$C_\ell(\theta) \propto \int g(\ell, k, \theta)^2 A(k)\frac{dk}{k}, \tag{42}$$

where g, the transfer function relating the two spectra, depends in a complicated way on the parameters in the model. See page 14 in Verde et al. (2003) for more detail.

7.2 WMAP Analysis

The starting point for our analysis is the estimate of the CMB power spectrum (for $2 \leq \ell \leq 900$) provided by the WMAP team (Hinshaw et al., 2003), denoted $\{\widehat{C}_\ell\}_{\ell=2}^{900}$. Although derived using a computationally-efficient approach, Hinshaw et al. (2003) show that the estimate is practically indistinguishable from the maximum likelihood estimate, and hence we will treat it as if it were the MLE. (See Appendix A in Hinshaw et al. (2003).) In another paper by the WMAP team (Verde et al., 2003), an approximation to the Fisher information matrix is derived as a function of the true power spectrum.

Next we apply the variance stabilizing transformation advocated by Bond et al. (2000). Our data vector $\mathbf{x}$ consists of the following transformed estimates of the CMB power spectrum:

$$x_{\ell-1} \equiv \log\left(\widehat{C}_\ell + N_\ell/B_\ell^2\right), \quad \ell = 2, 3, \ldots, 900. \tag{43}$$

Here, N_ℓ and B_ℓ represent contributions from measurement error and beam smearing, respectively, at wavenumber ℓ, and are taken as known quantities. ("Beam smearing" refers to the fact that measurements are actually the CMB field convolved with a kernel function.) Under an idealized model for the CMB, this transformation exactly stabilizes the asymptotic variance of the MLE.

Thus, in what follows we will appeal to standard asymptotic results for the MLE and assume that $\mathbf{x}$ is the realization of a multivariate Gaussian random vector with mean $\mu(\theta)$, where

$$\mu_{\ell-1}(\theta) \equiv \log\left(C_\ell(\theta) + N_\ell/B_\ell^2\right), \quad \ell = 2, 3, \ldots, 900, \tag{44}$$

and covariance matrix $\Sigma(\theta)$. Here, θ is the vector of cosmological parameters; because the variance stabilizing transformation is not exact, it is important to incorporate the dependence of the covariance matrix on θ.

As mentioned earlier, a strength of MES and MR is they handle restrictions on the parameter space in an optimal manner. This takes dual meaning in our CMB analysis. First, and more significantly, we test only power spectra that arise from the ΛCDM model. Second, we need to restrict the cosmological parameter space to a compact set in order to use Theorems 3.2 and 3.3. We use a two-stage confidence procedure with overall coverage probability $1 - \alpha$. In the first stage, a confidence estimator with coverage probability $1 - \alpha_1$ is employed, with α_1 very small, in our case 0.0001. This initial set is the compact parameter space Θ used in the application of the described MES and MR methods, except with coverage probability $(1 - \alpha - \alpha_1)$. (This small adjustment is of no practical significance.) The resulting confidence region is MES/MR *conditional on an event of probability* $1 - \alpha_1$. Since α_1 is so small, this is not viewed as a concern. This is not "data snooping" in the classic sense because the overall coverage probability remains at $1 - \alpha$.

We have a natural test with which to use in the first stage, a simple chi-square test. Although it is not optimal for the reasons given above, it is simple to implement in this

situation since under the stated assumptions,

$$(\mathbf{x} - \mu(\theta_0))^{\mathrm{T}} \Sigma(\theta_0)^{-1} (\mathbf{x} - \mu(\theta_0)) \tag{45}$$

has the chi-square distribution with 899 degrees of freedom. The initial test throws out parameter combinations for which the chi-square test statistic is too large, i.e. those combinations that differ substantially from the expected spectrum.

7.2.1 Computational Considerations

The described Monte Carlo technique is used to approximate the LFA/LRA. The vertices of the support of the prior are 300 randomly chosen cosmological parameter combinations that pass the first-stage chi-square test. Twenty-four thousand such parameter combinations are used as the randomly selected "nulls." The distribution under each null is approximated using 400 randomly simulated data vectors from each null. In the notation of Section 4 $p = 300$, $q = 24000$, and $n = 400$. Simulations results, given in Section 7.3, show that the Monte Carlo variability is not significant at these levels: the MR estimate is stable. However, at these dimensions the analysis requires substantial computational resources. In Section 9.1, included in the supplemental materials, we detail specific steps taken to make the implementation feasible.

7.3 Results

The primary analysis, constructing the 95% MR confidence region, consisted of testing 24,000 candidate parameter combinations (the "nulls"); 282 (1.1%) of them were accepted. Strictly speaking, the confidence set **is** this set of 282 parameter values/spectra, but this is of little practical use. We pursue a more useful projection, an acceptance rule such that tests are easy to perform, but faithful to the original results. We take a conservative viewpoint: The rule should accept any parameter combination which was accepted in the original set, at the expense of false positives.

    The darker portion of Figure 3 shows the 95% confidence band on the power spectrum formed by taking the outer envelope of all 282 accepted. The rule is thus that a spectrum needs to pass through the entire band in order to be accepted. The loss of precision is minimal: Of the 24,000 spectra, 126 are accepted under this rule that were not initially

