Approximate Bayesian Inference with the Weighted Likelihood Bootstrap
Author(s): Michael A. Newton and Adrian E. Raftery
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 56, No. 1 (1994), pp. 3-48
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2346025
Accessed: 10-10-2018 19:25 UTC

J. R. Statist. Soc. B (1994)
56, *No.* 1, *pp.* 3–48

Approximate Bayesian Inference with the Weighted Likelihood Bootstrap

By MICHAEL A. NEWTON† and ADRIAN E. RAFTERY

University of Wisconsin, Madison, USA *University of Washington, Seattle, USA*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, April 7th, 1993,* Professor B. W. Silverman *in the Chair*]

SUMMARY

We introduce the *weighted likelihood bootstrap* (WLB) as a way to simulate approximately from a posterior distribution. This method is often easy to implement, requiring only an algorithm for calculating the maximum likelihood estimator, such as iteratively reweighted least squares. In the generic weighting scheme, the WLB is first order correct under quite general conditions. Inaccuracies can be removed by using the WLB as a source of samples in the sampling–importance resampling (SIR) algorithm, which also allows incorporation of particular prior information. The SIR-adjusted WLB can be a competitive alternative to other integration methods in certain models. Asymptotic expansions elucidate the second-order properties of the WLB, which is a generalization of Rubin's Bayesian bootstrap. The calculation of approximate Bayes factors for model comparison is also considered. We note that, given a sample simulated from the posterior distribution, the required marginal likelihood may be simulation consistently estimated by the harmonic mean of the associated likelihood values; a modification of this estimator that avoids instability is also noted. These methods provide simple ways of calculating approximate Bayes factors and posterior model probabilities for a very wide class of models.

Keywords: BAYES FACTOR; BAYESIAN INFERENCE; DIRICHLET WEIGHTS; MONTE CARLO METHODS

1. INTRODUCTION

This paper investigates the extent to which a new bootstrap procedure—the weighted likelihood bootstrap (WLB)—can be used by applied Bayesian statisticians to approximate posterior distributions. This is a direct extension of the Bayesian bootstrap (Rubin, 1981) from nonparametric models to parametric and semiparametric models. It is a Monte Carlo method which is particularly simple to apply in models where maximum likelihood estimation is feasible, as in many regression models and generalized linear models, for example. Unlike Markov chain simulation algorithms (Gelfand and Smith (1990) and Tierney (1991), for example), the WLB used in isolation is not simulation consistent, i.e. it does not produce exact answers as the amount of computing resources increases without bound. Rather, it provides an approximation which improves as more data become available. The WLB produces a random sample of parameter values from a distribution on the parameter space which approximates a Bayesian posterior. When used in conjunction with other methods, like sampling–importance resampling (SIR) (Rubin, 1987, 1988), the WLB output can be modified to produce a sample from the posterior of interest, and so the adjusted WLB is simulation consistent. The WLB may also form a good starting point for the adaptive

†*Address for correspondence*: Department of Statistics, University of Wisconsin–Madison, 1210 West Dayton Street, Madison, WI 53706-1685, USA.

 0035–9246/94/56003

importance sampling algorithm (West, 1992) when the normal and t-approximations are poor.

The paper is organized as follows. Section 2 contains a description of the method for independent data, and an application to a non-linear regression problem. Exact properties of the WLB are studied in Section 3, and asymptotic properties in Section 4. The ease of implementation of the WLB is investigated in Section 5, where three other examples are worked out. An extension to dependent data models and to partial likelihood is presented in Section 6. In Section 7, we study a method for approximating marginal likelihoods by using samples from a posterior distribution. This method can be used to compute Bayes factors from the output of the WLB or any other posterior simulation device.

2. THE METHOD

Initially, consider modelling data $x_1, x_2, \ldots, x_n$ as independent, each x_i having probability density function $f_i(x_i; \theta)$ with respect to some σ-finite measure on the sample space. Bayesian inference ultimately requires some knowledge of the likelihood function for the parameter θ,

$$L(\theta) = \prod_{i=1}^{n} f_i(x_i; \theta),$$

having started with a prior $\pi(\theta)$. In lieu of analytical expressions for various integrals of the posterior density proportional to $L(\theta)\,\pi(\theta)$, approximate inference proceeds by considering empirical averages taken with respect to a sample drawn from some appropriate distribution on the parameter space. Monte Carlo methods, including the WLB, are based on this premise.

In the WLB method, a sample is produced by maximizing a weighted likelihood function

$$\tilde{L}(\theta) := \prod_{i=1}^{n} f_i(x_i; \theta)^{w_{n,i}}, \tag{1}$$

where the weight vector $w_n = (w_{n,1}, \ldots, w_{n,n})$ has some probability distribution determined by the statistician. The function $\tilde{L}$ is not a likelihood in the usual sense; it is merely a device for generating a sample on the parameter space. We denote by $\tilde{\theta}$ any parameter value satisfying $\tilde{L}(\tilde{\theta}) \geq \tilde{L}(\theta)$ for all θ in the parameter space. (Without some constraints, $\tilde{\theta}$ is not guaranteed to be unique.) Whereas the likelihood function L (and the maximum likelihood estimate $\hat{\theta}$) are fixed after the data are observed, the weighted likelihood $\tilde{L}$ (and its maximizer $\tilde{\theta}$) have randomness induced by the distribution of the weights. Our thesis is that, for certain weight distributions, the conditional distribution of $\tilde{\theta}$ given the data can provide a good approximation to a posterior distribution of θ. Although this conditional distribution is usually difficult to find exactly, it is straightforward to simulate when maximization of $\tilde{L}$ is feasible. The simulation amounts to repeatedly sampling weight vectors and maximizing $\tilde{L}$.

Motivated by inference for multinomial data, a natural weight distribution is the uniform Dirichlet distribution. Such weights are simulated by generating n independent exponentials Y_i and forming $w_{n,i} = Y_i / \overline{Y}$ where $\overline{Y}$ is the sample mean of the Y_i. When combining the WLB with importance sampling, it is computationally convenient

to form weights $w_{n,i} \propto Y_i^\alpha$ for some power $\alpha \neq 1$. If $\alpha > 1$, the weight distribution is *overdispersed* relative to the Dirichlet distribution, in the same sense as used by Gelman and Rubin (1992), although in a different context. Indeed, many weight distributions could be postulated, and the quality of the approximation certainly depends on the choice. We investigate several choices in this paper.

The following example illustrates the power and simplicity of the WLB method. A non-linear regression model, considered by Marske (1967), and studied in detail by Bates and Watts (1988), relates biochemical oxygen demand x of prepared water samples to incubation time t by the equation

$$x_i = \beta_1\{1 - \exp(-\beta_2 t_i)\} + \epsilon_i \qquad i = 1, 2, \ldots, n.$$

The errors ϵ_i are assumed to be independent normal errors with constant variance σ^2, on which we assign an improper prior $\pi(\sigma^2) \propto \sigma^{-2}$. The generic parameter θ incorporates both the regression parameter vector β and the scale parameter σ. As described in Bates and Watts (1988), a transformation invariant, design-dependent prior for $\beta = (\beta_1, \beta_2)$ is $\pi(\beta) \propto |V^T V|^{1/2}$ where V is the $n \times 2$ matrix having (i, j)th element $\partial E_\beta(x_i)/\partial \beta_j$. Contours of this prior are shown in Fig. 1(a). Given the small data set of Marske ($\{(t_i, x_i)\} = \{(1, 8.3), (2, 10.3), (3, 19), (4, 16), (5, 15.6), (7, 19.8)\}$, with units (days, mg l^{-1})), we can quite directly obtain maximum likelihood estimates for this model by using routines for non-linear optimization. We use the S function 'nls' (Bates and Chambers, 1992); built-in functions in other languages could also be used. Bayesian inference is not simple here because marginal posteriors pose difficult integration problems.

The WLB takes advantage of the available estimation technology to carry out the integration. As described more precisely in Section 5, maximizing $\tilde{L}$ for any particular set of weights is done by simply including a weight vector in the estimation routine. This weight vector needs to be known only up to a multiplicative constant, and so the unnormalized weights Y_i^α can be used. A raw sample of WLB parameter values is produced by repeatedly generating a random weight vector and applying the estimation code to the appropriately weighted cases. This sample provides a first-order correct approximation (at least) to the true posterior distribution in an asymptotic sense (as $n \to \infty$, and $\alpha \to 1$; see Section 4). The WLB samples do not come from the joint posterior exactly, because, for one thing, no prior information was used in the simulation. However, a simple adjustment based on importance sampling can correct this.

The basic idea is that g, the joint density of $\tilde{\beta}$, is a good approximation to the marginal posterior density of β and hence is a good choice for an importance sampling density. Although g is generally unavailable, it can be simulation consistently estimated by a kernel density estimate $\hat{g}$ using the sample of WLB parameter values. Such a density estimate is shown in Fig. 1(d) for 5000 $\tilde{\beta}$s simulated from a weight distribution with $\alpha = 1.6$. The kernel is normal, and its covariance matrix is determined by using Terrell's (1990) method of maximal smoothing, i.e. a particular scale multiple of the inverse Hessian of the log-likelihood at its maximum. For comparison, contours of the exact marginal likelihood (having integrated out σ^2) and marginal posterior for β are shown in Figs 1(b) and 1(c). The WLB sample has captured the structure of this highly non-elliptical posterior distribution. To finish the analysis, each $\tilde{\beta}^j$ in the raw WLB sample must be assigned an importance weight

$$u_j \propto r(\tilde{\beta}^j) = \pi(\tilde{\beta}^j)\, L_m(\tilde{\beta}_j)/\hat{g}(\tilde{\beta}_j)$$

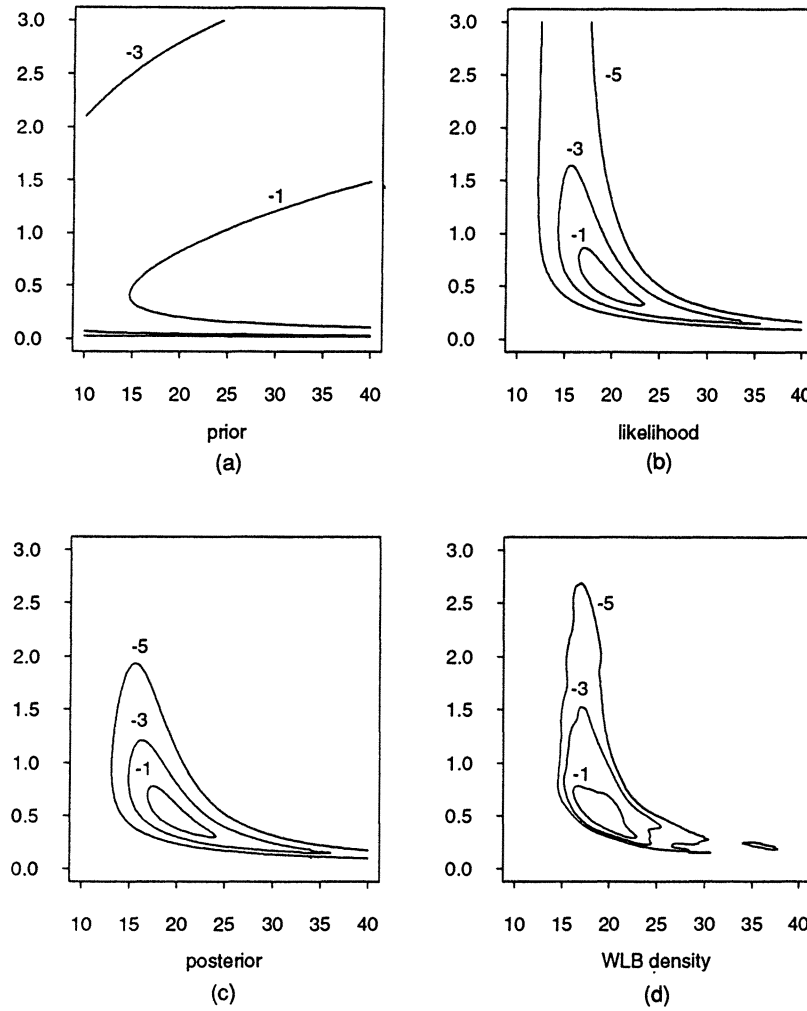

Fig. 1. Inference for the non-linear regression of Section 2: (a)–(c) contours of the prior, likelihood and posterior respectively (in integer units from the maximum on the log-scale) for the regression parameter β; (d) contours of the kernel density estimate produced by the WLB with 5000 draws (units are mg l^{-1} for θ_1 and days^{-1} for θ_2)

such that the weights sum to 1. (Since σ^2 has been integrated out, the likelihood L is replaced by a marginal likelihood L_m above.) To obtain a final sample, we might try rejection sampling based on these importance weights, but then a bound is needed on $r(\beta)$. Instead, we sample from the discrete distribution determined by these weights. This is precisely the SIR algorithm of Rubin (1988) (see also Gelfand and Smith (1992)). The final sample, an SIR-adjusted WLB sample, represents a simulation consistent estimate of the true posterior distribution of interest.

It is well known that, if an approximating density is not sufficiently close to the density of interest, then the importance weights can be dominated by a very small minority. Ritter *et al.* (1991) have studied this for the same non-linear regression example, showing that, in a t-approximation to the likelihood, 10 of 10000 weights

carry 60% of the weight. In contrast, the importance weights produced in our analysis are very stable, with the largest 40% carrying only half the weight.

Markov chain simulation methods, although almost always applicable in principle, are more difficult to apply than the SIR-adjusted WLB in this example. The entire analysis above, including generation of random weights, repeated maximization of $\tilde{L}$, density estimation, construction of importance weights and resampling to form the final posterior sample, was implemented in about 60 lines of S code, using built-in functions. The full conditionals required to run a Gibbs sampler, in contrast, do not have a simple form. Generally, for non-linear regressions, these conditionals are not log-concave, and so adaptive rejection sampling (Gilks and Wild, 1992) can fail. A 'griddy' Gibbs sampler can be run (Ritter and Tanner, 1992); indeed, a variety of approximations could be tried. By contrast, the WLB and the SIR-adjusted WLB are routine calculations in models like the non-linear regression above. In our experience with various regression models, it can take many hours to program a Gibbs sampler successfully.

3. ORIGINS OF WEIGHTED LIKELIHOOD BOOTSTRAP: MULTINOMIAL SAMPLING

The WLB with uniform Dirichlet weights simulates the posterior from an identifiable prior for the unconstrained multinomial model. Let $\{x_i\}$ be independent random variables, each taking one of k distinct values with probabilities $\theta=(\theta_1, \ldots, \theta_k)$. The likelihood and weighted likelihood functions collapse to become

$$L(\theta)=\prod_{j=1}^{k}\theta_j^{y_j}, \qquad \tilde{L}(\theta)=\prod_{j=1}^{k}\theta_j^{n\gamma_j}$$

where y_j counts the number of x_i equal to the jth distinct value, and similarly $n\gamma_j$ is the sum of the weights $w_{n,i}$ of data points x_i equal to the jth distinct value. From properties of Dirichlet random vectors, the vector $\gamma=(\gamma_1, \ldots, \gamma_k)$ has a Dirichlet distribution with parameters $y_1, \ldots, y_k$, i.e. it has probability density

$$p(\gamma)\propto\prod_{j=1}^{k}\gamma_j^{y_j-1}\mathbf{1}[y_j>0]. \tag{2}$$

When there are no modelling constraints on θ, $\tilde{\theta}=\gamma$, and so the WLB using uniform Dirichlet weights simulates the posterior of θ under the improper prior $\Pi_j\theta_j^{-1}$. Alternative Dirichlet distributions for the $w_{n,i}$ can produce posteriors under any conjugate prior. The Bayesian bootstrap (Rubin, 1981) follows from the above considerations under the *nonparametric* assumption that the unknown distribution of the data supports only observed values.

Models constrain probabilities, in particular the multinomial probabilities discussed in this section. A simple trinomial example from linkage analysis has probabilities (p_1, p_2, p_3) constrained by a parameter $\theta\in(0, 1)$:

$$(p_1, p_2, p_3)=\left(\frac{2+\theta}{4}, \frac{1-\theta}{2}, \frac{\theta}{4}\right). \tag{3}$$

Observed cell counts in this much-studied example (Rao, 1973; Dempster *et al.*, 1977;

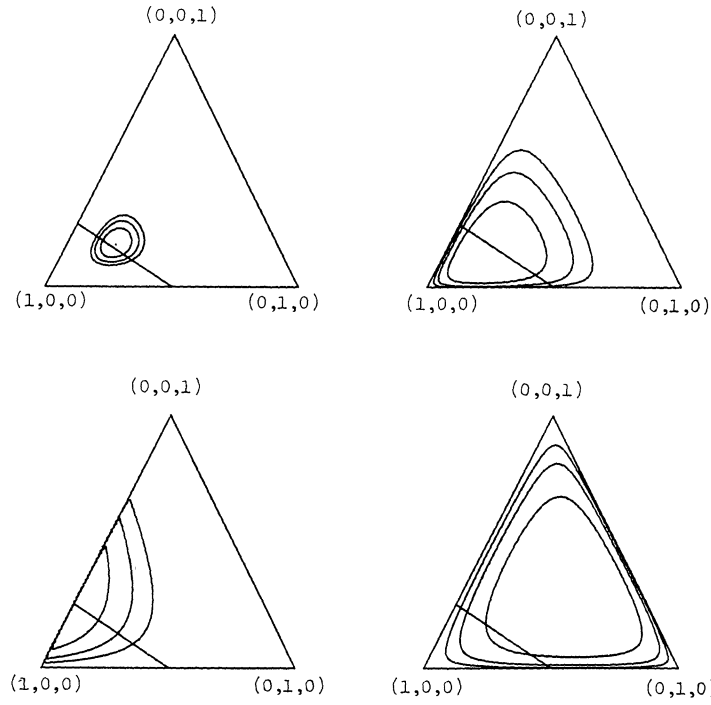

Fig. 2. Weight distribution: contours are at integer units of log-density from the maximum for the density function of equation (2); the four parts correspond to the four data sets of Table 1, and the line of negative slope in each part indicates the probability vectors satisfying the model constraints (equation (3))

TABLE 1
Four examples of linkage data, one per row

y_1	y_2	y_3	$n=\Sigma y_i$
125	38	34	197
13	4	3	20
14	1	5	20
3	4	3	10

Tanner and Wong, 1987) are shown in the top row of Table 1. The following three rows give particular subsets analysed by Tanner and Wong.

Fig. 2 shows contours of density (2) of the collapsed weight vector $(\gamma_1, \gamma_2, \gamma_3)$, which is the posterior density of (p_1, p_2, p_3) under an improper prior and when no model constraints are active. In terms of the collapsed weights γ_j, the maximizer of $\tilde{L}$ for this model is

$$\tilde{\theta} = -\tfrac{1}{2}(\gamma_2 - 2\gamma_1 + 1) + \tfrac{1}{2}\sqrt{\{(\gamma_2 - 2\gamma_1 + 1)^2 + 8\gamma_3\}}$$

which corresponds to a probability vector $\tilde{p} = (\tilde{p}_1, \tilde{p}_2, \tilde{p}_3)$ in the model given by

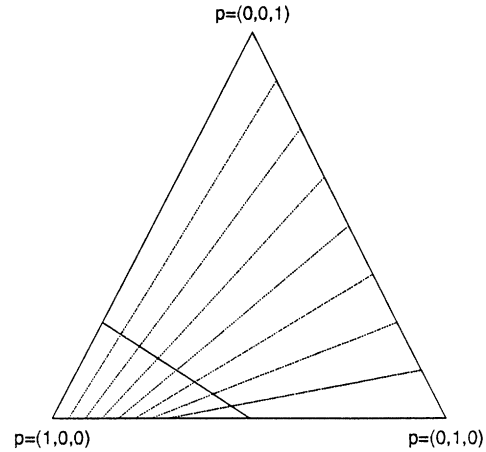

Fig. 3. The line of negative slope is the linkage model, the set of trinomial probability vectors satisfying certain constraints: in a WLB simulation, points are sampled from the simplex and then projected down into the model; the broken lines show how this projection happens (all the points on the same broken line are projected onto the same point in the model)

equation (3). Maximization of $\tilde{L}$ for a given vector $\gamma=(\gamma_1, \gamma_2, \gamma_3)$ is equivalent to a projection of γ into the model. This projection, shown in Fig. 3, is defined by the vector $\tilde{p}$ in the model which minimizes

$$\text{distance}(p, \gamma) = -\sum_{j=1}^{3} \gamma_j \log p_j.$$

Roughly, the posterior distribution of the parameter θ is induced by the conditional posterior of the unconstrained vector p given that p is in the model. This is the premise for the Dirichlet sampling process (Tanner and Wong, 1987). The WLB replaces conditioning by projection, making sampling much more efficient, but introducing errors into the method.

In Fig. 4, a histogram from 5000 simulated $\tilde{\theta}$s is compared with two posterior distributions for each of the data sets in Table 1. In one case, the prior is flat. The other is $\pi(\theta) \propto \{(2+\theta)(1-\theta)\theta\}^{-1}$, which is the restriction of the prior $\Pi_j p_j^{-1}$ to vectors in the model (i.e. $p=p(\theta)$). The approximation is reasonably good and improves with increasing sample size. The case shown in Fig. 4(c) is rather interesting because the sample is small and the data indicate that θ is close to the boundary of the parameter space; this is a case where inference is particularly sensitive to the prior distribution. The WLB with uniform weights provides a close approximation to the posterior under the prior $\pi(\theta)$ and is somewhat different from the likelihood function.

Fig. 5 shows the effect of changing the weight distribution in this simple linkage example. The particular change gives the collapsed weight vector γ a density proportional to $\Pi \gamma_j^{y_j}$ (compare with equation (2)). The induced distribution of $\tilde{\theta}$ is quite close to the posterior of θ under a flat prior. It is instructive to view the density of $\tilde{\theta}$ as a product of the likelihood L and some function $\pi_{e,n}$ which we call the effective prior. This function is not a prior in the usual sense, as it may depend on the data. Ideally, a weight distribution can be found so that $\pi_{e,n}$ is close to the prior of interest π. Lacking this, retrospective adjustment of WLB samples by SIR is often successful.

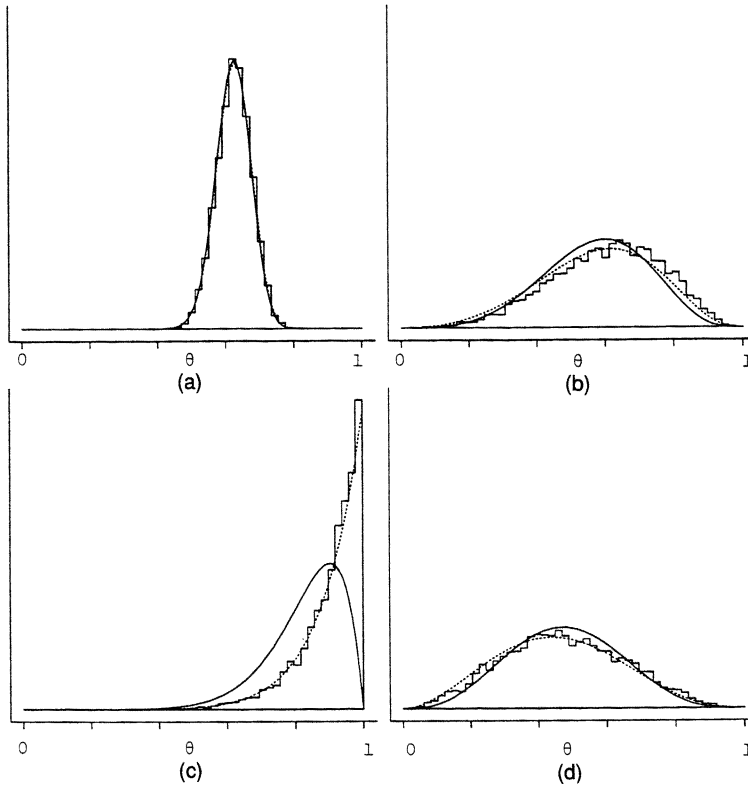

Fig. 4. Histograms (based on 5000 draws) from the WLB simulation compared with two posterior distributions for each of the four data sets in Table 1: ———, likelihood functions; ·····, posteriors under the prior $\pi(\theta) \propto 1/\theta(2+\theta)(1-\theta)$

4. ASYMPTOTIC ACCURACY

When the model satisfies sufficient regularity conditions (for details, see Newton (1991)), the asymptotic conditional distribution of $\tilde{\theta}_n$ can be studied and compared with known asymptotic properties of posterior distributions. First-order correctness of the WLB (with uniform Dirichlet weights) is embodied in the following two results, which assume independent and identically distributed data $x_1, x_2, \ldots$, from some unknown member f_{θ_0} of the model. Throughout this section, we are considering the distribution of an unadjusted WLB sample, i.e. before correction by SIR.

Let $\hat{\theta}_n$ be the maximum likelihood estimate (in R^k), and let $I_n(\hat{\theta}_n)$ be the observed information matrix, namely the $k \times k$ matrix of negative second partials of the log-likelihood times $1/n$.

Theorem 1. For each $\epsilon > 0$, as $n \to \infty$,

$$P(|\tilde{\theta}_n - \hat{\theta}_n| > \epsilon \,|\, x_1, x_2, \ldots, x_n) \to 0$$

along almost every sample path $x_1, x_2, \ldots$.

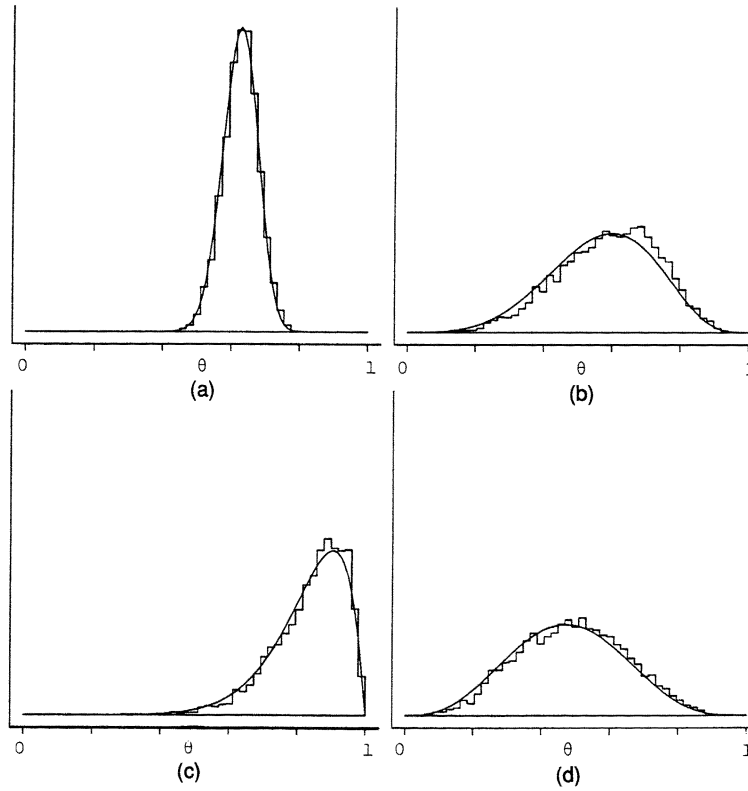

Fig. 5. Plot similar to Fig. 4, except that modified weights are used in the WLB simulation: modifying the weights makes the resulting histograms closer to the likelihood functions, i.e. it flattens the effective prior

Theorem 2. As $n \to \infty$, and for every Borel set $A \subset R^k$,

$$P(\sqrt{\{n I_n(\hat{\theta}_n)\}}(\tilde{\theta}_n - \hat{\theta}_n) \in A \mid x_1, x_2, \ldots, x_n) \to P(Z \in A)$$

along almost every sample path $x_1, x_2, \ldots$. Here, Z is a normal random vector with mean 0 and identity covariance matrix.

For both of these results, the probabilities refer to the distribution of $\tilde{\theta}_n$ induced by the random weights $w_{n,i}$. It is well known (Johnson, 1967, 1970) that the posterior distribution of $\sqrt{\{n I_n(\hat{\theta}_n)\}}(\theta - \hat{\theta}_n)$ is also asymptotically standard normal, and so the WLB is said to be first order correct. In performing the WLB, knowledge of the information matrix is not required.

In general, higher order approximations to the posterior involve the prior. Since the WLB with uniform Dirichlet weights does not use information from the prior, it is doubtful that the procedure will have good higher order properties. It is informative, however, to study higher order expansions, and we do this in the one-dimensional case. A Taylor series expansion gives

$$Z_n := \sqrt{\{n I_n(\hat{\theta}_n)\}}(\tilde{\theta}_n - \hat{\theta}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{n,i} \psi_{n,i}(\hat{\theta}_n) + R_n \tag{4}$$

where

$$\psi_{n,i}(\theta) = I_n(\theta)^{-1/2} \frac{\partial\{\log f_\theta(x_i)\}}{\partial\theta}$$

and R_n is an error term. The Edgeworth expansions of Haeusler *et al.* (1991) for weighted bootstraps can be applied to the dominant term in Z_n, under moment conditions on derivatives of the log-density. Therefore, if the error R_n is sufficiently small, the distribution function $\tilde{F}_n(t)$ of Z_n can be expressed as

$$\tilde{F}_n(t) = \Phi(t) + \frac{\phi(t)}{6\sqrt{n}} \left\{ \frac{h(\theta_0)}{I(\theta_0)^{3/2}} (t^2 - 1) \right\} + o\left(\frac{1}{\sqrt{n}}\right),$$

where ϕ and Φ are the standard normal density and distribution function respectively and

$$h(\theta) = 2E\left[\frac{\partial\{\log f_\theta(X)\}}{\partial\theta}\right]^3, \quad I(\theta) = -E\left[\frac{\partial^2\{\log f_\theta(X)\}}{\partial\theta^2}\right].$$

The errors R_n are sufficiently small if, for a sequence of positive numbers $\delta_n = o(1/\sqrt{n})$, $P(|R_n| > \delta_n | x_1, \ldots, x_n) = o(1/\sqrt{n})$ for almost every data sequence. This condition can fail, as it does for example with exponential data parameterized by $\theta = 1/E(x_i)$.

From Johnson (1970), the posterior distribution function $F_n(t)$ of $\sqrt{\{n I_n(\hat{\theta}_n)\}}(\theta - \hat{\theta}_n)$ can be expanded, for almost every data sequence, as

$$F_n(t) = \Phi(t) + \frac{\phi(t)}{6\sqrt{n}} \left\{ \frac{g(\theta_0)}{I(\theta_0)^{3/2}} (t^2 + 2) + \frac{6\dot{\pi}(\theta_0)}{\pi(\theta_0) I(\theta_0)^{1/2}} \right\} + o\left(\frac{1}{\sqrt{n}}\right)$$

where

$$g(\theta) = E\left[\frac{\partial^3\{\log f_\theta(X)\}}{\partial\theta^3}\right],$$

and π is the prior of interest with derivative $\dot{\pi}$.

The $n^{-1/2}$-terms in these two expansions are not always equal, in part because the expansion of the posterior involves the prior. However, there is a class of models for which the $n^{-1/2}$-terms in both expansions are equal when a particular prior π is being considered. To determine this class, recall a result from Bartlett (1953) that $h(\theta) = 6\dot{I}(\theta) + 4g(\theta)$, where $\dot{I}$ is the derivative of the Fisher information. Equating coefficients of t^2, we see that a necessary condition for the $n^{-1/2}$-terms to be equal is that $g(\theta) = -2\dot{I}(\theta)$. This constraint holds in several models, including those where θ is the location parameter of a symmetric distribution, and in exponential families where $\log f_\theta(x) = a(\theta)x + c(\theta) + d(x)$ and $\theta = E(X)$. We have found no model where $g(\theta) \neq -2\dot{I}(\theta)$ and where $\theta = E(X)$, and so perhaps the class is reasonably large. After equating constant terms, we have that $\tilde{F}_n(t) = F_n(t) + o(1/\sqrt{n})$ if both $g(\theta) = -2\dot{I}(\theta)$

and $\pi(\theta) \propto I(\theta)$, the square of the Jeffreys prior. So, if this prior is used in a model where $g = -2\dot{l}$, the WLB beats the normal approximation.

The prior of interest may not be the square of the Jeffreys prior, even if the model satisfies the necessary technical constraints, and so the practical use of this expansion is unclear. Again, our recommendation is retrospectively to adjust the WLB output by using a density estimate and the SIR algorithm to produce a simulation consistent estimate of the posterior density. Another approach would be to modify the weight distribution to incorporate model and prior information. No general recipe yet exists, but this idea may be workable given some results for frequentist bootstraps (Haeusler *et al.*, 1991).

As a practical matter, the quality of the WLB approximation is readily assessed by studying the stability of the importance sampling weights used in the SIR algorithm. If a small fraction of these weights dominates the others, then the density being simulated is not particularly close to the posterior of interest. By this method of assessment, we see that the WLB is quite satisfactory in the non-linear regression model studied in the first example, even though asymptotic expansions in this case have not been characterized.

The expansions of $\tilde{F}_n(t)$ and $F_n(t)$ provide insight into the structure of the effective prior introduced at the end of Section 3. The conditional density of $\tilde{\theta}$, given the data, is proportional to the likelihood L times some function $\pi_{e,n}$ which we call the effective prior. This function is not a prior in the usual sense, since it may depend on the data, but technically it plays the role of a prior by modifying the likelihood. Indeed $\pi_{e,n}$ may not equal the prior of interest π, although if the two are close then the WLB provides a good first approximation to the posterior. If the sequence of effective priors has a limit π_e with derivative $\dot{\pi}_e$, then in models where $g = -2\dot{l}$ this limit must be proportional to $I(\theta)$.

Equation (4) can also be used to study the effect of the overdispersion parameter α on the distribution of $\tilde{\theta}$. Suppose that weights $w_{n,i}$ are proportional to Y_i^{α} for some $\alpha \geqslant 1$, where Y_i are exponential. The uniform Dirichlet weights obtain for $\alpha = 1$, but the weights are more variable for larger α. Ignoring the error term in equation (4) and applying the delta method, we see that the conditional variance of $\tilde{\theta}$ is proportional to α^2, as an asymptotic approximation. One benefit of overdispersion is that attenuated regions of posterior mass are more adequately sampled. However, if α is too large, then a significant fraction of $\tilde{\theta}$s have small likelihood, causing a minority of points to dominate the importance weights. Our experience suggests that we should choose α rather close to 1. Importance weights can be compared from simulation under different α, and an empirical choice of α can be made.

5. IMPLEMENTATION AND EXAMPLES

5.1. *Iteratively Reweighted Least Squares*

Standard methods for computing maximum likelihood estimates can often be used to maximize a weighted likelihood function. The upshot of this in practice is that computer code for maximizing a likelihood can be invoked to perform the WLB simulation. One such method is iteratively reweighted least squares (IRLS) (Green, 1984).

Consider a weighted likelihood function $\tilde{L}$ (or its logarithm $\tilde{l}$) which is maximized by solving the (vector) *weighted likelihood equation*

$$\frac{\partial \tilde{l}}{\partial \theta}(\theta) = 0 \tag{5}$$

for $\tilde{\theta} \in R^k$. There is a close connection between a solution $\tilde{\theta}$ of equation (5) and the IRLS solution $\hat{\theta}$ of the likelihood equation

$$\frac{\partial l}{\partial \theta}(\theta) = 0,$$

where l is the logarithm of the likelihood function.

Following the general formulation described in Green (1984), $\tilde{l}$ is viewed as a function of an n-vector of predictors $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^{\mathrm{T}}$. These predictors, in turn, are viewed as functions of the parameter θ; thus $\eta = \eta(\theta)$. Letting u be the n-vector $(\partial l/\partial \eta)$ and D the $n \times k$ matrix $(\partial \eta/\partial \theta)$, the weighted likelihood equation (5) becomes simply

$$D^{\mathrm{T}} W u = 0,$$

where W is an $n \times n$ diagonal matrix with weights $w_{n,i}$ on its diagonal. The model densities are assumed to have the form $f_i(x_i; \theta) = \psi_i(\eta_i, x_i)$ where, for each i, ψ_i is a fixed, known, function determined by the model.

The iterative Newton–Raphson solution to the weighted likelihood equation is first to evaluate u, D and the second derivatives of $\tilde{l}$ at an initial guess $\tilde{\theta}_0$. Then an updated guess $\tilde{\theta}_1$ is computed by solving the linear system

$$-\frac{\partial^2 \tilde{l}}{\partial \theta \partial \theta^{\mathrm{T}}}(\tilde{\theta}_1 - \tilde{\theta}_0) = D^{\mathrm{T}} W u. \tag{6}$$

Iteration continues until convergence. In the standard Fisher scoring or IRLS solution, however, the matrix $(-\partial^2 \tilde{l}/\partial \theta \partial \theta^{\mathrm{T}})$ in equation (6) is replaced by an approximation $D^{\mathrm{T}} W A D$ where A is the expectation (under the current parameter value) of the $n \times n$ matrix $(-\partial^2 l/\partial \eta \eta^{\mathrm{T}})$. This approximation is derived from the expansion

$$\frac{\partial^2 \tilde{l}}{\partial \theta \partial \theta^{\mathrm{T}}} = D^{\mathrm{T}} W \frac{\partial^2 l}{\partial \eta \eta^{\mathrm{T}}} D + \sum_{i=1}^{n} w_{n,i} \frac{\partial l}{\partial \eta_i} \frac{\partial^2 \eta_i}{\partial \theta \partial \theta^{\mathrm{T}}}$$

and the fact that $E(\partial l/\partial \eta_i) = 0$. With this approximation, the Newton–Raphson algorithm involves evaluating u, D and A at an initial value $\tilde{\theta}_0$ and then solving the linear system

$$D^{\mathrm{T}} W A D(\tilde{\theta}_1 - \tilde{\theta}_0) = D^{\mathrm{T}} W u \tag{7}$$

for $\tilde{\theta}_1$. Again, iteration continues until convergence. We must assume that D is of full rank k and A is positive definite to ensure a unique solution at each iteration. By noting that equation (7) defines the normal equations for a regression problem, we can compute $\tilde{\theta}_1$ by regressing $A^{-1}u + D\tilde{\theta}_0$ on D with weight matrix WA, i.e.

$$\tilde{\theta}_1 = (D^{\mathrm{T}} W A D)^{-1} D^{\mathrm{T}} W A (A^{-1} u + D\tilde{\theta}_0).$$

The WLB simulation involves repeatedly generating weight matrices W and then performing the IRLS algorithm described above. By comparison, if W is the identity matrix, we have the standard algorithm to solve the likelihood equations. Also, the form of the estimating equations indicates that the diagonal entries of W need to be known only up to a constant of proportionality, so that unnormalized exponentials (or powered exponentials) may appear on the diagonal. By the addition of a random weight vector into a standard optimization routine, the WLB is readily implemented.

5.2. *Non-linear Regression*

The non-linear regression model discussed in Section 2 is illustrative because with two regression parameters the contours of the WLB density estimate are easily visualized. That the WLB works in a two-parameter model is not enough, however, to convince many statisticians of its general utility. In this section, we study a second non-linear regression model, this one having four regression parameters. Again, inference is quite straightforward with the SIR-adjusted WLB.

Abdollah (1986) and later Bates and Watts (1988) consider the following model for a problem from biochemistry:

$$x_i = \beta_1 + \frac{\beta_2}{1 + \exp\{-\beta_4(t_i - \beta_3)\}} + \epsilon_i \qquad i = 1, 2, \ldots, n.$$

The problem is concerned with the amount x_i of one chemical that binds to the surface of particular cells in the presence of a certain amount t_i of another chemical. We study this model by using data from the second tissue sample in Table A4.2 of Bates and Watts (1988). There are 16 observations, and four regression parameters, each with meaningful interpretations on the given scale. A scale parameter σ of the independent normal errors augments the regression parameters to give five unknown parameters in θ.