accepted. Despite the layers of Monte Carlo simulations, the estimate does appear to be stable. Figure 2 compares the results of five repetitions of the analysis, each time redoing all of the simulations, i.e., choosing different support points for the prior, different nulls, and different simulated data values. The error bars shown give the variability in the upper or lower endpoint at various values of ℓ; since the spectra are all smooth, the size of the error bars vary smoothly between those given. Despite this initial evidence of stability, a primary concern regarding the use of this method is the variability of the estimate.
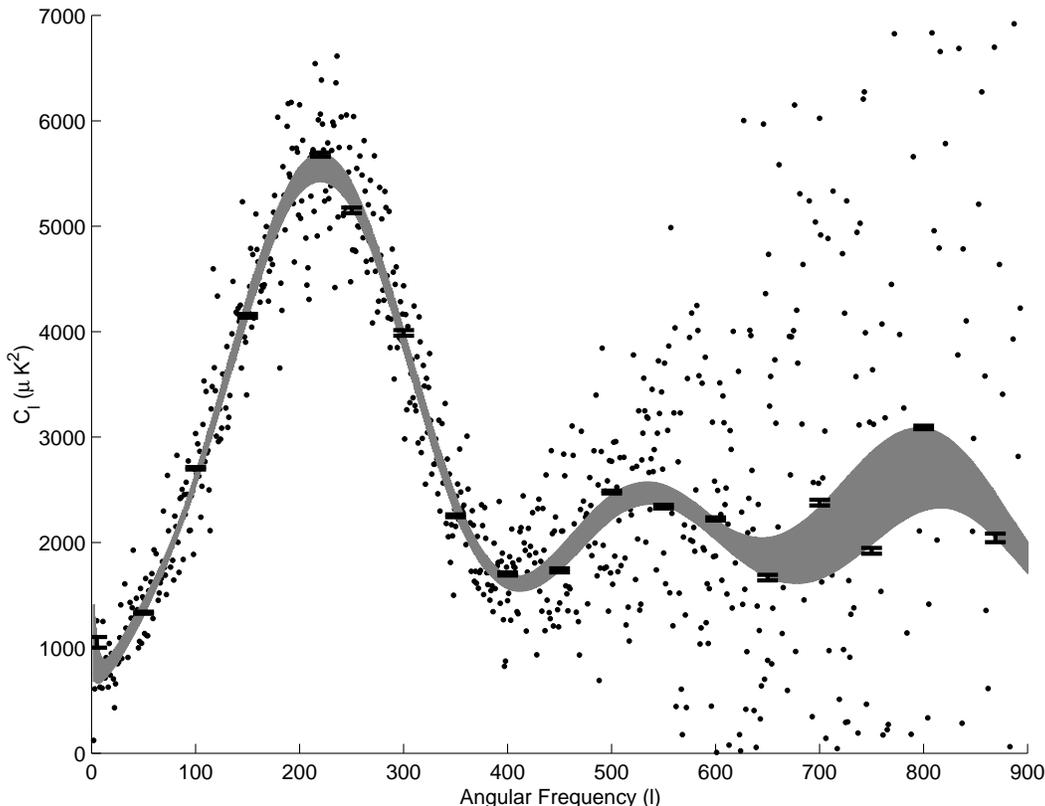
Figure 2: *A comparison of five repetitions of the MR 95% analysis. The error bars give the range of the bounds at different ℓ.*

Figure 3 also depicts the 68% MR confidence band; this confidence level is a common choice in the cosmology literature. Figure 4 shows both the 95% and 68% MES regions.