Bayesian inference for this problem starts with a prior. As with the example in Section 2, we put $\pi(\sigma^2) \propto \sigma^{-2}$ and $\pi(\beta) \propto |V^T V|^{1/2}$. The prior on β is locally uniform and transformation invariant. In all that follows, σ has been integrated out analytically. Using the prescription laid out in Section 2, we ran the WLB by generating 5000 vectors of exponential random variables and then applied an optimization routine with each to determine that many $\tilde{\beta}$s. In this example, the exponentials are powered up by $\alpha = 1.1$ to spread out the weights. For 147 of these vectors, no solution to the weighted likelihood equation was found, owing to ill conditioning of some kind. A normal-kernel density estimate was fitted to the remaining 4853, taking as its covariance matrix a scaled-down version of the inverse Hessian computed at the maximum likelihood estimate. We computed the scale factor in accordance with the principle of maximal smoothing (Terrell, 1990). Next, importance weights were computed by comparing the density estimate at each $\tilde{\beta}$ with the true posterior density (up to a constant) which is given by the known marginal likelihood and the given prior. These weights, when normalized to be a probability vector, are shown in Fig. 6(a). To obtain this plot, we sort the weights and plot their cumulative value against their rank. Perfectly uniform weights would fall on a straight line between the end points, and so we see that the weights in this case are quite stable. The heaviest 23% carry half the weight. By contrast, SIR weights based on the best fitting normal are such that 8% of the heaviest

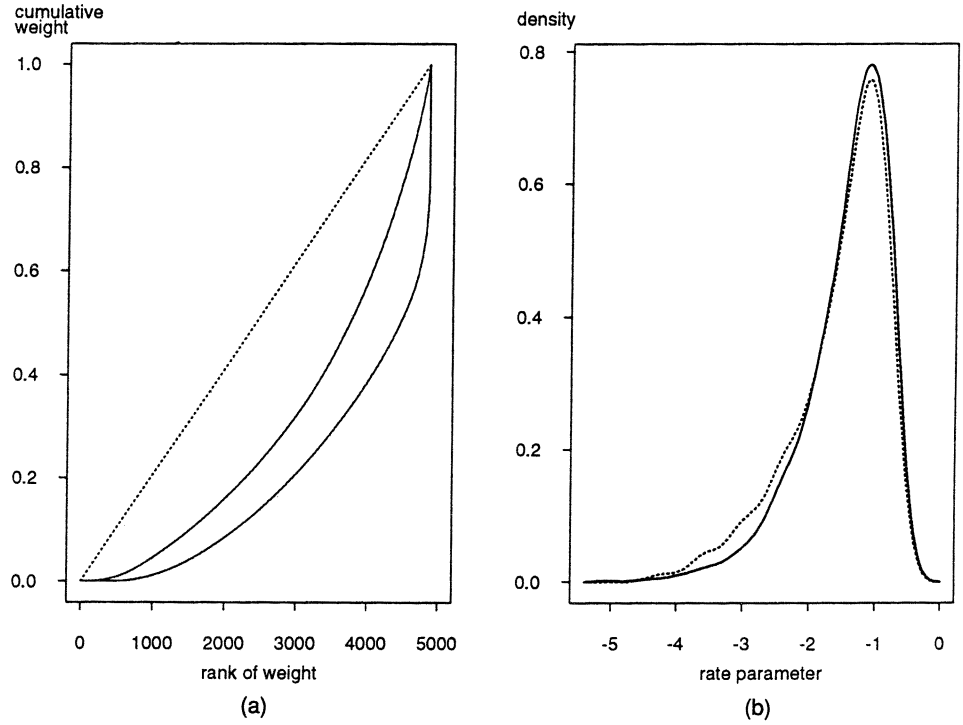

Fig. 6. Non-linear regression of Section 5.2: (a) importance weights (·····, perfectly uniform weights; the nearer curve gives the cumulative importance weights based on the WLB density estimate; the second, and much steeper, curve gives the cumulative weights based on a normal approximation to the posterior); (b) estimates of the marginal posterior for β_4 (———, density estimate based on 10 000 resamples in an SIR-adjusted WLB; ·····, similar estimate after two iterations of adaptive importance sampling have been applied to the initial kernel density estimate)

weights carry half the total, and this has been observed in several other non-linear regression models.

Marginal posterior inference for a parameter of interest follows immediately by treating the importance weights and the sampled $\tilde{\beta}$s as a discrete probability distribution. Fig. 6(b) shows an estimate of the marginal posterior distribution of the rate parameter β_4, produced by forming a density estimate from 10000 draws of the discrete distribution mentioned above. The true marginal is quite difficult to compute, and it is not shown in Fig. 6. However, by applying two iterations of adaptive importance sampling (West, 1992) to our estimate, we produce a better estimate of the true marginal. This second estimate, shown as a dotted line in Fig. 6(b), differs slightly from the SIR-adjusted WLB estimate in the mass that it assigns to the left-hand shoulder of the marginal.

The implementation of this entire analysis was done with a relatively small amount of S code. No special programs are required except estimation routines that already exist. The result is a simulation consistent estimate of the marginal posterior density of interest.

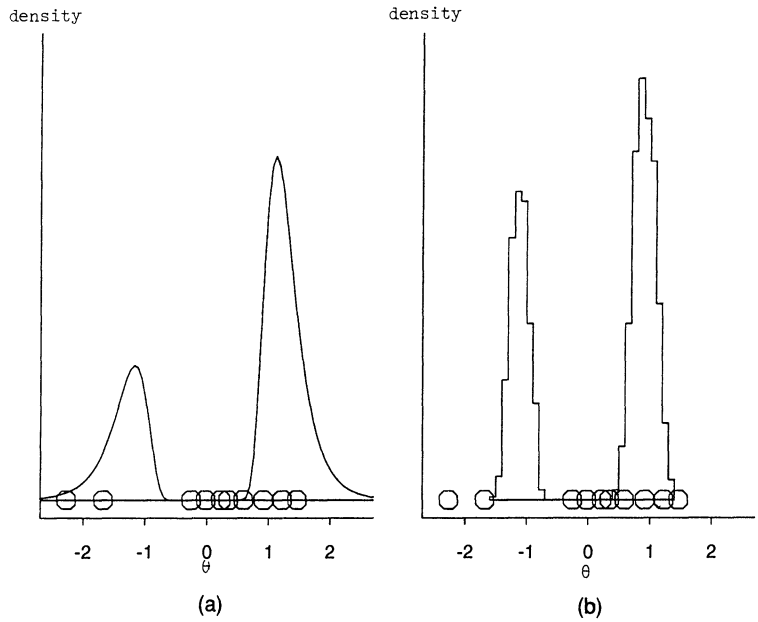

Fig. 7. (a) Normalized likelihood function for θ given the data (○) when the model is normal with mean θ and variance θ^2; (b) histogram from 1000 WLB draws

5.3. *Bimodal Posterior*

The asymptotic justification of the WLB method hinges on the approximate normality of the likelihood function. However, the two non-linear regression examples show that the WLB can capture more than just the approximating quadratic structure of this function. Fig. 7 shows another instance of this, where the WLB is applied to a model for constant coefficient of variation (Hinkley (1977), for example). The bimodal posterior is well approximated by the raw WLB sample.

5.4. *Generalized Linear Model*

Chambers and Hastie (1992) illustrate the S language by using data from an experiment to study mounting of electronic components to printed circuit boards (the solder.balance data set). A response is measured for each of 720 experimental conditions. This response x_i is a count, and a Poisson regression model is considered by Chambers and Hastie to explain structure in $\lambda_i = E(x_i)$:

$$\log \lambda_i = z_i^t \theta,$$

where z_i indicates the experimental conditions leading to x_i. With five factors, the main effects model has 18 regression coefficients. Bayesian inference under a flat prior was done with the WLB. Our analysis simply illustrates that the WLB has potential in moderately high dimensional problems. We do not address the scientific questions.

Several choices of dispersion parameter were considered, and, surprisingly, a value of $\alpha = 0.7$ proved most successful. These underdispersed weights projected the rather small sample of 2000 $\tilde{\theta}$s into the heart of the posterior mass. As in the other regression examples, a kernel density estimate combined with the likelihood produced

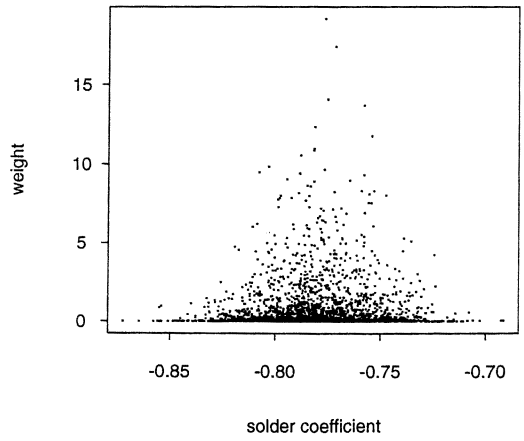

Fig. 8. Importance weights plotted against one dimension of the $\tilde{\theta}$-vector: this coefficient represents the effect of solder thickness (a two-level factor) on the response (maximum likelihood estimate, -0.78; importance weights scaled to have mean 1); highly weighted points tend to correspond to the centre of this marginal distribution and not to the tails, indicating that the WLB has not missed the tails of this 18-dimensional distribution

importance weights. Fig. 8 is a plot of these weights against one of the model coefficients. The largest 11% of the weights carry half the total mass, but we see from Fig. 8 (and other marginal plots not shown) that these large weights tend to be in the centre of the $\tilde{\theta}$s. The WLB has not missed the 'tails' of this 18-dimensional distribution.

6. EXTENSIONS OF WEIGHTED LIKELIHOOD BOOTSTRAP

6.1. *Dependent Data*

The definition of weighted likelihood (1) can be extended to models for dependent data as

$$\tilde{L}_n(\theta) := \prod_{i=1}^{n} f_\theta(x_i | x_1^{i-1})^{w_{n,i}}, \tag{8}$$

where $x_1^{i-1} = (x_1, x_2, \ldots, x_{i-1})$, and the factors in the product are the conditional densities of x_i given x_1^{i-1}. Different orderings of the data yield different weighted likelihood functions, although for time series there is the natural time ordering which we use below in two examples.

By analogy with the multinomial model of Section 3, maximization of $\tilde{L}$ in equation (8) simulates the posterior distribution of a transition probability matrix when $x_1, \ldots, x_n$ form a Markov chain on k states. In particular, this posterior comes from the square of the Jeffreys prior when there are no modelling constraints on the transition probabilities.

As a second example, consider simulating the predictive distribution of the future of a time series. Generally, such a distribution is a mixture of parameterized densities with respect to the posterior distribution of a parameter. This simulation is a two-step process. First, a parameter is simulated from its posterior distribution, and

Fig. 9. Prediction intervals for a US economic time series: annual change in log(money stock) from 1947 to 1970 with two sets of 95% prediction intervals (———, produced by the WLB; ·····, based on Box–Jenkins methodology); in performing the WLB, we first simulate AR(1) parameters by weighted regression; error variances are subsequently sampled from their posterior, and then finally future data sequences are sampled (the full curves represent upper and lower quantiles of 1000 simulated futures at each of seven time points; estimated AR parameter, 0.65; estimated prediction variance, 0.00035); diagnostics, including a check of the prediction residuals and a Portmanteau test (Box and Pierce, 1970) suggest that the AR(1) model is adequate; for the seven-period-ahead prediction, the predictive variance from the WLB analysis is about 70% greater than that from the standard Box–Jenkins method

then a future is simulated from its predictive distribution given the parameter. The WLB provides a simple approximation to the first step of this process. A typical economic time series modelled by Box–Jenkins methods is shown in Fig. 9 (from Nelson and Plosser (1982)). The data are yearly averages after World War II and consequently form fairly short series. Comparison with the standard non-Bayesian prediction intervals shows that there is considerable uncertainty about the autoregressive AR(1) parameters. Here, the WLB amounts to repeated weighted autoregressions.

The WLB applied to AR(p) models is essentially a special case of Künsch's (1989) blockwise bootstrap, and it inherits first-order correctness from Künsch's results. (Refer to section 5.2 of his paper, and use second-stage block size $l(n)=1$.) The WLB is based on Dirichlet weights, whereas the blockwise bootstrap involves multinomial weights. As noted by Rubin (1981), these two weight distributions have similar first- and second-moment properties, and so the resulting bootstraps will share first-order asymptotics. Of course, Künsch's method, by allowing a changing second-stage block size $l(n)$, does not rely on the correctness of the AR(p) model.

6.2. *Weighted Partial Likelihood*

For certain complex models, Cox (1975) introduced a factorization of the likelihood function into two parts. One part provides little information about the parameter θ of interest whereas the other part, the partial likelihood, does not depend on the

nuisance parameter and so is used in inference about θ. The WLB has a natural analogue for partial likelihood.

To construct a partial likelihood, Cox (1975) transforms data x into a sequence

$$(u_1, v_1, u_2, v_2, \ldots, u_n, v_n)$$

and then forms a particular partition of the full likelihood function:

$$L(\theta, \psi) = \prod_{i=1}^{n} f_{\theta,\psi}(u_i | u_1^{i-1}, v_1^{i-1}) \prod_{i=1}^{n} f_\theta(v_i | u_1^i, v_1^{i-1}),$$

the second product being the partial likelihood for θ based on (v_i) in the sequence (u_i, v_i). A randomly weighted full likelihood leads to a randomly weighted partial likelihood

$$\tilde{L}_P(\theta) = \prod_{i=1}^{n} f_\theta(v_i | u_1^i, v_1^{i-1})^{w_{n,i}}$$

and thus to a procedure for sampling values in the parameter space. In Newton (1991), this new bootstrap method is studied for Cox's proportional hazards model of survival analysis. It is shown that the conditional distribution of $\tilde{\theta}_n$ is asymptotically normal with the same variance as the partial likelihood estimator, at least in a simple case of this model. Because risk sets in each factor of this weighted partial likelihood never change, this WLB procedure is very different from Efron's (1979) nonparametric bootstrap.

7. USING SAMPLES FROM POSTERIOR TO EVALUATE THE MARGINAL LIKELIHOOD

Suppose that we wish to compare two models M_0 and M_1 (not necessarily nested) by using the Bayes factor, or ratio of posterior to prior odds,

$$B_{01} = \frac{p(x | M_0)}{p(x | M_1)}. \tag{9}$$

In equation (9),

$$p(x | M_j) = \int p(x | \theta_j, M_j)\, p(\theta_j | M_j)\, d\theta_j, \tag{10}$$

where θ_j is the (possibly vector) parameter of model M_j and $p(\theta_j | M_j)$ is the prior density of θ_j under model M_j ($j = 0, 1$). We call $p(x | M_j)$ the *marginal likelihood* of the data under model M_j. The integral in equation (10) is difficult to evaluate in general, especially when the dimension of θ_j is large. Exact or approximate analytical results are available for some specific models with particular classes of priors, such as linear models (Spiegelhalter and Smith (1982), and references therein), models arising in multivariate analysis (Smith and Spiegelhalter, 1980), log-linear models (Raftery, 1986), generalized linear models (Raftery, 1988a), general Poisson processes (Akman and Raftery, 1986), changepoint problems (Raftery and Akman, 1986) and software reliability models (Raftery, 1987, 1988b).

If one is interested in several models $\{M_j: j=0, 1, \ldots, J\}$, then Bayesian inference, prediction and decision-making all involve their posterior probabilities

$$p(M_j|x)=p(x|M_j)\,p(M_j)\Big/\sum_{k=0}^{J} p(x|M_k)\,p(M_k),$$

where $p(M_j)$ is the prior probability of the model M_j. Again, the marginal likelihoods $p(x|M_j)$ are the crucial components.

Dropping the notational dependence on M_j, equation (10) becomes

$$p(x)=\int p(x|\theta)\,p(\theta)\,\mathrm{d}\theta. \tag{11}$$

The Monte Carlo method for evaluating integrals of the form $I=\int g(\theta)\,p(\theta)\,\mathrm{d}\theta$ is to generate a sample $\{\theta^{(i)}: i=1, \ldots, m\}$ from a density $p^*(\theta)$. Under quite general conditions, a simulation consistent estimate of I is

$$\hat{I}=\sum_{i=1}^{m} w_i\,g(\theta^{(i)})\Big/\sum_{i=1}^{m} w_i, \tag{12}$$

where $w_i=p(\theta^{(i)})/p^*(\theta^{(i)})$; the function $p^*(\theta)$ is known as the *importance sampling function*.

The WLB gives us a sample approximately drawn from the posterior density $p^*(\theta)=p(\theta|x)=p(x|\theta)\,p(\theta)/p(x)$. For most functions $g(\theta)$ this would be a poor importance sampling function, but here we have $g(\theta)=p(x|\theta)$, and the importance sampling function $p^*(\theta)=p(\theta|x)$ is well suited for this case. Substituting into equation (12) yields, as an estimate for $p(x)$,

$$\hat{p}_1(x)=\left\{\frac{1}{m}\sum_{i=1}^{m} p(x|\theta^{(i)})^{-1}\right\}^{-1}, \tag{13}$$

the harmonic mean of the likelihood values. Thus, the marginal likelihood may be estimated by the harmonic mean of the likelihoods of a sample from the posterior distribution. This is true whether the posterior samples come from an SIR-adjusted WLB or any other sampling scheme, like the Markov chain Monte Carlo method.

It is readily verified that $\hat{p}_1(x)$ converges almost surely to the correct value $p(x)$ as $m\to\infty$. However, $\hat{p}_1(x)$ does not, in general, satisfy a Gaussian central limit theorem. This manifests itself by the occasional occurrence of a value of $\theta^{(i)}$ with a small likelihood and hence a large effect on the final result; it happens because $p(x|\theta)^{-1}$ is often not square integrable with respect to the posterior distribution.

An alternative to equation (13) is

$$\hat{p}_2(x)=\frac{1}{m}\sum_{i=1}^{m} p(x|\theta^{(i)}), \tag{14}$$

where $\{\theta^{(i)}: i=1, \ldots, m\}$ is a sample from the *prior* distribution rather than the posterior. This possibility was mentioned by Raftery and Banfield (1990) and was investigated in detail in particular cases by McCulloch and Rossi (1991). A major difficulty with $\hat{p}_2(x)$ is that most of the $\theta^{(i)}$ will have small likelihood values if the posterior is concentrated relative to the prior, so that the simulation process will be

quite inefficient. Thus the estimate will be dominated by a few large values of the likelihood, and so the variance of $\hat{p}_2(x)$ may be large and its convergence to a Gaussian distribution slow. These problems were apparent in the examples studied in detail by McCulloch and Rossi (1991); they are precisely the opposite of the difficulties with $\hat{p}_1(x)$.

These considerations suggest that we use as importance sampling function a *mixture* of the prior and posterior densities, $p^*(\theta) = \delta p(\theta) + (1-\delta) p(\theta|x)$, where δ is small. This yields a new estimate $\hat{p}_3(x)$, defined by the equation

$$\hat{p}_3(x) = \frac{\sum_{i=1}^{m} p(x|\theta^{(i)}) / \{\delta \hat{p}_3(x) + (1-\delta) p(x|\theta^{(i)})\}}{\sum_{i=1}^{m} \{\delta \hat{p}_3(x) + (1-\delta) p(x|\theta^{(i)})\}^{-1}}. \tag{15}$$

The estimator $\hat{p}_3(x)$ is appealing because it retains the efficiency of $\hat{p}_1(x)$, due to being based mostly on high likelihood values of θ, but avoids its unpleasant instability. It is readily verified that $\hat{p}_3(x)$ *does* satisfy a Gaussian central limit theorem, unlike $\hat{p}_1(x)$. However, $\hat{p}_3(x)$ has the irksome aspect that we must simulate from the prior as well as the posterior.

Simulation from the prior as well as the posterior may be avoided, without sacrificing the appealing aspects of $\hat{p}_3(x)$, by instead simulating all m values from the posterior distribution and *imagining* that a further $\delta_m/(1-\delta)$ values of θ are drawn from the prior, all with likelihoods $p(x|\theta^{(i)})$ equal to their expected value $p(x)$. This yields an approximation to $\hat{p}_3(x)$, namely

$$\hat{p}_4(x) = \frac{\delta m/(1-\delta) + \sum_{i=1}^{m} p(x|\theta^{(i)}) / \{\delta \hat{p}_4(x) + (1-\delta) p(x|\theta^{(i)})\}}{\delta m/(1-\delta)\, \hat{p}_4(x) + \sum_{i=1}^{m} \{\delta \hat{p}_4(x) + (1-\delta) p(x|\theta^{(i)})\}^{-1}}. \tag{16}$$

The estimator $\hat{p}_4(x)$ may be evaluated by using a simple and obvious iterative scheme; in our limited experience to date, this converges fast, often in a single step. In some small-scale numerical experiments, $\hat{p}_4(x)$ performed well for δ as small as 0.01 and did not display any of the instability of $\hat{p}_1(x)$.

The harmonic mean estimator $\hat{p}_1(x)$ is slightly reminiscent of Good's (1958) proposal for combining tests by taking the harmonic mean of the corresponding P-values; his argument was based on an analogy with Bayes factors. Also, the *arithmetic* mean of the likelihoods of a sample from the posterior is an unbiased estimator of the posterior mean of the likelihood function, $\int p(x|\theta) p(\theta|x)\, d\theta$, that underlies the 'posterior Bayes factors' of Aitkin (1991); it will typically be larger than $p(x)$. We share the misgivings of many discussants of Aitkin's paper about the interpretation of the posterior Bayes factors, but it is at least worth noting that they can be readily evaluated by using the WLB or Markov chain simulation methods.

8. DISCUSSION

We have introduced a bootstrap-like procedure for simulating approximately from a posterior distribution. For the generic weighting scheme (uniform Dirichlet weights)

the WLB is first order correct, and thus consistently estimates the mean and covariance structure of the posterior distribution. Higher order correctness is generally not available for these simple weights, although expansions elucidate the nature of the error. In practice, inaccuracies in the WLB simulation can be removed by combining the method with SIR and density estimation. This allows posterior simulation under particular priors of interest, and, as shown in a non-linear regression example, it is quite straightforward to implement. The WLB might be profitably combined with other importance sampling schemes, like the adaptive approximation method described in West (1992). Here, the WLB sample could provide a first approximation in cases where the normal or t first approximations are poor or unavailable. This is precisely how the dotted line approximation is produced in Fig. 6(b).

The WLB is similar to nonparametric bootstrapping of the maximum likelihood estimate, which is equivalent to applying the WLB with weights $m_n = (m_{n,1}, m_{n,2}, \ldots, m_{n,n})$ representing cell counts after classifying n objects randomly into n equally likely cells. As noted by Rubin (1981), these multinomial weights have similar first- and second-moment properties to the Dirichlet weights, and hence the simulated maximizers have distributions which are both first order correct. As shown in Weng (1989), second-order properties of the two weighting schemes are different. Here we are trying to simulate a posterior distribution, rather than a sampling distribution. The Dirichlet weights form the natural basis of a Bayesian simulation method because of their properties for multinomial data. However, except for unconstrained multinomials and Markov chains, and the class of models described in Section 4, the uniform Dirichlet weighting scheme is not second order correct. The Dirichlet weights are somewhat more convenient computationally, as they never equal 0, and often only require the generation of unnormalized exponential random variables. We are currently studying the possibility of having a simple recipe for specifying a good distribution for the weights which uses information in the prior and model structure. Many weighting schemes ensure first-order correctness (see Mason and Newton (1992)).

Other researchers have studied the use of bootstrapping for Bayesian inference. Boos and Monahan (1986) have studied the use of Efron's bootstrap to approximate a posterior through the sampling distribution of a pivot, and this line of research was developed more fully in Hall (1987). Zheng and Tu (1988) and references therein have also studied the use of weighted bootstraps to simulate pivotal distributions. Laird and Louis (1987) used a bootstrap sampling distribution to approximate a posterior in an empirical Bayes setting. More recently, Davison *et al.* (1992) have constructed a completely nonparametric likelihood based on bootstrapping. The WLB is somewhat different from this work, primarily because it is designed to solve a parametric problem approximately, but in its extension to semiparametric models there may be some deeper connections with other bootstrap methods.

On other connections—if the likelihood (of a distribution function) is defined nonparametrically as the joint probability assigned to the observed data under that distribution, then the WLB with uniform Dirichlet weights is exactly the same as the Bayesian bootstrap mentioned earlier. Weighting that same likelihood with multinomial weights gives Efron's (1979) nonparametric bootstrap. Thus randomly weighting the components of a likelihood is a unifying idea.

The WLB is easy to program for particular applications by using existing built-in functions in standard statistical languages and packages. However, as an example,

we have made available S code to perform the calculations for the Poisson regression model of Section 5.4; this can be used as a template and modified for other models. It is available by electronic mail from Statlib at no cost. Send a message to statlib@stat.cmu.edu containing the single line 'send wlb from S'.

ACKNOWLEDGEMENTS

This research was supported by Office of Naval Research contracts N-00014-88-K-0265 and N-00014-91-J-1074 and by the Applied Physics Laboratory, University of Washington. Earlier versions of this paper were presented at the Bootstrap Workshop, University of Washington, May 1988, the Statistical Society of Canada meetings in Victoria, British Columbia, in 1988 and St John's, Newfoundland, in 1990, and the Workshop on Bayesian Computations via Stochastic Simulation, Columbus, Ohio, February 1991. The authors are grateful to many people, including Douglas Bates, Charles Geyer, Hans Künsch, David Mason, Charles Nelson, Don Rubin, Jon Wellner, and no fewer than six referees for very helpful comments and discussions.

REFERENCES

Abdollah, S. (1986) The effect of doxorubicin on the specific binding of [^3H] nitrendipine to rat heart microsomes. *Master's Thesis*. Queen's University, Kingston.

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *J. R. Statist. Soc.* B, **53**, 111–142.

Akman, V. E. and Raftery, A. E. (1986) Bayes factors for non-homogeneous Poisson processes with vague prior information. *J. R. Statist. Soc.* B, **48**, 322–329.

Bartlett, M. S. (1953) Approximate confidence intervals. *Biometrika*, **40**, 12–19.

Bates, D. M. and Chambers, J. M. (1992) Nonlinear models. In *Statistical Models in S* (eds J. M. Chambers and T. J. Hastie). Pacific Grove: Wadsworth and Brooks/Cole.

Bates, D. M. and Watts, D. G. (1988) *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Boos, D. D. and Monahan, J. F. (1986) Bootstrap methods using prior information. *Biometrika,* **73**, 77–83.

Box, G. E. P. and Pierce, D. A. (1970) Distribution of residual autocorrelations in autoregressive–integrated moving average time series models. *J. Am. Statist. Ass.*, **65**, 1509–1526.

Chambers, J. M. and Hastie, T. J. (1992) *Statistical Models in S*. Pacific Grove: Wadsworth and Brooks/Cole.

Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.

Davison, A. C., Hinkley, D. V. and Worton, B. J. (1992) Bootstrap likelihoods. *Biometrika*, **79**, 113–130.

Dempster A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal posterior densities. *J. Am. Statist. Ass.*, **85**, 398–409.

——(1992) Bayesian statistics without tears: a sampling–resampling perspective. *Am. Statistn*, **46**, 84–88.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Good, I. J. (1958) Significance tests in parallel and in series. *Ann. Math. Statist.*, **53**, 799–813.

Green, P. J. (1984) Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). *J. R. Statist. Soc.* B, **46**, 149–192.

Haëusler, E., Mason, D. M. and Newton, M. A. (1991) Weighted bootstrapping of means. *Cent. Wisk. Inf. Q.*, **4**, 213–228.

Hall, P. (1987) On the bootstrap and likelihood based confidence regions. *Biometrika*, **74**, 481–493.
Hinkley, D. V. (1977) Conditional inference about a normal mean with known coefficient of variation. *Biometrika*, **64**, 105–108.
Johnson, R. A. (1967) An asymptotic expansion for posterior distributions. *Ann. Math. Statist.*, **38**, 1899–1907.
——(1970) Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.*, **41**, 851–864.
Künsch, H. R. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, **17**, 1217–1241.
Laird, N. M. and Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples. *J. Am. Statist. Ass.*, **82**, 739–757.
Marske, D. (1967) Biochemical oxygen demand data interpretation using sum of squares surface. *MS Thesis*. University of Wisconsin, Madison.
Mason, D. M. and Newton, M. A. (1992) A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, **20**, 1611–1624.
McCulloch, R. E. and Rossi, P. E. (1991) Bayes factors for nonlinear hypotheses and likelihood distributions. *Technical Report 101*. Statistics Research Center, University of Chicago Graduate School of Business, Chicago.
Nelson, C. R. and Plosser, C. I. (1982) Trends and random walks in macroeconomic time series: some evidence and implications. *J. Monet. Econ.*, **10**, 139–162.
Newton, M. A. (1991) The weighted likelihood bootstrap and an algorithm for prepivoting. *PhD Dissertation*. Department of Statistics, University of Washington, Seattle.
Raftery, A. E. (1986) A note on Bayes factors for log-linear contingency table models with vague prior information. *J. R. Statist. Soc.* B, **48**, 249–250.
——(1987) Inference and prediction for a general order statistic model with unknown population size. *J. Am. Statist. Ass.*, **82**, 1163–1168.
——(1988a) Approximate Bayes factors for generalized linear models. *Technical Report 121*. Department of Statistics, University of Washington, Seattle.
——(1988b) Analysis of a simple debugging model. *Appl. Statist.*, **37**, 12–22.
Raftery, A. E. and Akman, V. E. (1986) Bayesian analysis of a Poisson process with a change-point. *Biometrika*, **73**, 85–89.
Raftery, A. E. and Banfield, J. D. (1990) Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 32–43.
Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, p. 369. New York: Wiley.
Ritter, C., Bisgaard, S. and Bates, D. (1991) A comparison of approaches to inference for nonlinear models. In *Proc. 23rd Symp. Interface* (ed. E. M. Keramidas), pp. 148–155. Fairfax Station: Interface Foundation.
Ritter, C. and Tanner, M. A. (1992) Facilitating the Gibbs sampler: the Gibbs stopper and the griddy Gibbs sampler. *J. Am. Statist. Ass.*, **87**, 861–868.
Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
——(1987) Comment on "The calculation of posterior distributions by data augmentation", by M. A. Tanner and W. H. Wong. *J. Am. Statist. Ass.*, **82**, 543–546.
——(1988) Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 395–402. Oxford: Oxford University Press.
Smith, A. F. M. and Spiegelhalter, D. J. (1980) Bayesian approaches to multivariate structure. In *Interpreting Multivariate Data* (ed. V. Barnett). Chichester: Wiley.
Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *J. R. Statist. Soc.* B, **44**, 377–387.
Tanner, M. and Wong, W. (1987) The calculation of posterior densities by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528–550.
Terrell, G. R. (1990) The maximal smoothing principle in density estimation. *J. Am. Statist. Ass.*, **85**, 470–477.
Tierney, L. (1991) Markov chains for exploring posterior distributions. *Technical Report 560*. School of Statistics, University of Minnesota, Minneapolis.
Weng, C. S. (1989) On a second-order asymptotic property of the Bayesian bootstrap. *Ann. Statist.*, **17**, 705–710.

West, M. (1992) Modelling with mixtures. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.
Zheng, Z. and Tu, D. (1988) Random weighting method in regression models. *Sci. Sin.* A, **31**, 1442–1459.

DISCUSSION OF THE PAPER BY NEWTON AND RAFTERY

W. R. Gilks (Medical Research Council Biostatistics Unit, Cambridge): The authors propose the weighted likelihood bootstrap (WLB) for sampling from posterior distributions. This involves the following steps:

(a) sample weight vectors $\{w\}$ of length n from the uniform Dirichlet distribution; optionally power up the weights;
(b) for each w obtain $\tilde{\theta}$ maximizing the weighted likelihood in equation (1);
(c) calculate a kernel density estimate $\hat{g}(\theta)$ at each $\tilde{\theta}$;
(d) resample the $\{\tilde{\theta}\}$ using importance weights $\pi(\tilde{\theta})\, L(\tilde{\theta})/\hat{g}(\tilde{\theta})$,

where $\pi(\ )$ denotes the prior and $L(\ )$ denotes the likelihood. (The uniform Dirichlet distribution is given by equation (2) with $y_j = 1$ for all j.) Steps (c) and (d) can be iterated (adaptive importance sampling). The first two steps are a generalization of Rubin's (1981) Bayesian bootstrap. In a special case (see Section 3) steps (a) and (b) produce samples from the required posterior, but in general they produce samples from a posterior which corresponds to a data-dependent prior (the 'effective' prior), i.e. to no prior at all. The sampling–importance resampling (SIR) embodied in steps (c) and (d) is needed to rescue the method.

I would like to make the following comments.

Importance weights

SIR is a fragile method: if importance weights are very variable then an enormous initial sample will be needed for adequate richness in the resampled sample. Thus, if the WLB is to be useful, steps (a) and (b) must provide a reasonably accurate approximation to the posterior. Some reassurance is provided through the asymptotic arguments in Section 4, and examples with extremely non-quadratic likelihoods demonstrate surprisingly well-behaved importance weight distributions. However, it is clear that general reassurance cannot be given since steps (a) and (b) take no account of the prior. In the absence of clear guidelines about when the WLB is likely to behave well, one can only be advised to try it and see.

Model complexity

The authors provide asymptotic results for the following class of models (in the notation of Section 2):

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \tag{17}$$

This class is large, and the extension to dependent data (8) enlarges this class still further. However, having grown accustomed to the scope offered by Markov chain Monte Carlo (MCMC) methodology, many Bayesians would find equation (17) uncomfortably restrictive. For example, in modelling longitudinal data, each individual is usually assigned a (set of) random effect(s). In a trivial sense, random effects models are included in equation (17) since the vector of random effects can be included in θ. However, the asymptotic results will then no longer hold because the length of θ will depend on n. In some situations it might be possible to remove the random effects by integrating them out analytically, as in restricted maximum likelihood estimation (Laird and Ware, 1982), but in most situations this will not be possible. If the number of individuals (n) and the numbers of observations on each individual (m_i) are all large, then it should be possible to provide some asymptotic justification for a WLB based on the following weighted likelihood function:

$$\tilde{L}(\theta, \{\beta_i\}) = \prod_{i=1}^{n} \left\{ h(\beta_i \mid \theta)^{n w_i} \prod_{j=1}^{m_i} f(x_{ij} \mid \beta_i)^{m_i v_{ij}} \right\} \tag{18}$$

where the $n+1$ weight vectors $w = \{w_i\}$ and $v_i = \{v_{ij}\}$ are independently uniform Dirichlet distributed. Here β_i denotes a (vector of) random effects for individual i; $h(\ )$ denotes the population distribution of the random effects and $\{x_{ij}; j = 1, \ldots, m_i\}$ denotes the observations on individual i. Models of still greater generality are often required (graphical models) and can often be fitted by using MCMC methods

(Gilks *et al.*, 1993). It is not clear whether the WLB (perhaps further generalized along the lines of equation (18)) would be able to handle such situations.

Number of parameters

As model complexity increases, so too will the number of parameters. The authors have shown that the WLB can perform well with up to 20 parameters, but MCMC applications in which I am routinely involved typically contain hundreds, even thousands, of parameters. Letting k denote the number of parameters, obtaining adequate approximations to first and second moments of the posterior will involve $O(k^2)$ sampled weight vectors in step (a) above, and calculation of the matrix of second derivatives in step (b) and the kernel density in step (c) will involve $O(k^2)$ computations for each sampled weight vector. Thus computations might be expected to increase with $O(k^4)$, but this ignores the impact of k on importance weights. To gain some idea of the latter, suppose that we use the WLB to solve simultaneously k independent univariate problems. Importance weights would then be related geometrically to k. Thus increasing k could have a potentially devastating effect on importance weights. This can be illustrated by using the solder.balance data set analysed in Section 5.4. The authors report for the five-factor main effects model (containing 18 regression coefficients) that the largest 11% of the importance weights carried half the total mass. I repeated the analysis for the all-two-factor interactions model, which contains 113 regression coefficients, using S code supplied by the authors. In 300 WLB samples, the largest 1% of importance weights carried 59% of the total mass.

Bounds

The authors do not give a prescription for dealing with bounds on the parameters. Maximizing the weighted likelihood on a bounded domain in step (b) could produce poor kernel density estimates near the bounds. A better approach would be to maximize without bounds, and to attach zero importance weights in step (d) to points outside the bounds. However, this approach could be very inefficient if the bounds exclude most of the probability mass. In contrast, the Gibbs sampler with adaptive rejection sampling (Gelfand *et al.*, 1992; Gilks and Wild, 1992) can handle bounded domains quite efficiently.

Model choice

Section 7, providing a method for choosing between models using posterior simulation, is important and would be just as much at home in an MCMC paper. Many applied Bayesians will find this the most useful section of the paper.

Summary

I have indicated several potential disadvantages of the WLB. The principal advantages are that it can be rapidly deployed (requiring very little programming), it seems to work well for small but intricate problems and it generates independent posterior samples. In the short term, the WLB may enjoy some popularity until general purpose software for MCMC methods has been developed, but such software (for Gibbs sampling) is already reaching maturity (Gilks *et al.*, 1994). In the longer term, the WLB may find a niche in areas where the MCMC algorithm is prohibitively slow, such as in the optimal Bayesian design of complex studies.

It gives me great pleasure to welcome the WLB to the Bayesian toolkit, to wish it well and to propose the vote of thanks.

Gareth Roberts (University of Cambridge): In recent years, numerical techniques for Bayesian inference have improved significantly with the arrival of Markov chain Monte Carlo techniques such as the Hastings–Metropolis algorithm, the Gibbs sampler and their various hybrids and extensions. In fact many examples of applications of these techniques have been published. Therefore, it is refreshing to see that the authors do not consider Markov chain simulation to be the answer to all problems. They propose a method that is free from the uncertainty of dependent samples and the diagnosis of Markov chain convergence. Instead their method uses existing maximization routines to provide an approximate sample from the posterior, adjusting using sampling–importance resampling to provide a sample from the required distribution.

However, the performance of the weighted likelihood bootstrap must therefore be judged against that of Markov chain Monte Carlo. The paper suggests that a natural home for the weighted likelihood bootstrap is in low dimensional non-linear regression problems, where one-dimensional conditional distributions are not readily available, and moreover log-concavity of these conditionals cannot be relied on. Therefore application of the Gibbs sampler may not be straightforward. However, the simplest possible Markov chain Monte Carlo technique, the independence sampler (Tierney, 1991) is unaffected

by the unavailability of conditionals. It is a special case of the Hastings–Metropolis algorithm, where the proposal merely produces a sequence of independent and identically distributed (IID) random variables. I am indebted to Richard Gibbens, a colleague of mine at Cambridge, for helping me to establish that the independence sampler with independent normal proposals is sufficient to tackle this problem. Moreover, the routine is very easy to write and remarkably quick. Whereas independence sampling can rarely be recommended in higher dimensional problems, more sophisticated Markov chain Monte Carlo methods can. However, I share Dr Gilks's concern about the performance of the weighted likelihood bootstrap in higher dimensions.

Section 4 of the paper provides an asymptotic justification for the method. The most powerful result, theorem 2, demonstrates asymptotic correctness of the variance–covariance matrix of the weighted likelihood bootstrap under appropriate regularity conditions. Clearly these conditions will include essential smoothness conditions on the likelihood, and I urge the authors to state this result more fully in their reply, not only for mathematical correctness but also to give some idea of the sort of model for which the weighted likelihood bootstrap is applicable. Clearly the method is nonsense, for instance, when the model assumes IID observations from a uniform distribution on $[0, \theta]$.

To what extent does theorem 2 justify the weighted likelihood bootstrap? Certainly, in the examples given in the paper, analytical expressions of the observed Fisher information are readily available, and in general a simple numerical differentiation at the maximum likelihood estimate should be possible. Therefore a Laplace approximation (see for example Tierney and Kadane (1986)), generating multivariate normal observations with the prescribed variance–covariance matrix, will give the same level of asymptotic accuracy. However, at least in some models such as the constant coefficient of variation model of Section 5.3, the weighted likelihood bootstrap seems to show a remarkable tolerance to multimodality, whereas the Laplace approximation will not. The constant coefficient of proportionality model is intriguing and provides an interesting test case for perhaps a more pertinent asymptotic analysis in this context. No matter how many Taylor series expansion terms are in agreement, multimodality cannot be characterized. In multimodal posteriors, although asymptotic posterior normality still holds, the asymptotic behaviour of the minor modes is clearly of interest. Typically the minor modes will shrink to 0 at a geometric rate. The theory of large deviations allows us to study the rate of convergence to 0 of the probability mass contained in these minor modes.

Recalling the model for the weighted likelihood bootstrap,

$$[X_i|\theta] \equiv N(\theta, \theta^2), \qquad i=1, \ldots, n,$$

consider decay to 0 of probability in the minor mode. For the weighted likelihood bootstrap,

$$\tilde{\theta}_w = \pm\{(\bar{x}_w^2 + 4S_w)^{1/2} - \bar{x}_w\}/2$$

where

$$\sum_1^n w_i \bar{x}_w = \sum_1^n w_i x_i,$$

$$\sum_1^n w_i S_w = \sum_1^n w_i x_i^2$$

and the positive solution is taken if and only if $\bar{x}_w > 0$.

The model produces a bimodal posterior distribution, no matter how large the data set. A typical posterior is given in Fig. 7 of the paper. It is an extremely clean problem in that containment in either mode is very naturally defined by θ being greater than or less than 0. I have established the following large deviations results.

(a) Large deviations approximation of true posterior: let $p_n(x)$ be the proportion of posterior mass in the minor mode, given $x_1, \ldots, x_n$ (under perhaps a continuous positive everywhere prior)—

$$\frac{\log p_n(x)}{n} \to -2$$

for almost all x-sequences.

(b) Large deviations approximation based on a Laplace approximation:

$$\frac{\log p_n(x)}{n} \to -\frac{3}{2}$$

for almost all x-sequences.

(c) Large deviation approximation for the weighted likelihood bootstrap:

$$\frac{\log p_n(x)}{n} \to 0$$

for almost all x-sequences.