We stress that this estimate relies on the ΛCDM model described in Section 7.1. It is not valid for testing spectra arising from other models. The full collection of 135,000 spectra, each corresponding to a known combination of cosmological parameters, and each passing the first-stage test, are tested against these bands. A parameter vector is accepted if and only
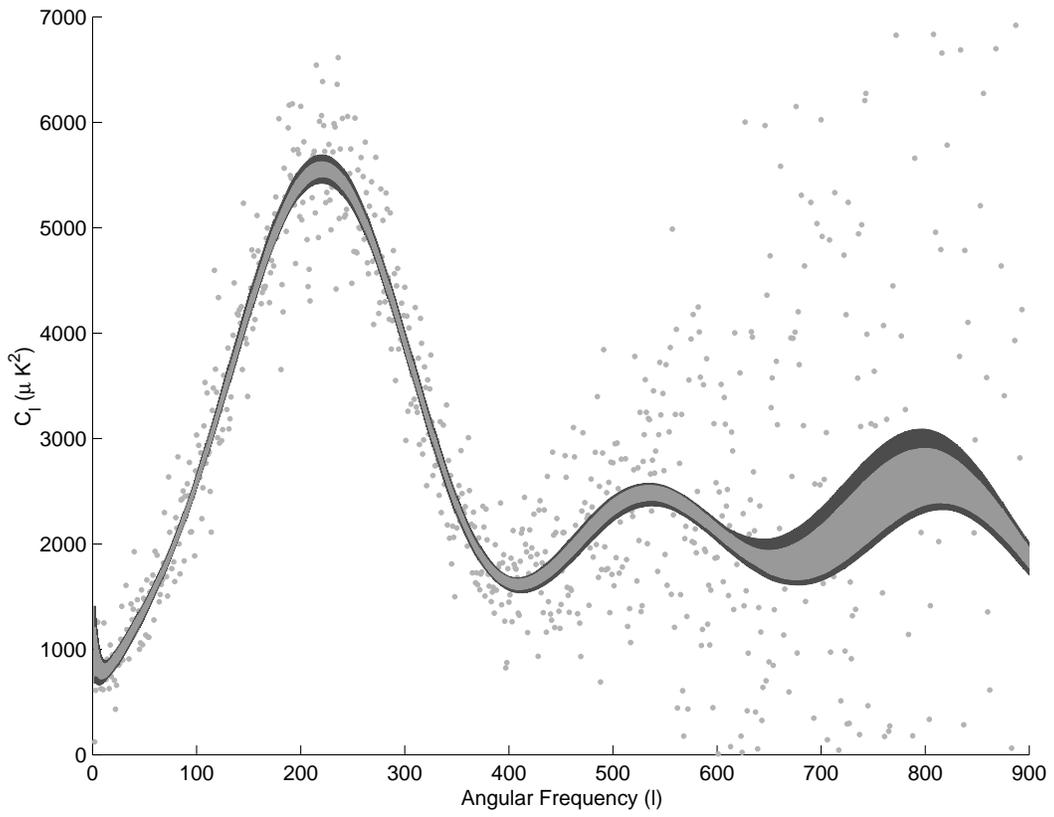
Figure 3: *95% (dark) and 68% (light) MR confidence bands on the spectrum.*
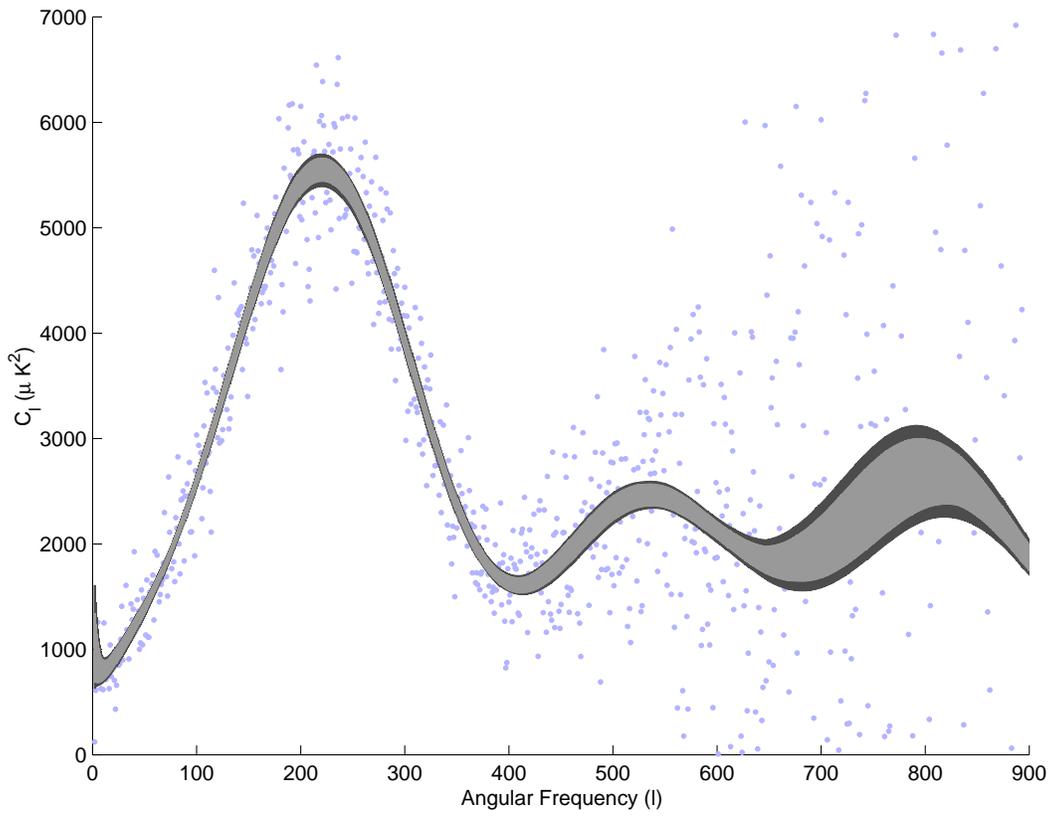
Figure 4: *95% (dark) and 68% (light) MES confidence bands on the spectrum.*

Parameter	MES	MR
Ω_b	$(0.032, 0.089)$	$(0.033, 0.083)$
Ω_m	$(0.15, 0.83)$	$(0.15, 0.74)$
H_0	$(53, 88)$	$(55, 88)$
A	$(0.66, 0.98)$	$(0.66, 0.98)$
n_s	$(0.92, 1.07)$	$(0.93, 1.07)$
τ	$(0.00, 0.27)$	$(0.00, 0.27)$
$\Omega_b h^2$	$(0.020, 0.028)$	$(0.021, 0.027)$
$\Omega_m h^2$	$(0.11, 0.24)$	$(0.11, 0.22)$

Table 3: *Ranges of accepted cosmological parameters, found by testing 135,000 models against the 95% MES and MR regions.*

if its spectrum lies entirely in the band. The acceptance rates were as follows: MR accepted 161 (0.1%) and 2039 (1.5%) for 68% and 95%, respectively, while MES accepted 909 (0.7%) and 4116 (3.0%). Tables 3 and 4 show the range of parameter values among the accepted spectra in each of the four cases. For example, there is a spectrum from the group of 135,000 that has $\Omega_b = 0.083$ which is accepted under the 95% MR procedure. These one-dimensional projections of the accepted spectra mask the complex tradeoffs between the parameters. It is typical to use the quantities $\Omega_m h^2$ and $\Omega_b h^2$, where $h \equiv H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})$; their ranges are shown in Tables 3 and 4.

For comparison, Table 4 includes the 68% credible intervals given in Spergel et al. (2003) based on their analysis of the WMAP data. The differences result partly from differences in the data sets. The WMAP experiment also measured the polarization of CMB photons; see Kogut et al. (2003) for details. Measurements of polarization are particularly useful in determining the value of τ since polarization is partly a result of reionization. Spergel et al. use the joint likelihood of the temperature and the cross-spectrum between temperature and polarization and they exclude $\tau < 0.089$. In their analysis of the cross-spectrum only, Kogut et al. (2003) construct a 68% credible interval for τ as $(13, 21)$. It is clear that our inability to reject small values of τ is related to our acceptance of lower values of H_0 since those two parameters have a strong degeneracy unresolvable by this data. The same can be said of the higher bounds on Ω_b and Ω_m. For us to include the temperature/polarization cross-spectrum would require finding a form of the joint likelihood that allows efficient Monte

Parameter	MES	MR	Spergel, et al.
Ω_b	$(0.038, 0.080)$	$(0.042, 0.071)$	$(0.041, 0.053)$
Ω_m	$(0.20, 0.69)$	$(0.24, 0.57)$	$(0.22, 0.36)$
H_0	$(55, 79)$	$(58, 75)$	$(67, 77)$
A	$(0.68, 0.95)$	$(0.69, 0.94)$	$(0.8, 1.0)$
n_s	$(0.94, 1.04)$	$(0.95, 1.04)$	$(0.95, 1.03)$
τ	$(0.00, 0.20)$	$(0.01, 0.18)$	$(0.089, 0.242)$
$\Omega_b h^2$	$(0.021, 0.027)$	$(0.022, 0.026)$	$(0.023, 0.025)$
$\Omega_m h^2$	$(0.13, 0.22)$	$(0.13, 0.19)$	$(0.12, 0.16)$

Table 4: *Ranges of accepted cosmological parameters, found by testing 135,000 models against the 68% MES and MR regions. Results are compared with the 68% Bayesian credible intervals reported in Spergel et al. (2003)*

Carlo simulations.

The differences observed in Table 4 also result from the different methods. Our one-dimensional projections are conservative in the sense that there only has to be one accepted spectrum for which $\Omega_b = 0.053$ for the interval to extend out to 0.053. The Bayesian credible interval would exclude that parameter value if the set of **all** parameter combinations with $\Omega_b = 0.053$ did not have sufficiently large posterior likelihood. We do not claim that our stated intervals are exact; they simply represent the range of accepted values from the set of spectra tested. The discrepancy in the intervals for A is interesting: Spergel et al. (2003) report that the maximum likelihood estimate for A is 0.78,[1] a value near the center of our range but completely excluded from theirs. This illustrates the pathological behavior possible when marginalizing the posterior over this parameter space.

8   Conclusion

A main goal of this research is to take a theoretically attractive idea, the construction of minimax expected size and minimax regret confidence procedures, and bring it to applications through a computationally feasible approximation procedure. Limiting the expected size of a confidence region has clear appeal, and we have allowed for broad generality in the definition of "size." In particular, the measure can depend on the truth, which is useful in

---

[1]Other MLE's are $H_0 = 68$, $\tau = 0.10$, $n_s = 0.97$, $\Omega_b h^2 = 0.023$, and $\Omega_m h^2 = 0.13$.

specific situations. The approximation is based on Monte Carlo simulations, but has theoretical backing: As the number of simulations increases, the maximum risk of the procedure converges to the real minimax risk.

The motivating application, the estimation of cosmological parameters from measurements of the cosmic microwave background radiation, is described. It was shown why this methodology is well-suited to inference in this problem.

References

Aitchison, J. (1966), "Expected-cover and Linear-utility Tolerance Intervals," *J. Roy. Stat. Soc., Ser. B*, 28, 57–62.

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

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999), *LAPACK Users' Guide*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 3rd ed.

Bennett, C., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S., Page, L., Spergel, D., Tucker, G., Wollack, E., Wright, E., Barnes, C., Greason, M., Hill, R., Komatsu, E., Nolta, M., Odegard, N., Peirs, H., Verde, L., and Weiland, J. (2003), "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results," *Astrophys. J. Suppl.*, 148, 1–27.

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

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

Bond, J., Jaffe, A., and Knox, L. (2000), "Radical Compression of Cosmic Microwave Background Data," *Astrophys. J.*, 533, 19–37.

Brown, G. (1951), "Iterative Solution of Games by Fictious Play," in *Activity Analysis of Production and Allocation*, ed. Koopmans, T., 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. Gupta, S. and Berger, J., New York: Springer-Verlag, pp. 237–251.

Chamberlain, G. (2000), "Econometric Applications of Maxmin Expected Utility," *J. of Appl. Econometrics*, 15, 625–644.

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

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

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

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

Guth, A. (1981), "Inflationary universe: A possible solution to the horizon and flatness problems," *Phys. Rev. D*, 23, 347–356.

Hinshaw, G., Spergel, D., Verde, L., Hill, R., Meyer, S., Barnes, C., Bennett, C., Halpern, M., Jarosik, N., Kogut, A., Komatsu, E., Limon, M., Page, L., Tucker, G., Weiland, J., Wollack, E., and Wright, E. (2003), "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: The Angular Power Sprectrum," *Astrophys. J. Suppl.*, 148, 135–159.

Hooper, P. (1982), "Invariant Confidence Sets with Smallest Expected Measure," *Ann. Stat.*, 10, 1283–1294.

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

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

Kempthorne, P. (1987), "Numerical Specification of Discrete Least Favorable Prior Distributions," *SIAM J. Sci. Stat. Comput.*, 8, 171–184.

Kogut, A., Spergel, D., Barnes, C., Bennett, C., Halpern, M., Hinshaw, G., Jarosik, N., Limon, M., Meyer, S., Page, L., Tucker, G., Wollack, E., and Wright, E. (2003), "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: TE Polarization," *Astrophys. J. Suppl.*, 148, 161–173.

Lewis, A. and Challinor, A. (2003), "CAMB: Code for Anisoptropies in the Microwave Background," Http://camb.info.

Nelson, W. (1966), "Minimax Solution of Statistical Decision Problems by Iteration," *Ann. Math. Stat.*, 37, 1643–1657.

Pratt, J. (1961), "Length of Confidence Intervals," *J. Am. Stat. Assoc.*, 56, 549–567.

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

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

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. Lyons, L., Mount, R., and Reitmeyer, R., SLAC.

Seljak, U. and Zaldarriaga, M. (1996), "A Line of Sight Approach to Cosmic Microwave Background Anisotropies," *Astrophys. J.*, 469, 437–444.