In fact decay in minor mode will be subexponential for all powered weights greater than or equal to 1.

Therefore the weighted likelihood is systematically giving too much weight to the minor mode. This suggests that a qualitative property of the weighted likelihood bootstrap in many problems will be to overemphasize areas of small probability, a heuristic supported strongly by the generalized linear model example in Section 5.4, as well as the non-linear regression of Section 2, as well as the theory and practice of the above example. Of course this will generally be good for sampling–importance resampling as the caption under Fig. 8 points out.

In conclusion, the authors propose an interesting new technique for numerical integration. However, guidelines for the suitability of the weighted likelihood bootstrap to any particular problem and for any particular prior are needed if the method is to become a reliable tool. It appears that the weighted likelihood bootstrap has useful properties for non-Gaussian posteriors. Asymptotic investigations such as the large deviations calculations above are needed to verify this. I have great pleasure in seconding the vote of thanks to Michael Newton and Adrian Raftery.

The vote of thanks was passed by acclamation.

Trevor Sweeting (University of Surrey, Guildford): The weighted likelihood bootstrap (WLB) is in general only first order correct as an approximation to the true posterior distribution and, in this sense, is no better or worse than a local normal approximation. The examples given, however, suggest that there is more to the method than just first-order correctness. I am interested in posterior approximations that can cope with multimodality, so I found the example of Section 5.3 particularly intriguing.

To understand just how the WLB manages to pick up bimodality, I generated samples from a normal mixture model with components $N(\theta, 1)$, $N(\theta+3, 1)$ and mixing probabilities $\frac{1}{2}$ each. Taking a uniform prior for θ, the posterior density of θ is frequently bimodal for small to moderate sample sizes. Fig. 10 shows the likelihood from a sample of six observations generated under $\theta=0$, whereas Fig. 11 shows a kernel density estimate obtained by using the WLB. The behaviour exhibited here is quite typical

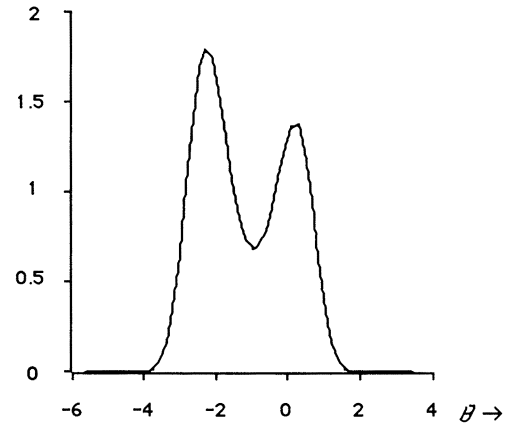

Fig. 10. Likelihood function for a sample of six observations from a normal mixture model

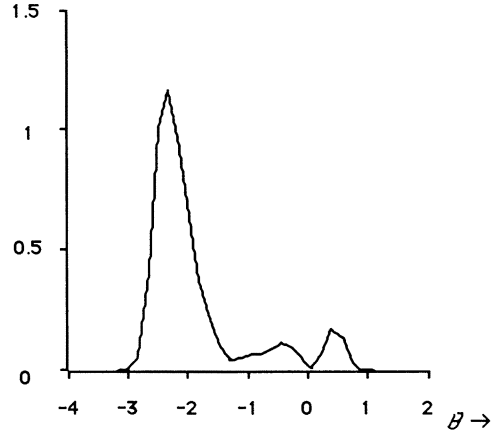

Fig. 11. Kernel density estimate of the likelihood shown in Fig. 10 based on 1000 WLB draws

and there are two points to note. Firstly, regarding computation, there is a danger that a second mode may be missed if local maximization is routinely used at each WLB draw. Secondly, although the WLB suggests that there is an interesting feature near 0, an insufficient number of WLB simulations result in the maximum located near 0 being the global maximum and the method gives a rather poor picture of the actual posterior.

A WLB result like that of Fig. 11, however, does not mean that there is necessarily some interesting posterior behaviour near 0. This can be seen by noting that, in this example, very similar WLB results to Fig. 11 can be obtained even when the actual likelihood has almost normal shape! The WLB is picking up the fact that there is a feature of the likelihood near 0 for some samples which were not actually observed. Of course, a straight normal approximation will be almost exact in such a case.

It would seem that, in general, we have no way of telling whether or not various irregular features of the WLB are spurious, in terms of the data observed. Perhaps this type of behaviour is less important when the WLB approximation is used only as an input to some other fully simulation consistent method, but I find this anti-Bayesian aspect of the method rather disquieting.

Douglas Bates (University of Wisconsin, Madison) **and Christian Ritter** (Université Catholique de Louvain, Louvain-la-Neuve): We extend our congratulations to the authors on their innovative addition to the arsenal of procedures for creating sample points from likelihoods, posteriors or approximations thereof. We tested the method on the biochemical oxygen demand example where it took only two hours to do the programming and to obtain 5000 sample points. The resulting sample (using the overdispersion $\alpha = 1.6$) traces the important features of the likelihood. However, it contains additional structure that is not contained in the likelihood. In particular, the cloud of sample points looks like a combination of several clouds of different shapes. The reason might be that the data set is quite small and that the overdispersion effectively suppresses observations at random and produces results similar to those from the jackknife for small sample sizes. Although this should be researched further, it indicates to us that the authors' suggestion of a sampling–importance resampling adjustment after the initial weighted likelihood bootstrap is recommended, at least in small samples.

Moreover, if the weighted likelihood bootstrap is to be used with non-linear regression problems, conventions need to be adopted on how to handle cases where the fitting algorithm does not converge for some selections of weights and where parameters drift off to infinity.

But these are comparatively minor matters. The ease of implementation of this method combined with our ever increasing computing power should make this a valuable addition to inference techniques for non-linear models.

B. J. Worton (University of Essex, Colchester): I would like to consider the bootstrap likelihood methods discussed in Section 8 in more detail and to give some recommendations on the accurate and efficient application of these methods. In Davison *et al.* (1992), we constructed a bootstrap likelihood by using data $y_1, \ldots, y_n$, an estimator T for a parameter θ and the following nested bootstrap algorithm.

(a) Use the bootstrap to generate populations $\mathscr{P}_1^*, \ldots, \mathscr{P}_M^*$ with parameter values $\theta_1^*, \ldots, \theta_M^*$.
(b) For each population $\mathscr{P}_i^*$, use a second level of bootstrapping to generate T^{**}s for the parameter value θ_i^*.
(c) Smooth by kernel density estimation each set of T^{**}s and evaluate at t the observed value of the statistic for the data, to produce likelihood points at $\theta_1^*, \ldots, \theta_M^*$.
(d) Smooth the scatterplot of likelihood points to compute a likelihood curve.

Application of this widely applicable algorithm is obviously very time consuming and can be made more efficient at both the first and the second levels of bootstrapping. At steps (b) and (c) of the algorithm, avoid the use of Monte Carlo simulation and kernel density estimation if possible, and instead use saddlepoint density approximations (see Davison and Hinkley (1988) and Daniels and Young (1991)) to compute the T^{**}-densities evaluated at t. However, even with these accurate approximations, considerable variation remains in the likelihood point estimates. This is not due to inaccuracies of the saddlepoint approximations, but the inherent variability of the first-level bootstrap-generated $\mathscr{P}_1^*, \ldots, \mathscr{P}_M^*$ populations which are central to the method. Therefore, it is advisable not to use $\mathscr{P}_1^*, \ldots, \mathscr{P}_M^*$ directly, but to smooth them before use. This can be achieved for a target θ^0-value and bandwidth ϵ by using a kernel smoother

$$p_j^*(\theta^0, \epsilon) \propto \sum_{i=1}^{M} w\{(\theta^0 - \theta_i^*)/\epsilon\} p_{ij}^*, \qquad j = 1, \ldots, n,$$

where p_{ij}^* is the probability associated with point y_j in population $\mathscr{P}_i^*$ and $p_j^*(\theta^0, \epsilon)$ is the corresponding smoothed probability. The *precise* value of the parameter θ for the smoothed population should then be used for θ^*. A grid of target θ^0-values can be used to generate M populations at step (a). Whether used in conjunction with Monte Carlo simulation and kernel density estimation methods or saddlepoint methods, this approach works well, and for moderate levels of smoothing produces a smooth curve at step (c).

The connection between these bootstrap likelihood methods and the weighted likelihood bootstrap is unclear at present. However, could the authors elaborate on the extension of the weighted likelihood bootstrap to semiparametric models and the deeper connections with other bootstrap likelihood methods?

G. A. Barnard (Colchester): The concatenation of 'weighted' with 'bootstrap' and 'likelihood' joins three ideas, each good in itself, but puzzling in combination.

In the first two examples the likelihood functions could be easily calculated. If the data points for the first had been given as they should have been, a likelihood resembling Fig. 1(b) would have been obtained, suggesting strongly that the ratio and the product of the βs might reasonably be estimated, but attempting to estimate either parameter separately would be dangerous.

In the second example the analysis produces a posterior relative to a doubly improper prior with singularities at 0 and 1. Psychological considerations would suggest a prior restricted to a closed proper subset of [0, 1], but the authors' prior does the exact opposite. Straightforward likelihood plotting would correspond to a uniform prior, and posteriors relative to more appropriate priors could easily be obtained from this.

We badly need to explore ways of exhibiting likelihood functions involving many parameters. The technologies involved in 'virtual reality' could help in this.

But I share with Tukey the view that we are wise to feel that, if we cannot formulate our thoughts by using five parameters or fewer, we had better think again before proceeding. Until then we should engage in exploratory activities such as projection pursuit. An exception to this arises in those problems where we can recognize the correct solution as soon as we see it; but I fail to see the use of the weighted likelihood bootstrap here either.

The soldering example reminded me of when the Plackett–Burman experimental designs were invented during the war. There was an urgent need to proceed with the manufacture of fuses which helped to bring down some of the V1 weapons. Urgencies of that sort could render meaningful justifications such as that the method uses only three lines of code. But in normal times the experiments generating our data involve far more effort than do three lines of code.

I hope that the authors will be able to present the background and details of a realistic problem where their proposed method has real advantages over other methods. Until that happens I remain unconvinced.

Richard Gibbens (University of Cambridge): I would like to make some brief remarks concerning an experiment that Gareth Roberts and I were motivated to consider by this interesting paper on Bayesian

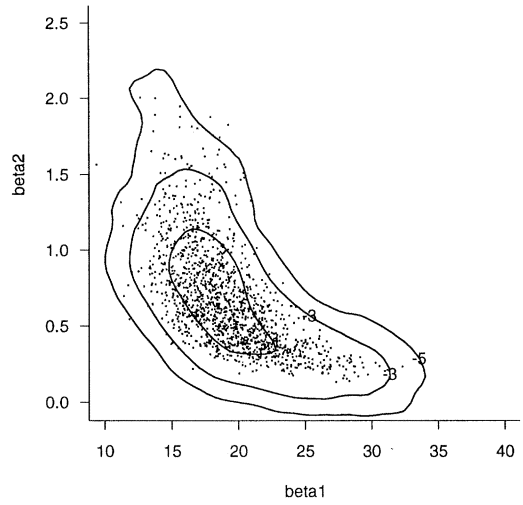

Fig. 12. Kernel density estimate

methods. We decided to consider a simple form of Markov chain Monte Carlo technique for the non-linear regression example of Section 2 of the paper. As Gareth Roberts has already mentioned we used an independence sampler to construct a large sample from the posterior distribution for β_1, β_2 and σ^2. Fig. 12 shows this sample (of 15000 observations) together with contours describing a kernel density estimate of the distribution. We found our results to be in broad agreement with those shown in Fig. 1(c). (The choice of contours reflects those from Fig. 1 but they should be viewed with caution given the clear lack of data in the tails of our sample.)

The authors describe some of the software engineering aspects of their work and this is an area of particular interest to us. They found in one example that they could implement their approach in about 60 lines of S code together with built-in functions. In our experiment we had similar experiences in that we used about 30–40 lines of S code to construct the various statistical quantities and then used a short C routine (of four lines) called by means of the interface between S and C to implement the independence sampler.

Bernard Silverman (University of Bath): It is interesting to see so much use of kernel density estimation, both in the paper and in the discussion. The maximal smoothing method of Terrell (1990), as used by the authors in their Fig. 1(d), is of course constructed by reference to the accuracy of the density estimate as a whole. Extreme contours of f are of most interest, and so it may be worth considering an adaptive method such as that described in Silverman (1986). Even Terrell's generally conservative method seems to undersmooth in the tails. Richard Gibbens (previous contribution) is brave or foolhardy to show both the original data and the contours of his density estimate, which indicate a need for less severe smoothing perhaps linked with some adaptivity. Overall there are perhaps two points: firstly, in contrast with most bootstrap applications, a somewhat larger bootstrap sample than 5000 may be useful if extreme contours of a two-dimensional likelihood are of interest; secondly, some more thought about the most appropriate way to estimate these extreme contours may be worthwhile.

A. C. Davison (University of Oxford): One feature that distinguishes this work from the host of other recently proposed methods for approximate Bayesian inference is that, as the authors remark in Section 1, the raw weighted likelihood bootstrap is not an approximation to an exact Bayesian procedure. Put more bluntly, it does not condition fully on the data. Instead Dirichlet weights are used to perturb estimating equations that determine a suitable M-estimate of the parameter of interest. This lack of full conditioning may worry some Bayesian statisticians, but it might be a positive advantage from a frequentist point of view. However, just as it induces a prior on the parameter space, it induces a measure on the observation space. Have the authors any comments on this aspect of their idea, particularly when the data are dependent?

For models where the log-likelihood for a parameter β given observations y_j with prior weights w_j may be written

$$\sum_{j=1}^{n} w_j \frac{a(y_j;\ \beta)}{\phi} + c(y_j,\ w_j;\ \phi), \tag{19}$$

such as linear and non-linear regressions, generalized linear models, and so forth, the type of data perturbation proposed by the authors, which amounts to jittering the w_j, seems natural. It corresponds to asking how the estimator $\hat{\beta}$ resulting from model (19) would have been different had the initial weights been different. But this straightforward interpretation is not universal, and I have misgivings about whether the method can always give sensible answers—particularly in problems where the contributions to the estimating equations are non-independent, though the partial likelihood discussed in Section 6.2 suggests that my reservations may be unfounded.

The regression framework considered in Section 5 of the paper is very general, but here too I smell a rat. Suppose that the unweighted estimating equations used in equation (5) corresponded to a quasi-likelihood or a robust estimator. Then, although each estimating equation might be sensible in itself, there would often be no unique objective function whose derivatives yielded the estimating equations. To apply the weighted likelihood bootstrap would produce an approximate posterior distribution, but for a likelihood that was not unique. The authors do not attempt this, but I find the possibility alarming, and wonder whether the authors could comment.

The following contributions were received in writing after the meeting.

Mostafa Bacha and Gilles Celeux (Institut National de Recherche en Informatique et en Automatique, Le Chesnay): We illustrate the behaviour of the weighted likelihood bootstrap (WLB) in a small sample setting with highly censored data and provide a numerical illustration of the influence of the parameter α when the WLB is combined with the sampling–importance resampling (SIR) algorithm. For this, we ran the WLB–SIR method for a simple example. The results are very sensitive to α. Small α can lead to a very sharp weight distribution for which the SIR adjustment does not provide a satisfactory approximation to the posterior distribution; large α will provide an overdispersed weight distribution. The choice of a good α-value has to be done in a rather empirical manner by trying different values.

Our example is a commonly encountered situation in failure time analysis. We simulated a right-censored sample of size 30 from a Weibull distribution with cumulative density function $F(x) = 1 - \exp\{-(x/\eta)^{\beta}\}$. We consider the shape parameter $\beta = 2$ and the scale parameter $\eta = 1000$. The censoring time was $c = 700$. The seven uncensored points were 185.3, 341.5, 388.4, 541.2, 580.8, 597.3 and 668.6. The maximum likelihood estimators were $\beta_{\mathrm{ml}} = 2.36$ and $\eta_{\mathrm{ml}} = 1279$. Their approximate standard deviations from the observed information matrix were 0.81 and 187 respectively.

We used a gamma(λ, μ) prior for β with a shape parameter $\lambda = 25$ and a scale parameter $\mu = 10$ so that the mass of the prior distribution is concentrated on [1, 4]. For η, we used a prior $\pi(\eta) \propto 1/\eta$ and the parameters β and η are assumed to be independent. Using numerical integration, we computed the true mean β_{B} and the standard deviation $\mathrm{Sd}(\beta_{\mathrm{B}})$ of the posterior distribution of the shape parameter. We obtained $\beta_{\mathrm{B}} = 2.37$ and $\mathrm{Sd}(\beta_{\mathrm{B}}) = 0.42$.

We ran the WLB–SIR algorithm with different values of α from 0.2 to 4.0, using 2000 draws from the weight distribution. For brevity, we focus on the β posterior distribution. The results are summarized in Table 2 for two significant values of α. The first value $\alpha = 0.4$ is a value for which the WLB distribution

TABLE 2

Means and standard deviations of the Weibull parameters from the WLB–SIR approximate posterior distribution for two values of α

α	*Shape*		*Scale*	
	Mean	*Standard deviation*	*Mean*	*Standard deviation*
0.4	2.41	0.23	1273	70
1.4	2.34	0.37	1355	234

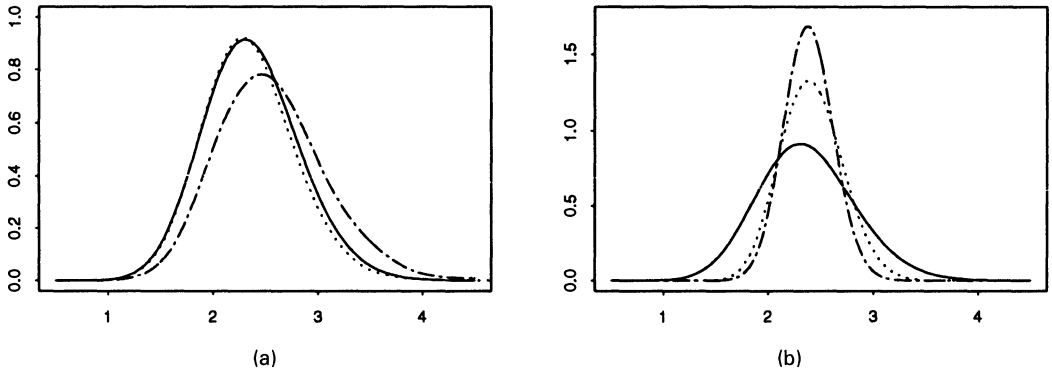

Fig. 13. True posterior distribution (———), WLB–SIR approximation (· · · · · ·) and WLB approximation (— · —) for (a) $\alpha = 1.4$ and (b) $\alpha = 0.4$

is sharp and concentrated on [2.15, 2.64]. Consequently the WLB–SIR distribution approximates the actual posterior distribution poorly. The second value $\alpha = 1.4$ provides the best approximate posterior distribution. This is quite satisfactory (see Fig. 13), especially in light of the very small amount of complete data. From the examples of Newton and Raftery and this simple example, the problem of finding a successful value of α is highly data dependent. Thus, a data-driven determination of α may be useful to improve the performance of the WLB in Bayesian inference.

Bradley P. Carlin (University of Minnesota, Minneapolis): First, I congratulate Newton and Raftery on a fine and intriguing paper in the rapidly developing area of computational methods for Bayesian statistics. The weighted likelihood bootstrap's (WLB's) ability to make use of existing maximum likelihood code is a key advantage over other such methods, and one that will increase its appeal among practitioners. Also, the WLB appears to require less tuning than Markov chain Monte Carlo (MCMC) methods, especially in the area of convergence diagnosis. Still, the WLB's several levels of iteration (repeated weight generation, maximization, density estimation and final sampling–importance resampling adjustment) make me wonder about the differential time, both in human and computer terms, involved in running WLB *versus* MCMC. Implementation of the WLB for high dimensional problems also seems problematic.

In Section 6.1, the extension of the WLB to time series of dependent data is described. But notice that in the further extension to series with *missing* dependent data (as in a series where only record breaking events are observed), the likelihood itself involves a potentially high dimensional integral. In such cases, Carlin and Gelfand (1993) show that an MCMC algorithm (sampling over the missing data) is often the only feasible way to obtain maximum likelihood estimates of the unknown model parameters. Fortunately, making the jump to a fully Bayesian analysis is straightforward simply by also including the parameters in the sampling order.