Smoot, G., Bennett, C., Weber, R., Maruschak, J., Ratliff, R., Janssen, M., Chitwood, J., Hilliard, L., Lecha, M., Mills, R., Patschke, R., Richards, C., Backus, C., Mather, J., Hauser, M., Weiss, R., Wilkinson, D., Gulkis, S., Boggess, N., Cheng, E., Kelsall, T., Lubin, P., Meyer, S., Moseley, H., Murdock, T., Shafer, R., Silverberg, R., and Wright, E. (1990), "COBE Differential Microwave Radiometers - Instrument design and implementation," *Astrophys. J.*, 360, 685–695.

Spergel, D., Verde, L., Peiris, H., Komatsu, E., Nolta, M., Bennett, C., Halpern, M., Jarosik, N., Kogut, A., Limon, M., Meyer, S., Page, L., Tucker, G., Weiland, J., Wollack, E., and Wright, E. (2003), "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters," *Astrophys. J. Suppl.*, 148, 195–211.

Van Zwet, W. (1980), "A Strong Law for Linear Functions of Order Statistics," *Ann. Prob.*, 8, 986–990.

Verde, L., Peiris, H., Spergel, D., Nolta, M., Bennett, C., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S., Page, L., Tucker, G., Wollack, E., and Wright, E. (2003), "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Parameter Estimation Methodology," *Astrophys. J. Suppl.*, 148, 195–211.

Vidakovic, B. (2000), "Γ-Minimax: A Paradigm for Conservative Robust Bayesians," in *Robust Bayesian Analysis*, eds. Ríos Insua, D. and Ruggeri, F., New York: Springer-Verlag, pp. 241–259.

Winkler, R. (1972), "A Decision-Theoretic Approach to Interval Estimation," *J. Am. Stat. Assoc.*, 67, 187–191.

9 Online Supplemental Material

9.1 Computational Details

A library of 135,000 cosmological parameter combinations that passed the chi-square test was formed using the the software package CAMB (Lewis and Challinor, 2003). Based on the standard subroutine CMBFAST (Seljak and Zaldarriaga, 1996), CAMB reduces the time from minutes to seconds (for flat models) with little loss of accuracy. It took between 20 and 30 seconds to find the spectrum for approximately 95% of the parameter combinations using CAMB. [2] The library was constructed using random walks through the parameter space. Random steps are taken and accepted only if the first-stage test accepts the new combination. Step sizes were adjusted to have a high rate of rejection (approximately 50%), in an attempt to ensure that the entire space is searched. Forty-two independent walks were constructed and compared, again in hopes of detecting irregular behavior such as missing portions of the space. Regardless, due to degeneracies in the parameters, it is a challenge to fully explore the space. Initial simulation results discussed later indicate that the confidence set is stable even after selecting a new suite of parameter combinations in this manner. This is not surprising: The integrals that are approximated via the Monte Carlo simulations require dense sampling of the space of spectra, not the cosmological parameter space. The role of the measure ν is to direct the sampling so that it is uniform in the parameter space. Steps taken in the random walks are perturbations in the values of the cosmological parameters. Out of this group of 135,000 spectra, $p = 300$ are chosen randomly to be the support for the prior (labeled $\{\theta_i\}_{i=1}^p$) and $q = 24,000$ are chosen to be the "nulls" (labeled $\{\eta_i\}_{i=1}^q$), i.e. the spectra that are accepted or rejected in the to form the confidence set. Following the development in Section 4, for each η_j there is a matrix $\mathbf{A_j}$ whose (i,k) entry is the likelihood ratio

$$\frac{|\Sigma(\theta_i)|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x_{jk}} - \mu(\theta_i))^{\mathrm{T}} \Sigma(\theta_i)^{-1} (\mathbf{x_{jk}} - \mu(\theta_i))\right)}{|\Sigma(\eta_j)|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x_{jk}} - \mu(\eta_j))^{\mathrm{T}} \Sigma(\eta_j)^{-1} (\mathbf{x_{jk}} - \mu(\eta_j))\right)}, \tag{46}$$

where $\{\mathbf{x_{jk}}\}_{k=1}^n$ are simulated data values distributed as η_j. (Here the penalty function ϕ is constant.) Equation (46) can be rewritten as

$$\frac{\exp\left(-\frac{1}{2}(\mathbf{G}(\eta_j)^{\mathrm{T}} \mathbf{z_{jk}} + \mu(\eta_j) - \mu(\theta_i))^{\mathrm{T}} \mathbf{D}(\theta_i)^{\mathrm{T}} \mathbf{D}(\theta_i) (\mathbf{G}(\eta_j)^{\mathrm{T}} \mathbf{z_{jk}} + \mu(\eta_j) - \mu(\theta_i))\right)}{|\Sigma(\theta_i)|^{1/2} |\Sigma(\eta_j)|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{z_{jk}}^{\mathrm{T}} \mathbf{z_{jk}}\right)} \tag{47}$$

---

[2]On a Sun UltraSparc II, 360 MHz, 128 MB of RAM

where $\mathbf{G}(\theta)$ is the Cholesky decomposition of $\Sigma(\theta)$, $\mathbf{D}(\theta)$ is the Cholesky decomposition of $\Sigma(\theta)^{-1}$ and $\mathbf{z_{jk}}$ is a vector of standard normal variates. This produces matrices $\mathbf{A_j}$ which are identically distributed to those using Equation (46), as can be shown easily using basic properties of the normal distribution.

Equation (47) facilitates the computation of the matrices $\mathbf{A_j}$:

1. Nearby ℓ, ℓ^* for which $\ell - \ell^*$ is even are binned in groups of four, summed to reduce the size of $\mathbf{x}$ further to 221 data values. This introduces little loss of information since the spectra are typically very smooth, and, as described above, such modes are moderately correlated. As this is a linear transformation of a normal vector, $\mathbf{x}$ remains Gaussian, and we still let μ and Σ denote the mean vector and covariance matrix, respectively, now assumed to be of reduced size.