The authors mention a connection of their work to that of Laird and Louis (1987); the WLB's substitution of maximizations for integrations indeed gives it an empirical Bayes (EB) flavour. Subsequent work by Carlin and Gelfand (1990) shows an even stronger connection. Carlin and Gelfand point out that the Laird and Louis bootstrap algorithm for widening 'naïve' EB confidence intervals is essentially a method for matching a fully Bayesian solution under an 'effective hyperprior' which is not necessarily simple nor natural for the problem at hand. They then develop a generalized bootstrap which can often be used in conjunction with importance sampling to match a given hyperprior Bayes solution, or alternatively to correct the bias in using the naïve intervals. Similar guidance for the WLB (in the form of the general recipe for the weight distribution, mentioned by the authors in Section 8) would be welcome.

P. Clifford (Oxford University): Like Professor Barnard I feel a little uneasy about this paper. The recent revival of interest in Bayesian methods arises not from a wholesale conversion among statisticians to Bayesian ideology but from a need to deal with high dimensional parameter spaces in modern applications such as image restoration and the analysis of extensive epidemiological databases. In these applications, it is convenient to co-ordinate parameters by imposing probabilistic structures which are analogous to those which would be introduced by Bayesian theory. The structures may involve relatively

few hyperparameters which themselves, in principle, are amenable to estimation by classical methods. The entrepreneurial atmosphere in which classical statistics has developed has produced many sharp-edged tools. When only a few parameters are involved, statisticians feel confident in using these tools. When the number of parameters goes into the thousands it is less obvious how to proceed and the clear centralist dogma of Bayesianism becomes an attractive starting point. It is unfortunate that the authors' examples do not fall into this class of problem: their first example has six observations and three parameters. I suspect that the majority of statisticians would not hesitate to use classical methods of inference for these data and would see no need to tamper with the likelihood by introducing prior weighting.

Putting these general remarks aside I would like to turn to the subject of Markov chain simulation. The authors claim that such simulations are more difficult to apply than sampling–importance resampling adjusted weighted likelihood bootstrap. This assertion is based on a comparison with the Gibbs sampler. The authors do not seem to have considered using the Metropolis sampler which despite its venerability is almost invariably trivial to apply. Apart from a few special cases there seems to be no good reason for ever using the Gibbs sampler in preference to the Metropolis sampler. In its simplest form, the Metropolis sampler produces a Markov chain with equilibrium density $\pi(\theta_1, \ldots, \theta_p)$ by cycling through the parameter indices, successively altering parameters in a manner similar to the Gibbs sampler. When parameter θ_k is selected a proposal θ_k' is generated by adding to θ_k an independent random variable *uniformly* distributed on the interval $(-\Delta_k, \Delta_k)$. If $\pi(\theta_1, \ldots, \theta_k', \ldots, \theta_p)/\pi(\theta_1, \ldots, \theta_k, \ldots, \theta_p) > U$, where U is independently uniformly distributed on (0, 1), then the proposal is accepted; otherwise the parameter is unchanged. If the proposal falls outside the support of the density it is automatically rejected. Traditionally, in the physics community, the vector $\Delta = (\Delta_1, \ldots, \Delta_p)$ is adjusted dynamically until proposals are ultimately being accepted about 50% of the time. Coding this procedure in Basic for the first example requires only *one* line of code (with 12 colons).

Lu Cui, Michael Sherman and Martin A. Tanner (University of Rochester): We have found it useful to rethink the authors' approach in terms of the method of composition and importance sampling (Tanner, 1991). Let

$$L(\beta \mid Y) = \prod_{i=1}^{n} f_i(\beta; x_i), \quad L_w(\beta \mid Y) = \prod_{i=1}^{n} f_i(\beta; x_i)^{w_i},$$

where w is the vector $(w_1, w_2, \ldots, w_n)$ with density $g(w)$. Sampling w from the Dirichlet(1, 1, . . ., 1) distribution simplifies the following algorithms. Note that

$$L(\beta \mid Y) = \int \frac{L(\beta \mid Y)}{c(w)\, L_w(\beta \mid Y)}\, c(w)\, L_w(\beta \mid Y)\, g(w)\, \mathrm{d}w,$$

where $c(w)$ is equal to the reciprocal of $\int L_w(\beta \mid Y)\, \mathrm{d}\beta$ for a given value of w. This identity suggests the following *ideal weighted likelihood bootstrap (WLB) algorithm*:

(a) draw a w^*-vector from $g(w)$;
(b) sample a β^*-vector from $c(w^*)\, L_{w^*}(\beta \mid Y)$;
(c) assign mass $L(\beta^* \mid Y)/c(w^*)\, L_{w^*}(\beta^* \mid Y)$ to β^*.

By repeating steps (a)–(c) we have a sample from the likelihood of interest. The method of composition is used in steps (a) and (b), whereas importance sampling is used in step (c). If the distribution of these masses is skewed, then as noted by the authors (see also Tanner and Wong (1987)) one will need to use an iterative algorithm (e.g. data augmentation) to correct the deficiency.

It will typically be difficult to sample directly a β^*-vector from $L_{w^*}(\beta \mid Y)$. One approach would be to approximate $L_{w^*}(\beta \mid Y)$ by a matching multivariate normal density function centred at the mode of $L_{w^*}(\beta \mid Y)$, i.e. $\tilde{\beta}$ with variance–covariance matrix $\tilde{\Sigma}$ defined by the curvature of $L_{w^*}(\beta \mid Y)$ at $\tilde{\beta}$. We refer to this distribution as $\phi_{w^*}(\ ,\ \tilde{\beta},\ \tilde{\Sigma})$. More complicated algorithms could be defined by using the multivariate t- or the split t-distribution. The modified algorithm (the *normal WLB algorithm*) would then be

(a′) draw a w^*-vector from $g(w)$;
(b′) sample a β^*-vector from $\phi_{w^*}(\ ,\ \tilde{\beta},\ \tilde{\Sigma})$;
(c′) assign mass $\omega^* = L(\beta^* \mid Y)/\phi_{w^*}(\beta^*,\ \tilde{\beta},\ \tilde{\Sigma})$ to β^*.

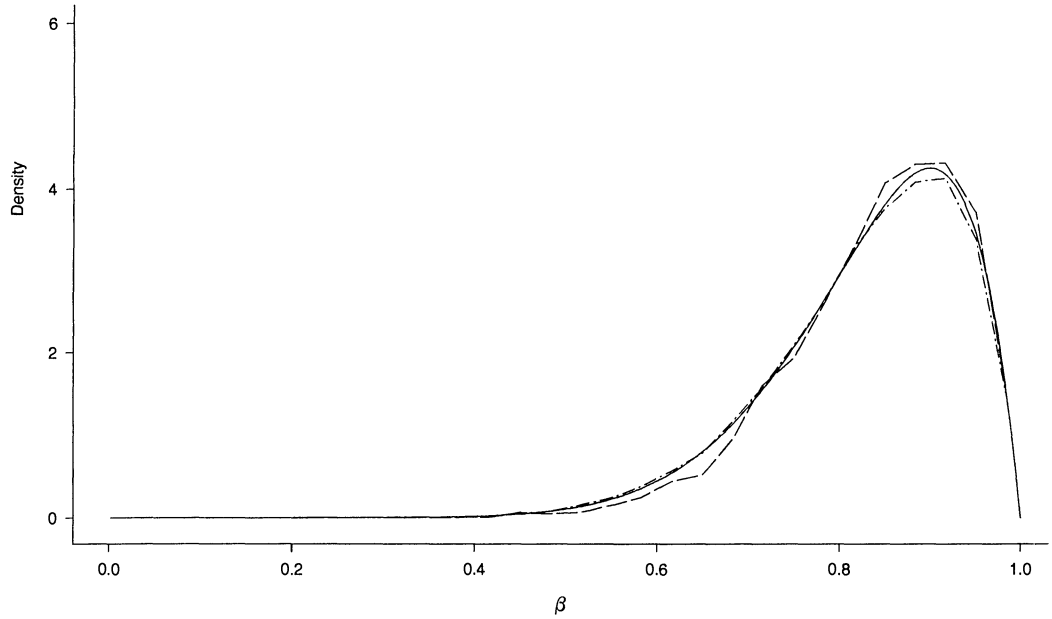

Fig. 14. Posterior distribution of β (flat prior)

By repeating steps (a′)–(c′) we have a sample $(\beta^{(1)}, \omega^{(1)}), \ldots, (\beta^{(m)}, \omega^{(m)})$, where $\omega^{(i)}$ is the mass assigned to $\beta^{(i)}$, from the likelihood of interest. This normal WLB algorithm assigns appropriate weights to the βs without the need to form a nonparametric density estimate. Also, on the basis of this algorithm, a Rao–Blackwell estimate of $L(\beta \mid Y)$ is given by the ω-weighted mixture of multivariate normals.

To illustrate these algorithms we consider the likelihood $\beta^9(1-\beta)$. Because the true likelihood is proportional to a β-distribution, the ideal WLB algorithm can be easily implemented. w was generated by drawing 10 independent unit exponentials and dividing each by the sum, following the authors. Sampled deviates which fell outside [0, 1] were rejected. The acceptance rate was about 80% in this example. Fig. 14 presents a histogram of the output from the ideal (chain curve) and normal (broken curve) WLB algorithms, as well as the true (full) curve, where each density estimate is based on 20000 simulated values. Both curves seem to track the true likelihood. Further work to examine the performance and utility of these algorithms and the possibility of converting the normal WLB algorithm to an iterative algorithm is needed.

Alan E. Gelfand and Bani K. Mallick (University of Connecticut, Storrs): The following example is a special case of a class of doubly semiparametric proportional hazards models discussed in Mallick and Gelfand (1993).

Let $t_1, t_2, \ldots, t_n$ be such a set of uncensored survival times having covariates $\mathbf{x}_i$ associated with time t_i. Let the integrated hazard for t_i have the form $H_0(t_i)\, g(\mathbf{x}_i^{\mathrm{T}}\beta)$ where H_0 is strictly increasing differentiable from R^+ to R^+ and g is strictly decreasing differentiable from R^1 to R^+. H_0 and g are assumed unknown and modelled as follows. $H_0(t) = J(t)/\{1-J(t)\}$ where $J(t) = w\,\mathrm{IB}\{J_0(t);\ 1,\ 2\} + (1-w)\,\mathrm{IB}\{J_0(t);\ 2,\ 1\}$, $J_0(t) = t/(1+t)$, $w \in [0, 1]$ where IB(; c, d) denotes the incomplete beta function associated with a Be(c, d) density. Hence specification of w determines H_0. If w is random, say Be(α, α), then $E\,H_0(t) \approx t$, i.e. the base-line hazard is 'centred' on the exponential hazard. $g(\eta) = 1/\{1+K(\eta)\}$ where $K(\eta) = \nu\,\mathrm{IB}\{K_0(\eta);\ 1,\ 2\} + (1-\nu)\,\mathrm{IB}\{K_0(\eta);\ 2,\ 1\}$, $K_0(\eta) = \exp\eta/(1+\exp\eta)$, $\nu \in [0, 1]$. Hence ν determines g and, if $\nu \sim \mathrm{Be}(\alpha, \alpha)$, $E\,g(\eta) \approx \exp(-\eta)$, the usual covariate link. Taking β, ω and ν as unknown results in the likelihood

$$L(\beta,\ w,\ \nu) \propto \prod_{i=1}^{n} H_0'(t_i)\, g(\mathbf{x}_i^{\mathrm{T}}\beta) \exp\{-H_0(t_i)\, g(\mathbf{x}_i^{\mathrm{T}}\beta)\},$$

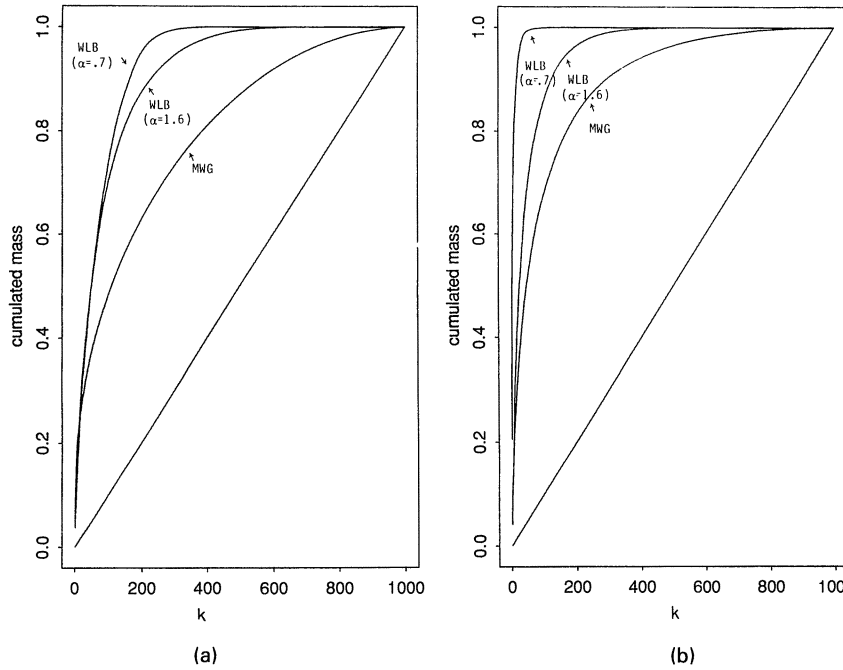

Fig. 15. Total mass assigned to the k largest weights *versus* k for (a) a flat prior for β and (b) Jeffreys's prior for β

evaluation of which requires $4n$ IB evaluations; maximization can be done via grid search or possibly a derivative-based method since $\partial L/\partial\beta_i$, $\partial L/\partial w$ and $\partial L/\partial \nu$ can be directly computed. The Bayesian model is completed by taking independent uniform priors for w and ν and, given w, a flat or Jeffreys prior for β.

We applied this model to a data set described in Aitkin *et al.* (1989) ($n=33$). Allowing for interaction, β becomes four dimensional. Aitkin *et al.* (1989) fit, among other models, an exponential generalized linear model which corresponds to our base-line model. We fit the six-dimensional Bayesian model in two ways. We ran a weighted likelihood bootstrap (WLB) using 1000 weights drawn from Dir($\alpha\mathbf{1}$) with $\alpha=0.7$ and $\alpha=1.6$ thus obtaining 1000 samples approximately from the posterior under a flat prior. We ran an adaptive Metropolis-within-Gibbs (MWG) algorithm for both a flat and a Jeffreys prior following Müller's (1991) suggestions, stopping each of five strings after 500 iterations using the last 200 iterations from each again to obtain 1000 samples approximately from the posterior. Run times for the WLB on an IBM 3094 computer were 220 min for each α: for the MWG 40 min. Using either the WLB or the MWG samples we obtained the importance weights u_j as defined by the authors.

In Fig. 15 we plot the total mass assigned to the k largest weights against k for the two WLBs and the MWG. The MWG weights are dramatically better than the WLB weights even under a flat prior. The latter has much more mass attached to far fewer points.

Andrew Gelman (University of California, Berkeley): As the authors point out, the weighted likelihood bootstrap (WLB) can be used in place of the Gibbs sampler or Metropolis algorithm in a wide class of problems for which it is a fairly close fit to the target posterior distribution (so that, after importance resampling, the distribution will be almost exactly correct). One tricky point in application seems to be determining whether the importance ratios are sufficiently close for the method to be accurate. As with any approximate method that is based on overdispersion, the WLB has the potential for an even wider range of practical applicability: for problems in which the WLB simulation distribution is not a close fit, it may still be useful to use the sampling–importance resampling (SIR) samples as starting points for a Markov chain simulation. For example, Section 4.2 of Gelman and Rubin (1992) illustrates SIR samples (although not from the WLB) used successfully as a starting distribution for parallel runs of a Gibbs sampler.

Variation of importance ratios also seems like a potential difficulty in the estimates of the marginal likelihood presented in Section 7. Meng and Wong (1993) present a similar method for estimating marginal likelihoods and normalizing constants that uses samples from two distributions to achieve a lower variance than the harmonic mean estimator.

Let $p_i(\theta)$, $\theta \in \Theta_i$, $i=1, 2$, be two densities, each of which is known up to a normalizing constant: $p_i(\theta)=q_i(\theta)/z_i$, $i=1, 2$. The following identity is fundamental to our approach:

$$\frac{z_1}{z_2}=\frac{E_2\{q_1(\theta)\,\alpha(\theta)\}}{E_1\{q_2(\theta)\,\alpha(\theta)\}}, \tag{20}$$

where E_i denotes the expectation with respect to the p_i ($i=1, 2$) and $\alpha(\theta)$ is an arbitrary function defined on the common support, $\Theta_1 \cap \Theta_2$, such that the two expectations are finite and non-zero. Given draws from both densities and a choice of α, the numerator and denominator on the right-hand side of equation (20) can be simulated easily.

The 'harmonic mean' (13) corresponds to choosing $q_1(\theta)=p(x|\theta)\,p(\theta)$, $q_2(\theta)=p(\theta)$ and $\alpha(\theta)=\{p(x|\theta)\,p(\theta)\}^{-1}$. Similarly, equation (14) corresponds to the same choice of q_i, $i=1, 2$, with $\alpha(\theta)=p(\theta)^{-1}$. For these choices of α, as the authors noted, the resulting simulation may be unstable because the corresponding integrands are not necessarily square integrable. Furthermore, they provide legitimate estimates only when the prior $p(\theta)$ is proper. By suitable choices of α in equation (20), however, all these problems can be avoided. For example, choosing $\alpha=1/\sqrt{(q_1 q_2)}$ leads to

$$\frac{z_1}{z_2}=\frac{E_2[\sqrt{\{q_1(\theta)/q_2(\theta)\}}]}{E_1[\sqrt{\{q_2(\theta)/q_1(\theta)\}}]}, \tag{21}$$

where both integrands are always square integrable with respect to the corresponding densities. Implementing equation (20) with the optimal choice of α is also straightforward, as is detailed in Meng and Wong (1993). Continuous extensions of these methods are presented in Gelman and Meng (1993).

A. P. Grieve (ZENECA Pharmaceuticals, Macclesfield): On the basis of their experiences the authors suggest that the overdispersion parameter α should be chosen 'rather close to 1'. How are we to interpret this in the light of their use of values for α of 0.7 and 1.6? Is this effectively a recommendation to use purely uniform Dirichlet weights?

To understand to what extent a blanket adoption of uniform Dirichlet weights is reasonable I have applied the weighted likelihood bootstrap (WLB) to the family of distributions

$$\exp\left\{\frac{X\theta-b(\theta)}{a(\phi)}+c(X,\,\phi)\right\}$$

and have assumed

(a) that ϕ is known, which gives the exponential family with canonical parameter θ, and
(b) that interest centres on making inferences about the mean value parameter $\mu=E(X)=b'(\theta)$.

For data x_i ($i=1, \ldots, n$) and a given set of random uniform Dirichlet weights y_i ($i=1, \ldots, n$), the WLB gives, as a random estimate of μ,

$$\tilde{\mu}=\sum_{i=1}^{n} y_i x_i. \tag{22}$$

Using standard properties of the uniform distribution on an n-dimensional simplex it is possible to show that

$$E(\tilde{\mu})=\bar{x}, \quad \mathrm{var}(\tilde{\mu})=\frac{n-1}{n+1}\frac{s^2}{n}.$$

There are two features of this result which are potentially disturbing, both of which arise because the WLB does not use all available information. Firstly, irrespective of which member of the exponential family I am interested in, the WLB simulates from the same data-based distribution, although the posteriors are very different in shape. Secondly in the concrete example of a normal distribution with known variance σ^2 we would presumably be wanting to simulate from a distribution with mean $\bar{x}$ and

variance σ^2/n. In practice s^2/n will deviate from σ^2/n and the degree to which it does so will influence the choice of overdispersion parameter α.

Arising from these observations I have further questions for the authors. Can they expand on their comments concerning an empirical choice for α and provide guidelines for choosing one particular value? Have they investigated any other weighting distributions, apart from the α-modified uniform Dirichlet? If so how do they compare with the uniform Dirichlet weights and can one make an empirical choice between them? For the exponential family we could presumably work with equation (22) to choose in some sense an optimal weight distribution to give a random sample from the true posterior, but of course in this case we know the true posterior and therefore do not need the WLB.

Hans R. Künsch (Eidgenössische Technische Hochschule, Zürich): Trying to summarize this interesting paper in one sentence, one could say (ignoring the difference between multinomial and Dirichlet weights) that the bootstrap of the maximum likelihood estimator $\hat{\theta}(x)$ gives a good approximation not only to the sampling distribution $\mathcal{L}(\hat{\theta}-\theta\,|\,\theta)$ but also to the posterior $\mathcal{L}(\theta\,|\,x)$. Since the bootstrap does not involve a prior, such a result would counter a frequent argument against Bayesian methods, namely the possibly crucial dependence of the conclusions on a subjective prior. However, extreme choices of the prior do change the posterior substantially. So the main question seems to me to be, for which priors is the weighted likelihood bootstrap (WLB) approximation sufficiently good? Here I liked the idea of looking at how unequal the importance weights given to the bootstrap values are. Another concern is the assumption that the model considered is correct. For example in the time series analysed in Fig. 9 the AR(1) model is chosen mainly for convenience. If some other model generated the data the WLB will not approximate the sampling distribution. My main motivation for proposing the blockwise resampling in Künsch (1989) was to retain the model-free nature of Efron's original proposal also under dependence. Note that when we choose the weights $w_{n,1}, \ldots, w_{n,n}$ in the WLB to be dependent then we obtain a procedure which behaves like blockwise resampling with increasing block size $l(n)$; see formula (2.12) in Künsch (1989).

Thomas Leonard (University of Wisconsin, Madison) **and John S. J. Hsu** (University of California, Santa Barbara): The weighted likelihood function (1) provides a very interesting, eclectic, idea. Have the authors taken this idea one stage further, and tried to incorporate the simulated value of this function in the denominator of the importance sampling weight function? This modification would avoid the need for the more complicated density estimates that they use, and it would be neat if it turned out to converge more rapidly.

In many special cases, for example the non-linear regression model, on the third page, it is possible instead to use conditional maximization techniques to obtain continuous approximations with saddlepoint accuracy to marginal posterior densities or moments (e.g. Leonard (1982), Tierney and Kadane (1986), Leonard *et al.* (1989) and Tierney *et al.* (1989)). The conditional maximization can be completed with standard packages, and a variety of adjustments to the profile posterior density included to ensure convincing finite sample accuracy. Since the computing time is very small, and the approach is algorithmic, are simulations always necessary?

If we do decide to simulate, then it is important to use a procedure which has reasonable asymptotic properties, as the number of simulations M increases, with n fixed. In Section 4, the authors let $n \to \infty$, so that theorem 2 does not always help us. Suppose instead that to compute the expectation of a function $\omega(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$, with respect to a fully specified posterior density $\pi_y(\boldsymbol{\theta})$, we perform importance sampling by simulating from an approximation $\pi_y^*(\boldsymbol{\theta})$ to the posterior density which has positive support on the parameter space Θ. Then (e.g. Geweke (1989) and Leonard and Hsu (1992)), the simulated posterior expectation ω_M will converge almost surely to the exact posterior expectation $\tilde{\omega}$, as $M \to \infty$, if $\tilde{\omega}$ is finite. Furthermore, whenever the quantity

$$V = \int_\Theta \omega(\boldsymbol{\theta})\{\pi_y(\boldsymbol{\theta})^2/\pi^*(\boldsymbol{\theta})\}\,\mathrm{d}\boldsymbol{\theta} - \tilde{\omega}^2$$

is finite, $M^{1/2}V^{-1/2}(\omega_M - \omega)$ converges in distribution to a standard normal random variate. The importance function $\pi^*(\boldsymbol{\theta})$ should be chosen to ensure that V is likely to be as small as possible, and V can itself be calculated during the simulations. A small value for

$$W = \int_\Theta \pi_y(\boldsymbol{\theta})^2/\pi^*(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} - 1$$

ensures good convergence for the expectation of any bounded function. Can the authors demonstrate a similar result for their procedure? We believe that a requirement of asymptotic normality can be used to clarify large subsections of the literature on Bayesian simulations. It can, for example, suggest situations where the Gibbs sampler would converge within a practical time limit.

We congratulate the authors on some outstandingly novel ideas.

Jun S. Liu and Donald B. Rubin (Harvard University, Cambridge): Newton and Raftery's examples appear to provide striking evidence for the potential utility of the weighted likelihood bootstrap (WLB) for simulating non-normal likelihoods or posterior distributions, but their arguments do not provide explanations for this subasymptotic effect. Their implicit claim is that the WLB tracks the posited likelihood of θ, whereas it approximately tracks the posterior distribution of θ under a discrete approximation to the model that generated the data, irrespective of the posited model!

Consider two likelihoods, normal(θ, 1) and exponential(θ), where θ is the population mean, with fixed data. For both posited likelihoods, the value of θ being simulated is the weighted mean of the sample values, where the weights are independent of the posited likelihood; hence, the WLB distribution of θ is the same when the likelihood is normal as when it is exponential. Because the implied WLB specification for the data is a discrete approximation to the model that generated the data, the WLB distribution of θ tends to follow the shape of the posited likelihood for θ when the empirical distribution of the data approximates the model underlying this likelihood. Thus, regardless of which likelihood is posited, if the data look like a normal sample (or exponential sample), the WLB distribution of θ will tend to look like the posterior distribution of θ under a normal likelihood (or exponential likelihood) with a diffuse prior on θ. For large n, both WLB distributions look normal but will only have the correct scales if $\alpha \to 1$, thus implying that the Bayesian bootstrap (BB) specification is the only asymptotically acceptable weight distribution. Because this large sample restriction on the WLB holds in general, in the following general argument we assume a BB weight distribution.

When doing the WLB, a distribution function P with point masses on the observed data points is generated by the BB, denoted $P \sim [P|X, \text{BB}]$. Let $M_0 = \{f_\theta : \theta \in \Theta\}$ denote the posited model and $\mathcal{F}$ the space of all distributions. The maximization step in the WLB is equivalent to finding a θ such that the distance from f_θ to P is minimized. The model assumption M_0 affects the WLB only by inducing a particular mapping from $\mathcal{F}$ to the parameter space Θ, i.e. $\tilde{\theta} = \theta(P)$. Conditional on a fixed data set, and given a fixed function $\tilde{\theta}$, the WLB distribution of $\tilde{\theta}$ is the same for all posited models that induce the same function $\tilde{\theta}$. If the space $\mathcal{F}$ is partitioned into different classes of models indexed by M, then

$$[\theta(P)|X, \text{BB}] = \int [\theta(P)|X, \text{BB}, M]\,[M|X, \text{BB}]\,dM = \int [\theta(P)|X, M]\,[M|X, \text{BB}]\,dM,$$

because $[\theta(P)|X, \text{BB}, M] = [\theta(P)|X, M]$ for all M with positive support for the observed X. Thus, the WLB draws $\tilde{\theta}$ from a posterior distribution that mixes over all possible models under a diffuse prior. If models that are relatively well supported by the data under the BB specification, i.e. models with relatively large values of $[M|X, \text{BB}]$, yield posterior distributions $[\theta(P)|X, M]$ similar to $[\theta(P)|X, M_0]$, then the WLB distribution of $\tilde{\theta}$ will be close to $[\theta(P)|X, M_0]$, which is the targeted posterior distribution of θ.

Albert Y. Lo (State University of New York, Buffalo): The paper demonstrates advantages of the weighted likelihood bootstrap (WLB) for posterior inference in smooth parametric models. The choice of non-Dirichlet weights, i.e. non-exponential Y_i, affects the quality of the WLB approximations and is particularly interesting. It turns out that the accuracy of the WLB depends on the weights only through the coefficient of skewness of Y_1. (This is also found when using non-exponential weights in Rubin's (1981) Bayesian bootstrap; see Lo (1991, 1993).) In the WLB setting, the maximization of the weighted likelihood amounts to finding the roots θ^* of $\Sigma\, Y_i l_i'(\theta) = 0$, where $l_i'(\theta) = (\partial/\partial\theta) \log f(X_i|\theta)$ and $Y_1, \ldots, Y_n$ are independent and identically distributed non-negative random variables. A Taylor argument gives

$$\sigma(Y_1)^{-1} n^{1/2} I_n(\hat{\theta})(\theta^* - \hat{\theta}) = n^{-1/2} \Sigma\{Y_i/\sigma(Y_1)\} l_i'(\hat{\theta}) + R_n. \tag{23}$$

Conditional on the data, the distribution of $n^{-1/2} \Sigma\{Y_i/\sigma(Y_1)\} l_i'(\hat{\theta})\{n^{-1} \Sigma\, l_i'(\hat{\theta})^2\}^{-1/2}$, and hence of $\sigma(Y_1)^{-1} n^{1/2} I_n(\hat{\theta})^{1/2}(\theta^* - \hat{\theta})$, has an expansion

$$F^*(t|\mathbf{x}) = \Phi(t) + \tfrac{1}{6} n^{-1/2}(1-t^2)\,\phi(t)\,\tau(Y_1)\,E\{l_1'(\theta_0)\}^3\,I(\theta_0)^{-3/2} + o(n^{-1/2}) \tag{24}$$

where $I(Y_1)$ is the coefficient of skewness of Y_1. (Assume that R_n decreases to 0 sufficiently fast; see the discussion in Section 4.) Dirichlet weights correspond to $\sigma(Y_1)=1$ and $\tau(Y_1)=2$; equation (24) specializes to equation (4). Define a coefficient of asymptotic accuracy by

$$e_\pi(\Phi, F^*) = \lim_{n\to\infty}\{n^{1/2}|F_n(t|\mathbf{x}) - \Phi(t)|\}/n^{1/2}|F_n(t|\mathbf{x}) - F^*(t|\mathbf{x})|, \tag{25}$$

where $F_n(t|\mathbf{x})$ is the expansion for the posterior distribution obtained by the authors. It follows then that $e_\pi(\Phi, F^*) = |2-\tau(Y_1)|^{-1}$. Hence, for $\tau(Y_1)\in(1, 3)$, the WLB beats the normal approximation to the authors' posterior distributions. WLBs with $\tau=2$ yield the 'most accurate' approximations with $o(n^{-1/2})$ errors (e.g. exponential Y_i). WLBs with $\tau=1$ or $\tau=3$ tie with the normal approximation.

The WLB can also be used to approximate a sampling distribution of the maximum likelihood estimate $\hat\theta$, a property also enjoyed by the nonparametric bootstrap (Efron, 1982; Lo, 1987). In this regard, the WLB is a relative of the 'randomly weighted M-method' of Rao and Zhao (1992). Let us look at the accuracy problem. Use $(1, \hat\theta, \theta_0)$ to play the role of $(Y_i, \theta^*, \hat\theta)$ in the arguments leading to equations (23) and (24). Hence, the sampling distribution of $n^{1/2} I_n(\theta_0)^{1/2}(\hat\theta - \theta_0)$ admits the expansion

$$\begin{aligned}\bar F(t|\mathbf{x}) &= \Phi(t) + \tfrac{1}{6} n^{-1/2}(1-t^2)\,\phi(t)\,E\{l_1'(\theta_0)\}^3\,I(\theta_0)^{-3/2} + o(n^{-1/2}) \\ &= F^*(t|\mathbf{x}) + \tfrac{1}{6} n^{-1/2}(1-t^2)\,\phi(t)\{1-\tau(Y_1)\}\,E\{l_1'(\theta_0)\}^3\,I(\theta_0)^{-3/2} + o(n^{-1/2}),\end{aligned} \tag{26}$$

where the last equality follows from equation (24). The coefficient of asymptotic accuracy is

$$e(\Phi, F^*) = \lim_{n\to\infty}\{n^{1/2}|\bar F(t) - \Phi(t)|\}/n^{1/2}|\bar F(t|\mathbf{x}) - F^*(t|\mathbf{x})| = |1-\tau(Y_1)|^{-1}. \tag{27}$$

For $\tau(Y_1)\in(0, 2)$, the WLB beats the normal approximation. WLBs with $\tau=1$ are the most accurate with $o(n^{-1/2})$ errors (e.g. gamma(4; β) Y_i). WLBs with $\tau=0$ or $\tau=2$ tie with the normal approximation.

Thomas A. Louis (University of Minnesota School of Public Health, Minneapolis): I congratulate Dr Newton and Professor Raftery for their intriguing paper on generalizing the Bayesian bootstrap. Their weighted likelihood bootstrap (WLB) adds another method to our options for generating samples that can be used to construct posterior distributions and Bayesian inferences. The WLB's ability to use standard weighted likelihood maximization routines is a potential benefit, but its success depends strongly on an effective, possibly data-dependent, choice of α. Although Markov chain Monte Carlo (MCMC) methods might be easier to implement and to 'tune' than the WLB, they are certainly no panacea. Situations causing problems in maximizing the weighted likelihood are also likely to produce slow or misdirected MCMC convergence.

I am principally concerned with what we have when the WLB has done its work. Achieving a first-order correct approximation is hardly sufficient: the maximum-likelihood-based Gaussian approximation accomplishes this. A prior-augmented approximation will do even better, to say nothing of more sophisticated Laplace or saddlepoint approximations. It appears that, for most realistic applications, on its own the WLB will do no more than to deliver a sample of parameter values that can be smoothed and used to feed an importance sampling adjustment (ISA). As the authors note, if this sample is not close to the posterior distribution, the ISA will be inefficient and possibly not worth all the work that it took to obtain it.

So, I am left with questions for the authors. In what situations will the WLB approach save time (both human and computer central processor unit) in achieving a desired accuracy of approximation? More specifically, assuming that an ISA will be used, how does the WLB compare with generating samples from the prior distribution, or from the prior-augmented likelihood, or from the mixture distribution advocated by West (1992), etc.? Is the WLB an effective approach for obtaining good starting values for an MCMC method? Until these and a host of other questions and issues are addressed, I view the WLB as an interesting generalization of the Bayesian bootstrap that reveals quite fascinating structures but awaits methodological and applied roles.

Radford M. Neal (University of Toronto): I shall comment on the estimators for the marginal likelihood

in Section 7, describe an alternative method using an importance sampler and refer to past work on this problem in physics.

Of the estimators for $p(x)$ discussed, $\hat{p}_1$ and $\hat{p}_4$ are based solely on the likelihoods of values sampled from the posterior. In many problems, the posterior is determined largely by the likelihood, the effect of the prior being small. Even replacing the prior with an improper distribution will often have little effect on what a typical sample drawn from the posterior looks like. The value of $p(x)$ depends strongly on the prior, however—with an improper prior, it is reduced to 0. In such situations, $\hat{p}_1$ and $\hat{p}_4$ cannot work well. As the authors note, there are also problems with $\hat{p}_2$. I am not convinced that the hybrid of $\hat{p}_3$ avoids these difficulties.

Better results may be obtainable by using an importance sampler that approximates the posterior, perhaps constructed using the weighted likelihood bootstrap as in Section 2. If $\hat{g}$ is the normalized importance sampling density, $p(x)$ may be estimated by

$$\frac{1}{m}\sum_{i=1}^{m} p(\theta^{(i)})\, p(x|\theta^{(i)})/\hat{g}(\theta^{(i)})$$

where $\theta^{(1)}, \ldots, \theta^{(m)}$ are drawn freshly from $\hat{g}$ (points used in constructing $\hat{g}$ should not be reused).

For complex distributions, constructing a good importance sampler becomes infeasible. Techniques developed for the equivalent problem of estimating the 'free energy' of a simulated physical system can be applied, however. Generally, a series of intermediate distributions that connect the posterior to an analytically tractable reference distribution must be introduced (though see Voter (1985)). For the *acceptance ratio* method (Bennett, 1976), successive distributions must overlap significantly. For the *thermodynamic integration* (Bash *et al.*, 1987) and *interpolation* (Bennett, 1976) methods, they need not, but a smoothness assumption must be justified. With *umbrella sampling* (Torrie and Valleau, 1977), the intermediate distributions are implicit; a single simulation run visits all intermediate regions.

Art B. Owen (Stanford University): To maximize equation (1) with respect to θ, assuming regularity, we solve the estimating equations $0 = \Sigma_i w_i \nabla \log f_i(x_i, \theta)$ where ∇ denotes gradient with respect to θ. The usual bootstrap repeatedly solves these equations with $w = (w_1, \ldots, w_n)'$ drawn from a multinomial distribution with parameters n and $(1/n, \ldots, 1/n)'$. The weighted likelihood bootstrap substitutes a continuous distribution for the w_i.

One advantage of continuously distributed weights is that they make certain Monte Carlo variance reduction techniques possible. Graham *et al.* (1990) describe some balanced bootstrap sampling techniques. Their example 4 illustrates a method based on Bose's (1938) construction for mutually orthogonal Latin squares (MOLS). The method may be applied to samples of size $n = p^r$ where p is a prime number and r is a positive integer, but it cannot be used for general sample sizes. By using continuously distributed weights, it is possible to obtain second- and higher order balance through sampling schemes based on orthogonal arrays generated via MOLS and other techniques. Suppose that $w_i = Y_i/\Sigma_j Y_j$ where the Y_i are independent and identically distributed. We can further write $Y_i = g(U_i)$ where $U = (U_1, \ldots, U_n)'$ is uniform on $[0, 1]^n$. For standard exponential Y, $g(u) = -\log u$. Finally suppose that $\int h(U) \Pi_i \, dU_i$ is of interest for some function h. The mean, variance, skewness and cumulative density function of functions of $\tilde{\theta}_n$ may be written this way. Such an integral may be estimated by the mean of h over levels taken from a randomized orthogonal array as described in Owen (1992a). Any set of $n-2$ MOLS can be used, by mapping U_1 onto rows, U_2 onto columns and $U_3, \ldots, U_n$ onto the letters of the orthogonal Latin squares. Bose's (1938) construction provides $p^r - 1$ MOLS of p^r levels (p prime) and can thus be used for any $n \leqslant p^r + 1$. There is no need for the number of levels in the Latin squares to be a multiple of n.

We can use a step function for g to obtain discretely distributed weights. This is not equivalent to the approach taken in Graham *et al.* (1990). In their orthogonal array the elements are indices and not weights.

An expression for the Monte Carlo variance of means over randomized orthogonal array samples is given by Owen (1992b). Some software for generating randomized orthogonal arrays may be obtained from the directory /pub/oa on playfair.stanford.edu.

Another way to obtain first-order correct inferences from these estimating equations is to define a nonparametric profile likelihood ratio function

$$\mathscr{R}(\theta) = \sup\left\{ \prod_{i=1}^{n} n w_i \,|\, 0 = \sum_{i=1}^{n} w_i \nabla \log f_i(x_i, \theta),\ 0 \leqslant w_i \leqslant 1 = \sum w_i \right\}.$$

Owen (1990) showed how standard χ^2-asymptotics often apply to this likelihood ratio and connected it to the posterior in Rubin's (1981) Bayesian bootstrap when a non-informative prior is used and to the nonparametric tilting bootstrap of Efron (1981).

Dongsheng Tu (Academia Sinica, Beijing): I congratulate the authors for this significant advance in Bayesian bootstrap and random weighting methodology. My contribution will focus only on the second-order properties of the proposed weighted likelihood bootstrap (WLB). For this I first review briefly some related results for the frequentist weighted bootstrap. A more detailed review may be found in Tu and Zheng (1991).

A random weighting scheme which is identical with Rubin's Bayesian bootstrap was first proposed by Zheng (1987) as an alternative to Efron's bootstrap for approximating the sampling distribution of pivotal quantities. Tu and Zheng (1987) showed that this approach is second order accurate if the random weights are jointly distributed as Dirichlet(4, 4, . . ., 4) (Weng (1989) also obtained this result). Later, Zheng and Tu (1988) generalized this scheme to linear models to approximate the sampling distribution of least square estimators and also proved the second-order accuracy. In a further study, however, Tu (1986) found that a natural generalization of this method to minimum contrast estimators, which is similar to the approach proposed in this paper for the maximum likelihood estimator, fails to be second order accurate, even though the weights are distributed as Dirichlet(4, 4, . . ., 4). Instead of looking for new weights, a method of transformation was proposed to modify the random weighting method so that it can achieve the second-order accuracy. An example can be found in Tu (1992) for differentiable functional statistics, in which the transformation is estimated by the jackknife, another resampling method.

Now let us return to the WLB proposed in this paper. In Section 4, the authors also found that the WLB fails in general to be second order accurate. I believe that the above-mentioned idea involving transformation can also be applied to improve the accuracy of the WLB, i.e. the posterior distribution of $\sqrt{\{n I_n(\hat{\theta})\}}(\theta - \hat{\theta}_n)$ may be approximated by the conditional distribution of $H[\sqrt{\{n I_n(\hat{\theta}_n)\}}(\tilde{\theta}_n - \hat{\theta}_n)]$, where the transformation H depends on θ_0, which can be estimated by $\hat{\theta}_n$, and prior π. The functional form of H can be found by comparing the asymptotic expansions of $F_n(t)$ and $\tilde{F}_n(t)$. By this adjustment we may be able to approximate the posterior distribution of θ for any given prior π with second-order accuracy. I hope that the explicit form of H can be worked out in the near future.

The **authors** replied later, in writing, as follows.