2. The matrices $\{\mathbf{D}(\theta_i)\}_{i=1}^p$ are calculated and stored to disk. p is small enough that this does not introduce significant storage demands.

3. For each $1 \leq j \leq q$, the matrix $\mathbf{G}(\eta_j)$ is constructed, and then used over a sequence of iterations $k = 1, 2, \ldots, n$ where different vectors $\mathbf{z_{jk}}$ of standard normal variates are formed using the NAG routine `G05FDF`, its inner product is stored as b_{jk}. Set

$$\mathbf{M_{jk}} \equiv \mathbf{G}(\eta_j)^\mathsf{T} \mathbf{z_{jk}} + \mu(\eta_j). \tag{48}$$

Then, for each of $i = 1, 2, \ldots, p$,

$$\mathbf{N_{ijk}} \equiv \mathbf{D}(\theta_i) \left(\mathbf{M_{jk}} - \mu(\theta_i)\right) \tag{49}$$

is calculated using the matrices stored at step two. Each of these matrix multiplications exploits the upper triangular form of the Cholesky decomposition. Finally, the (k, i) entry of $\mathbf{A_j}$ is initialized with $\mathbf{N_{ijk}}^\mathsf{T}\mathbf{N_{ijk}} - b_{jk}$.

4. The quantity

$$\log |\Sigma(\eta_j)| - \log |\Sigma(\theta_i)| \tag{50}$$

is added to all entries of the ith column of $\mathbf{A_j}$. This is simple using already-generated quantities since for general positive definite matrix $\mathbf{A}$,

$$\log(|\mathbf{A}|) = 2 \sum_i \log(\text{Chol}(\mathbf{A})(i, i)) = -2 \sum_i \log\left(\text{Chol}\left(\mathbf{A}^{-1}\right)(i, i)\right). \tag{51}$$

LAPACK (Anderson et al., 1999) routines were used to compute Cholesky decompositions and matrix inverses. After the above steps, $\mathbf{A_j}$ holds minus two times the logarithm of the likelihood ratios shown in Equation (46). Working with the logarithms allows for much more stable calculations, avoiding products of very small or very large numbers. For this reason, the `LFA_Search` subroutine accepts these log ratios as its main argument.

The process of building the $\mathbf{A_j}$ matrices and approximating the LFA/LRA was done on a subset of 32 processors of Seaborg, a parallel computer housed at NERSC at Lawrence Berkeley Laboratory. All steps of the process are parallelized; separate processors build different matrices $\mathbf{A_j}$, and `LFA_Search` is likewise designed to run with different processors handling different subsets of the nulls η_j. It takes an average of 27 seconds to form one matrix $\mathbf{D}(\theta_i)$, and 16.5 seconds to completely construct one of the $\mathbf{A_j}$ matrices. With all 32 processors, it takes approximately 210 minutes to do both of the above steps for all 24,000 nulls. The iterations that comprise the `LFA_Search` procedure take about 0.03 seconds each. In a batch of runs of the entire process, total CPU time consumed ranged from 6.75 to 9.00 hours, the variability due to differences in time to convergence for the iterative search for the LFA/LRA.

9.2 Proofs

9.2.1 Proof of Theorem 1

We prove a stronger theorem: The assumptions that ν is a probability measure, that Γ has a finite number of vertices, and that $v_\theta(\cdot)$ is bounded from above are not necessary. It is necessary, however, that ν be σ-finite and that $0 \leq v_\theta(\eta) < \infty$. These results are taken directly from EHS, modified to fit this situation.

Lemma 2. *As a function of d (fix $\pi \in \Gamma$), $\mathbf{R}_\pi(d)$ is a weak-star lower semicontinous mapping of $L_\infty[\nu \times \mu]$ into $[0, \infty]$.*

Proof. This is almost verbatim the proof of Lemma 1 in EHS. A slight change was needed to allow for the additional generality introduced by the penalty function.

Fix $\pi \in \Gamma$. Form $\{A_j\}_{j=1}^\infty$, an increasing sequence of nested ν-measurable subsets of Θ such that $\nu(A_j) < \infty$, $v_\theta(\eta) \leq j$ for $\eta \in A_j$, and $\cup_j A_j = \Theta$. Then,

$$\mathbf{R}_\pi(d) = \int_\Theta \int_{\mathcal{X}} f_\eta(x)\, r_\pi(\eta, x)\, d(\eta, x)\, \mu(dx)\, \nu(d\eta)$$

$$= \sup_j \int_\Theta \int_{\mathcal{X}} \left[1_{A_j}(\eta)\, f_\eta(x)\, r_\pi(\eta, x) \right] d(\eta, x)\, \mu(dx)\, \nu(d\eta)$$

by monotone convergence. For fixed j the double integral is a weak-star continuous functional in d because the term in brackets is a member of $L_1[\nu \times \mu]$. As the supremum of a collection of weak-star continuous functionals, $\mathbf{R}_\pi(d)$ is weak-star lower semicontinuous. $\square$

The following two results come directly from EHS (Theorem 4 and Lemma 2, respectively), and they are stated without proof.

Lemma 3. *Let M be a convex set and let $\mathcal{T} : M \times N \to [-\infty, \infty]$ be linear in M and convex-like in N, in the sense that for each $n_0, n_1 \in N, \kappa \in (0, 1)$, there is $n_\kappa \in N$ such that*

$$\kappa \mathcal{T}(m, n_0) + (1 - \kappa)\, \mathcal{T}(m, n_1) \geq \mathcal{T}(m, n_\kappa)$$

for all $m \in M$. If N is a compact topological space and $\mathcal{T}(m, n)$ is lower semicontinuous in n for each m, then

$$\inf_{n \in N} \sup_{m \in M} \mathcal{T}(m, n) = \sup_{m \in M} \inf_{n \in N} \mathcal{T}(m, n). \tag{53}$$

Lemma 4. *If $\alpha \in [0, 1]$, then $\mathcal{D}_\alpha \subseteq L_\infty[\nu \times \mu]$ is weak-star compact.*

Combining the above results, Lemma 1, and the definition of π_0 gives

$$\inf_{d \in \mathcal{D}_\alpha} \sup_{\pi \in \Gamma} \mathbf{R}_\pi(d) = \sup_{\pi \in \Gamma} \inf_{d \in \mathcal{D}_\alpha} \mathbf{R}_\pi(d) = \sup_{\pi \in \Gamma} \mathbf{R}_\pi(d_\pi). \tag{54}$$

This proves Theorem 1.

9.2.2 Proof of Theorem 2

We will assume that $n_m = m$. It should be clear that this has no effect on the limiting results.

Lemma 5 (Van Zwet (1980)). *Suppose $J, J_1, J_2, \ldots$ are each Lebesgue measurable functions $[0, 1] \to I\!\!R$, are uniformly bounded, and are 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 uniform$(0,1)$ random variables. Define $U_{1:m}, U_{2:m}, \ldots, U_{m:m}$ to be $U_1, U_2, \ldots, U_m$ placed in increasing order. Next, let $g \colon [0,1] \to \mathbb{R}$ be a Borel measurable, integrable function and define

$$g_m(t) \equiv g\big(U_{[mt]+1:m}\big)$$

where $[x]$ denotes the integer portion of x. Then,

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

Lemma 6. *Fix $\eta \in \Theta$ and π. Then,*

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

Proof. We will apply Lemma 5 defining $J_m(u)$ 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 cumulative distribution function for $r_\pi(\eta, X)$ when X is distributed as $\mathbb{P}_\eta$, i.e. $R(x) = \mathbb{P}_\eta(r_\pi(\eta, X) \le x)$. The function $g(\cdot)$ of Lemma 5 is $g(u) = \inf\{x \colon R(x) \ge u\}$. Thus, if U is a uniform$(0,1)$ random variable, $g(U)$ is a random variable with cdf $R(\cdot)$ since $\inf\{x \colon R(x) \ge u\} \le y$ if and only if $R(y) \ge u$. (This is the familiar "inversion" method for generating random variates from an arbitrary CDF.) We know $g(\cdot)$ is integrable since

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

Next define $u' = \inf\{u \colon g(u) = g(1-\alpha)\}$, $a = g(1-\alpha)$, and

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

34

Then,

$$
\begin{aligned}
\int_0^1 J(u)\,g(u)\,du &= \int_0^{u'} g(u)\,du + \int_{u'}^{1-\alpha} g(u)\,du \\
&= \mathbb{E}\big(g(U)\,\mathbf{1}_{\{U<u'\}}\big) + \mathbb{E}\big(g(U)\,\mathbf{1}_{\{u'\le U\le 1-\alpha\}}\big) \\
&= \mathbb{E}\big(g(U)\,\mathbf{1}_{\{g(U)<g(u')\}}\big) + a(1-\alpha-u') \qquad (55) \\
&= \mathbb{E}\big(g(U)\,\mathbf{1}_{\{g(U)<a\}}\big) + a(1-\alpha-u') \\
&= \mathbb{E}_\eta\big(r_\pi(\eta,X)\,\mathbf{1}_{\{r_\pi(\eta,X)<a\}}\big) + c\,\mathbb{E}_\eta\big(r_\pi(\eta,X)\,\mathbf{1}_{\{r_\pi(\eta,X)=a\}}\big) \\
&= \int_{\mathcal{X}} r_\pi(\eta,x)\,d^*(\eta,x)\,\mathbb{P}_\eta(dx) \\
&= \inf_{d\in\mathcal{D}_\alpha} \int_{\mathcal{X}} r_\pi(\eta,x)\,d(\eta,x)\,\mathbb{P}_\eta(dx) \qquad (56) \\
&= \inf_{d\in\mathcal{D}_\alpha} \int_{\Theta} \gamma_d(\theta,\eta)\,v_\theta(\eta)\,\pi(d\theta) \qquad (57)
\end{aligned}
$$

where

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

Here, equation (55) holds because $g(U) < g(u')$ if and only if $U < u'$; equation (56) holds because $d^* \in \mathcal{D}_\alpha$.

Next, begin by considering the function $U\colon \mathcal{X} \times [0,1] \to [0,1]$ 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 is distributed as $\mathbb{P}_\eta$. Then, if $W_1, W_2, \ldots$ are independent uniform$(0,1)$ random variables, and $X_1, X_2, \ldots$ are independent random variables distributed as $\mathbb{P}_\eta$, we have that $U_1 \equiv U(X_1, W_1), U_2 \equiv U(X_2, W_2), \ldots$ are independent uniform$(0,1)$ random variables. Further,

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

Let $X_{1:m}, X_{2:m}, \ldots, X_{m:m}$ denote $X_1, X_2, \ldots, X_m$ ordered by the value of $r_\pi(\eta, X_i)$ (the way in which ties are handled is unimportant). Likewise, $U_{1:m}, U_{2:m}, \ldots, U_{m:m}$ denotes

$U_1, U_2, \ldots, U_m$ placed 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,

$$
\begin{aligned}
\int_0^1 J_m(u)\, g_m(u)\, du &= \int_0^{1-\alpha_m} g_m(u)\, du \\
&= \frac{1}{m} \sum_{k=1}^m g(U_{k:m})\, d^*(\eta, k) \\
&= \frac{1}{m} \sum_{k=1}^m r_\pi(\eta, X_{k:m})\, d^*(\eta, k) \\
&= \inf_{d \in \mathcal{D}'_{\alpha_m}} \frac{1}{m} \sum_{k=1}^m r_\pi(\eta, X_k)\, d(\eta, X_k) \quad (58)
\end{aligned}
$$

where

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

with $k' = \inf\{k \in \mathbb{Z} \colon k \geq (1 - \alpha_m)m\}$.

Combining Lemma 5 with equations (57) and (58) we find that

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

Next, define

$$
\bar{K} \equiv \left[K \times \left(\frac{1}{m} \sum_i \sum_k r_\pi(\eta, X_k) \right)^{-1} \right] \bigwedge 1. \quad (59)
$$

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

$$
\mathbb{E}\left[\sum_i r_\pi(\eta, X_k) \right] \leq pM < K. \quad (60)
$$

Hence,

$$
Z_{m,\pi}(\eta) \equiv \inf_{d \in \mathcal{D}'_{\alpha_m}} \frac{1}{m} \sum_{k=1}^m r_\pi(\eta, X_k)\, d(\eta, X_k)\, \bar{K} \xrightarrow{a.s.} \inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta)\, v_\theta(\eta)\, \pi(d\theta)\,. \quad (61)
$$

$\square$

Lemma 7. *As $m \to \infty$,*

$$
\mathbb{E}[Z_{m,\pi}(T_{jm})] \longrightarrow \mathbf{R}_\pi(d_\pi)\,. \quad (62)
$$

36

Proof. Through two applications of the bounded convergence theorem, we can see that for fixed $\eta \in \Theta$

$$\mathbb{E}[Z_{m,\pi}(\eta)] \longrightarrow \inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta) \, v_\theta(\eta) \, \pi(d\theta) \tag{63}$$

and that

$$\int_\Theta \mathbb{E}[Z_{m,\pi}(\eta)] \, \nu(d\eta) \longrightarrow \int_\Theta \left[\inf_{d \in \mathcal{D}_\alpha} \int_\Theta \gamma_d(\theta, \eta) \, v_\theta(\eta) \, \pi(d\theta) \right] \nu(d\eta) . \tag{64}$$

But

$$\int_\Theta \mathbb{E}[Z_{m,\pi}(\eta)] \, \nu(d\eta) = \mathbb{E}[Z_{m,\pi}(T_{jm})] \tag{65}$$

and

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

The infimum and integral can be switched because, as established in Lemma 2.1, the infimal d is formed by minimizing for fixed η. $\square$

Lemma 8. $\{U_m\}_{m=1}^\infty$ *is a sequence of random variables such that*

$$U_m = \frac{1}{q_m} \sum_{j=1}^{q_m} V_{jm} \tag{66}$$

where

1. $\{V_{jm}\}_{j=1}^{q_m}$ *are i.i.d. for each m and independent across different 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 $U_m \xrightarrow{a.s.} \mu$.

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

$$\mathbb{P}[|U_m - \mu| > \epsilon] \leq \mathbb{P}[|U_m - \mu_m| > \epsilon/2] \leq \left(\frac{16}{\epsilon^4} \right) \mathbb{E}\big[(U_m - \mu_m)^4\big] \tag{67}$$

using the Markov inequality. Setting $W_{jm} \equiv V_{jm} - \mu_m$,

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

where c is a constant independent of m. (See page 85 in Billingsley (1995), the proof of Theorem 6.1). Hence, by Borel-Cantelli,

$$
\mathbb{P}[|U_m - \mu| > \epsilon \ \ \text{i.o.}] = 0. \tag{68}
$$

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

The results above combined imply that as $m \to \infty$,

$$
\widehat{\mathbf{R}}_\pi(d_{\pi,m}) = \frac{1}{q_m}\sum_{j=1}^{q_m} Z_{m,\pi}(T_{jm}) \xrightarrow{a.s.} \mathbf{R}_\pi(d_\pi) \tag{69}
$$

for any probability distribution π on Θ.

Lemma 9. *Let* $\pi, \pi' \in \Gamma$ *such that* $\pi = \sum_i w_i \pi_i$ *and* $\pi' = \sum_i w_i' \pi_i$. *Then, for all* m,

$$
\left|\widehat{\mathbf{R}}_\pi(d_{\pi,m}) - \widehat{\mathbf{R}}_{\pi'}(d_{\pi',m})\right| \leq K\|w - w'\|_1. \tag{70}
$$

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

$$
\sum_k r_\pi(T_{jm}, X_{jkm})\ d(T_{jm}, X_{jkm}) \tag{71}
$$

and

$$
\sum_k r_{\pi'}(T_{jm}, X_{jkm})\ d(T_{jm}, X_{jkm}). \tag{72}
$$

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

$$
\begin{aligned}
|Z_{m,\pi}(T_{jm}) - Z_{m,\pi'}(T_{jm})| &\leq \sum_i |w_i - w_i'|\left(\frac{1}{m}\sum_k r_{\pi_i}(T_{jm}, X_{jkm})\ d'(T_{jm}, X_{jkm})\ K_{jm}\right) \\
&\leq K\sum_{i=1}^{p} |w_i - w_i'| \\
&= K\,\|w - w'\|_1.
\end{aligned}
$$

The fact that

$$\widehat{\mathbf{R}}_\pi(d_{\pi,m}) = \frac{1}{q_m} \sum_{j=1}^{q_m} Z_{m,\pi}(T_{jm}) \tag{73}$$

leads to the desired result. $\qquad\qquad\square$

Lemma 9 implies that the family $\{\widehat{\mathbf{R}}_\pi(d_{\pi,m})\}_{m=1}^{\infty}$ is equicontinuous as functions of the weight vector w associated with π. The space of possible weights is compact, so the pointwise convergence established above for fixed π is uniform in π. (See Royden (1988), page 168, Lemma 39). This completes the proof of Theorem 2. $\square$