What are the advantages of the weighted likelihood bootstrap (WLB)? Or, as Louis puts it, when will it save human and central processor unit (CPU) time relative to competitors, mainly Markov chain Monte Carlo (MCMC) and analytical approximations? As summarized by Gilks, we made the case that

(a) the WLB can be rapidly deployed, requiring little programming (see the experience of Bates and Ritter),
(b) it seems to work well for 'small' (i.e. up to at least 18 parameters) but intricate problems,
(c) it generates independent posterior samples and
(d) it is self-monitoring, in that a highly dispersed distribution of the importance weights indicates that it is not working well.

The discussants have added two more, namely

(e) it will tend not to miss minor modes in low dimensional multimodal likelihoods (Roberts) and
(f) it is 'more accurate' than the normal approximation under quite general conditions (Lo).

Liu and Rubin suggest that the raw WLB output is informative by itself, as a robustified form of parametric inference.

With respect to (e), Sweeting reports poor performance of the WLB for a multimodal likelihood based on six observations. However, his use of the WLB is not what we recommend in practice. We reanalysed Sweeting's first example with $\alpha = 1.7$ to spread out the weights and found a much better distribution of $\tilde{\theta}$s between the two modes. On sampling–importance resampling (SIR) adjustment, we obtained a very close approximation to the true posterior. Further, if the WLB puts mass where no real mode is, then that mass will be downweighted on SIR adjustment.

Roberts suggests that the natural home for the WLB is in non-linear regression problems, presumably including generalized linear models, robust regression, survival analysis and autoregressive models for

time series. The list of examples so far treated suggests the scope to be somewhat wider than this, and Gilks in his equation (18) makes a fascinating suggestion for extending the scope of the WLB to hierarchical models with very many parameters. We look forward to further investigation of this idea.

Barnard hopes that the method will be tested on a realistic problem, and so do we. Bacha and Celeux have told us that their discussion is part of a larger project with Electricité de France on reliability in nuclear power-stations; this may eventually provide such a major application. The first circulated version of our paper (Newton and Raftery, 1991) gives several other examples that were removed because of space constraints: logistic regression, Markov chains, normal mixtures, spectral analysis and calibration. This is still available on request from us.

Leonard and Hsu and Louis mention analytical approximations: Laplace, saddlepoint and prior augmented. These will usually use less CPU time than the WLB (or any other Monte Carlo method) for the same accuracy, but they will often require much more human time than the WLB.

Comparison with Markov chain Monte Carlo methods

The WLB produces independent samples from the posterior, and so avoids the convergence issues that are inherent in MCMC methods. The WLB is often easier to apply than the Gibbs sampler, although, as Clifford points out, the larger class of Metropolis–Hastings algorithms has members that are faster and easier to implement than the Gibbs sampler, and perhaps the Gibbs sampler itself should be avoided. The WLB will not beat MCMC methods when optimization is difficult.

The WLB has only one user-specified control parameter, and thus may spare *human* time. Gibbens and Roberts apply the MCMC independence sampler with normal proposals to our first example, which has $p=3$ parameters. Their algorithm has nine (or more generally $p(p+3)/2$) control parameters. The independence sampler is closely related to importance sampling (Smith and Roberts (1993), section 3.3), and so we would expect its performance to be sensitive to the control parameters. Thus choosing them well is both important and potentially difficult, especially in higher dimensions (such as $p=18$!). Similar comments apply to the MCMC algorithms used by Clifford, and Gelfand and Mallick.

Gelfand and Mallick claim that their MCMC algorithm works better than the WLB in their example, based on plots of importance weights. However, these plots are misleading because the WLB weights are for the unadjusted WLB sample, whereas the MCMC weights are for the final sample, which is claimed to be from the posterior. A fairer comparison would involve the importance weights for the final SIR-adjusted sample from the WLB. Indeed, it is surprising that their weights were so far from constant, casting some doubt on the quality of their approximation to the posterior.

Gelman points out an important possible synthesis of WLB and MCMC methods, namely that the (unadjusted) WLB with $\alpha>1$ can provide an overdispersed distribution from which to generate starting values for MCMC algorithms. Even if the WLB or its generalization in Gilks's equation (18) does not yield a good approximation to the posterior, it may provide a satisfactory and routinely available overdispersed distribution of starting values.

Choice of α

As Grieve and Louis point out, the choice of α is crucial, and Bacha and Celeux's example shows how inefficient the WLB can be with a very bad choice of α. We have suggested looking at the cumulative importance weight plots (e.g. Fig. 6(a)) for various values of α. This can be formalized by choosing the value of α for which the area Δ between the constant and WLB cumulative importance weight curves is smallest. This approach is not foolproof, although *looking* at the cumulative importance weight curves is useful. This is because, for the SIR adjustment to work well, we need the initial WLB sample to cover the entire posterior distribution, and so we should be more tolerant of non-constant cumulative importance weight curves that are due to overdispersion than to underdispersion. In Bacha and Celeux's example, Δ is calculated for six values of α from 0.4 to 1.4 and is smallest for $\alpha=0.8$. However, this misses part of the posterior, leading to underestimated variances and so on; a higher value of α gave a better performance even though Δ was slightly larger.

We would like to suggest a different approach to the choice of α. The idea is that the WLB sample should cover the posterior fully; if so, then we expect the largest importance weights u_j to be near the middle of the WLB sample. This can be assessed by plotting the u_j against the Mahalanobis distances from the sample mean, d_j. We have found the following *10–90% rule* to work fairly well: choose the smallest α such that the largest weights that sum to 10% are in the lower 90% of the distribution of d_j. We have also found it useful to calculate $\mathrm{corr}(u_j, d_j)$, restricted to values such that $u_j>2/m$. This

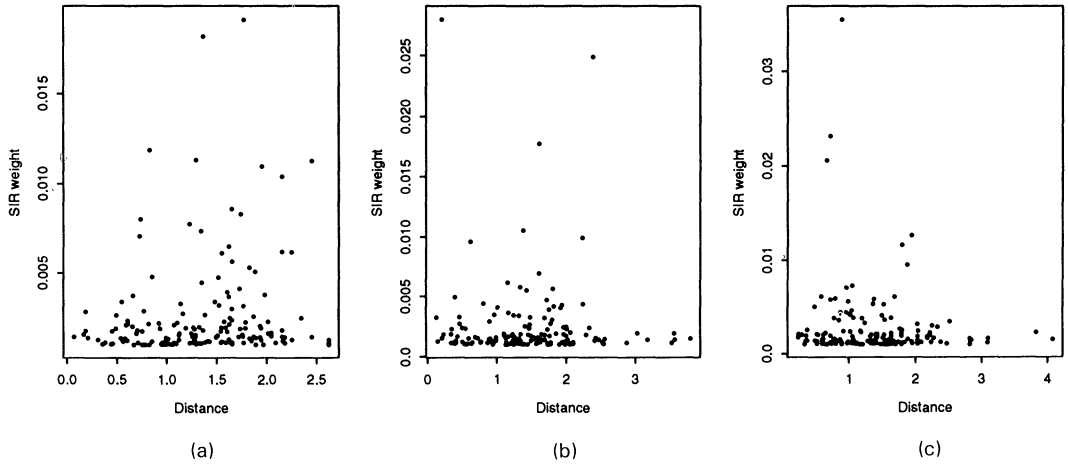

Fig. 16. Importance weights u_j plotted against Mahalanobis distances from the mean of the WLB sample, d_j, for (a) $\alpha = 0.4$, (b) $\alpha = 1.2$ and (c) $\alpha = 1.4$, in Bacha and Celeux's example: only weights greater than twice the average weight (i.e. $u_j > 2/m$) are shown

should be negative, and in our experience good values of α correspond to $\text{corr}(u_j, d_j) < -1/\sqrt{m^*}$, where $m^* = \#\{j: u_j > 2/m\}$. These rules of thumb need further investigation and refinement.

With the help of Mostafa Bacha, we applied these ideas to the two-parameter example of Bacha and Celeux. Fig. 16 shows improvement as α increases. Table 3 shows that only $\alpha = 1.4$ satisfies the 10–90% rule. For these αs we have $\text{corr}(u_j, d_j) = 0.15, -0.03, -0.10$ with $m^* = 160$, so that again only $\alpha = 1.4$ satisfies $\text{corr}(u_j, d_j) < -1/\sqrt{m^*}$. It turns out that, empirically, $\alpha = 1.4$ is indeed a better choice.

Modifications

Silverman notes that the conservative kernel method we adopt may still undersmooth the tails, and so an adaptive kernel smoother and a larger sample size are advised. Leonard and Hsu suggest that we use the weighted likelihood function itself to form importance weights, thus avoiding time consuming kernel density estimation. The importance of this idea can be seen through Cui, Sherman and Tanner's 'reinterpretation'. Our $\tilde{\beta}$ is simply the mode of the β^*-distribution. Further, the covariance of the normal approximation for β^* is generally a by-product of the maximization routine to compute $\tilde{\beta}$, making such a modified WLB potentially quite fast and accurate.

TABLE 3
Largest importance weights in Bacha and Celeux's example†

$\alpha = 0.4$		$\alpha = 1.2$		$\alpha = 1.4$	
Weight	*Distance percentile*	*Weight*	*Distance percentile*	*Weight*	*Distance percentile*
0.019	23	0.028	98	0.035	77
0.018	48	0.025	7	0.023	86
0.012	78	0.018	37	0.021	87
0.011	53	0.011	50	0.013	18
0.011	2	0.010	10	0.012	23
0.011	13	0.010	89		
0.010	6				
0.009	31				

†The distance percentile is the percentile of the observation in the distribution of Mahalanobis distances d_j from the mean of the WLB sample. Small distance percentiles correspond to large distances.

We agree with Leonard and Hsu that some estimate of Monte Carlo error must accompany the WLB output. Owen's suggestions for balancing the weights may be an efficient way to reduce this error.

In the non-linear regression examples, a fraction of weight vectors produced no $\tilde{\theta}$-values owing to ill conditioning of some kind. Bates and Ritter ask for guidance here, but all that we can suggest is to ignore those cases. We can try to tune the optimization code to nurse convergence in some cases, but this does not always work. Retrospective adjustment by SIR may alleviate problems.

Connections to other bootstraps

Liu and Rubin suggest that the law of $\tilde{\theta}$ (unadjusted) is a meaningful posterior distribution in its own right. We tried to formulate this but decided instead to view the WLB as an approximation to parametric inference. The WLB nonparametrically robustifies a parametric posterior by averaging over all models which induce the same map $\tilde{\theta}$. The fact that different models induce the same map is also noticed by Davison and Grieve, and can be viewed as a positive feature. Liu and Rubin's idea is compelling, but care is needed, for example, to ensure that $[\theta(P)|X, \mathrm{BB}, M]$ can be equated to $[\theta(P)|X, M]$ without regard to discreteness in the former. Künsch's suggestion that model-independent inference can be assured under a different weight construction implies that Liu and Rubin's ideas may carry over to time series.

Suppose that the true law governing an independent and identically distributed sample is supported on the observed sample points. A point P in this model is equivalent to a vector λ of n probabilities summing to 1. The nonparametric likelihood of P is simply $L(P)=\Pi_{i=1}^{n}\lambda_i$. A parameter θ depends on P by $\theta=t(P)$. Owen's empirical likelihood of θ results from profiling $L(P)$ at each θ. The Bayesian bootstrap is equivalent to integrating the posterior $L(P)\,\Pi\lambda_i^{-1}$ at each θ. The WLB basically provides the transformation t.

Worton asks about the relationship to bootstrap likelihood, which is presented as an approximate partial likelihood $p\{t(\hat{P})|\theta\}$ where $\hat{P}$ is the observed estimator of P. Suppressed in the notation is the fact that this probability depends on the true λ which means that it is not actually a partial likelihood. To overcome this, a rather odd assumption is made that a different true λ exists for each θ. The double bootstrap is then invoked to estimate the resulting 'true bootstrap likelihood' at each θ. It seems more natural to use the double bootstrap through prepivoting (Beran, 1988) to produce an approximate pivot $h_n(x, \theta)$ having density g and then to form the partial likelihood $g\{h_n(x, \theta)\}$ for fixed x.

The connection between the WLB and Laird and Louis's bootstrap is that they are both bootstrap procedures applied in Bayesian inference. Carlin suggests that effective priors form another connection. Perhaps any frequentist procedure is Bayesian under some effective prior, and so the notion may not be particularly useful. A more cogent issue for future study is the formation of approximate pivots. In fact the bias correction method of Carlin and Gelfand (1990) is an application of Beran's prepivoting (Newton, 1991). Whereas a classical pivot is a function of data and parameters having a known sampling distribution, we can define a Bayesian pivot as a function of data and parameters having a posterior distribution which is independent of the data.

Theory

In response to Roberts, theorem 1 is proved by using a generalization of the argument of Foutz (1977) to establish consistency of the maximum likelihood estimator (MLE). Theorem 2 is proved by extending the Cramér proof of normality of the MLE. Essentially the same smoothness conditions are required. Lo generalizes our second-order expansion (providing an answer to Grieve). Tu observes the importance of a proper scale for the parameter. The WLB works better on some scales than on others, so transformation to approximate normality is advised.

Model choice

Since we circulated the first version of our paper (Newton and Raftery, 1991), considerable research has been done on evaluating Bayes factors by simulating from the posterior; see Kass and Raftery (1993) for a review. Rosenkrantz (1992) evaluated the estimators of the marginal likelihood, $\hat{p}_1(x)$, $\hat{p}_2(x)$, $\hat{p}_3(x)$ and $\hat{p}_4(x)$ from Section 7 of our paper, in the contexts of hierarchical Poisson models, outlier identification and a multinomial model with latent variables. She found that analytical approximations via the Laplace method give greater accuracy for much less computation *and* human time than the posterior simulation estimators; the problem is that the Laplace method is not always applicable. Among the posterior simulation estimators, she found that $\hat{p}_3(x)$ with a large value of δ (close to 1) had the best performance. This differs from the earlier advice given in our paper where we recommended a small value of δ.

To respond to the discussants' comments, we need some notation. Let $\|X\|_h = m^{-1}\Sigma X(\theta^{(j)})$, where $\{\theta^{(j)}\}$ is a sample of size m from the density $h/\int h$. Let L denote the likelihood, π the prior, 'post' the posterior, g a positive function and f a normalized density. Then the general importance sampling estimator of $p(x)$, with importance sampling function g, is $\|L\pi/g\|_g/\|\pi/g\|_g$; when g is normalized, this is just $\|L\pi/g\|_g$, as Neal points out. When $g = L\pi$ this becomes $\hat{p}_1(x) = \|1/L\|^{-1}_{\text{post}}$, whereas when g is the prior we obtain $\hat{p}_2(x) = \|L\|_\pi$.

A simple modification of the harmonic mean estimator $\hat{p}_1(x)$ is $\hat{p}_5(x) = \|f/L\pi\|^{-1}_{\text{post}}$; this is mentioned by Gelfand and Dey (1993). It is unbiased and simulation consistent, and has a central limit theorem if the tails of f are sufficiently thin, specifically if $\int\{f^2/L\pi\} < \infty$. If θ is one dimensional, if the prior and posterior distributions are both normal and if f is normal with mean equal to the posterior mean and variance equal to $\varkappa$ times the posterior variance, then the mean-squared error of $\hat{p}_5(x)$ is minimized when $\varkappa = 1$. This suggests that high efficiency is most likely to result if f is roughly proportional to the posterior density. In a small numerical study that we described at the meeting, $\hat{p}_5(x)$ had very good performance.

Meng and Wong (1993), via Gelman, propose the alternative $\hat{p}_6(x) = \|L\pi g\|_\pi/\|\pi g\|_{\text{post}}$; we look forward to further investigation of its properties. As a practical matter, it is somewhat awkward because it involves simulation from the prior as well as the posterior.

We thank the discussants for their insightful comments and regret not addressing all the points raised.

REFERENCES IN THE DISCUSSION

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Models in GLIM*, p. 35. Oxford: Oxford Scientific.

Bash, P. A., Singh, U. C., Langridge, R. and Kollman, P. A. (1987) Free energy calculations by computer simulation. *Science*, **236**, 564–568.

Bennett, C. H. (1976) Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.*, **22**, 245–268.

Beran, R. (1988) Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J. Am. Statist. Ass.*, **83**, 687–697.

Bose, R. (1938) On the application of the theory of Galois fields to the problem of construction of hyper-Graeco-Latin squares. *Sankhya*, **3**, 323–338.

Carlin, B. P. and Gelfand, A. E. (1990) Approaches for empirical Bayes confidence intervals. *J. Am. Statist. Ass.*, **85**, 105–114.

——(1993) Parametric likelihood inference for record breaking problems. *Biometrika*, **80**, in the press.

Daniels, H. E. and Young, G. A. (1991) Saddlepoint approximation for the Studentized mean, with an application to the bootstrap. *Biometrika*, **78**, 169–179.

Davison, A. C. and Hinkley, D. V. (1988) Saddlepoint approximations in resampling methods. *Biometrika*, **75**, 417–431.

Davison, A. C., Hinkley, D. V. and Worton, B. J. (1992) Bootstrap likelihoods. *Biometrika*, **79**, 113–130.

Efron, B. (1981) Nonparametric standard errors and confidence intervals (with discussion). *Can. J. Statist.*, **9**, 139–172.

——(1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Foutz, R. V. (1977) On the unique consistent solution to the likelihood equation. *J. Am. Statist. Ass.*, **72**, 147–148.

Gelfand, A. E. and Dey, D. K. (1993) Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc.* B, to be published.

Gelfand, A. E., Smith, A. F. M. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems. *J. Am. Statist. Ass.*, **87**, 523–532.

Gelman, A. and Meng, X. L. (1993) Path sampling for computing normalizing constants. *Technical Report*. Department of Statistics, University of California, Berkeley.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Geweke, J. (1989) Bayesian inference in econometric models, using Monte Carlo integration. *J. Econometr.*, **57**, 1317–1339.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993) Modelling complexity: applications of Gibbs sampling in medicine. *J. R. Statist. Soc.* B, **55**, 39–52.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994) A language and program for complex Bayesian modelling. *Statistician*, **43**, in the press.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Graham, R. L., Hinkley, D. V., John, P. W. M. and Shi, S. (1990) Balanced design of bootstrap simulations. *J. R. Statist. Soc.* B, **52**, 185–202.

Kass, R. E. and Raftery, A. E. (1993) Bayes factors and model uncertainty. *Technical Report 254*. Department of Statistics, University of Washington, Seattle.
Künsch, H. R. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, **17**, 1217–1241.
Laird, N. M. and Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples. *J. Am. Statist. Ass.*, **82**, 739–757.
Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
Leonard, T. (1982) Comment on "A simple predictive density function". *J. Am. Statist. Ass.*, **77**, 657–658.
Leonard, T. and Hsu, J. S. J. (1992) Bayesian inference for a covariance matrix. *Ann. Statist.*, **20**, 1669–1696.
Leonard, T., Hsu, J. S. J. and Tsui, K. W. (1989) Bayesian marginal inference. *J. Am. Statist. Ass.*, **84**, 1051–1058.
Lo, A. Y. (1987) A large sample study for the Bayesian bootstrap. *Ann. Statist.*, **15**, 360–375.
———(1991) Bayesian bootstrap clones and a biometry function. *Sankhya* A, **53**, 320–333.
———(1993) Bayesian method for weighted sampling models. *Ann. Statist.*, **21**, in the press.
Mallick, B. K. and Gelfand, A. E. (1993) Doubly semiparametric proportional hazards models. Submitted to *J. Am. Statist. Ass.*
Meng, X. L. and Wong, W. H. (1993) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Technical Report 365*. Department of Statistics, University of Chicago, Chicago.
Müller, P. (1991) A generic approach to posterior integration and Gibbs sampling. *Technical Report 91-09*. Department of Statistics, Purdue University, West Lafayette.
Newton, M. A. (1991) The weighted likelihood bootstrap and an algorithm for prepivoting. *PhD Dissertation*. Department of Statistics, University of Washington, Seattle.
Newton, M. A. and Raftery, A. E. (1991) Approximate Bayesian inference via the weighted likelihood bootstrap. *Technical Report 199*. Department of Statistics, University of Washington, Seattle.
Owen, A. B. (1990) Empirical likelihood confidence regions. *Ann. Statist.*, **18**, 90–120.
———(1992a) Orthogonal arrays for computer experiments, integration and visualization. *Statist. Sin.*, **2**, 439–452.
———(1992b) Lattice sampling revisited: Monte Carlo variance of means over randomized orthogonal arrays. Submitted to *Ann. Statist.*
Rao, C. R. and Zhao, L. C. (1992) Approximation to the distributions of M-estimates in linear models by randomly weighted bootstrap. *Sankhya* A, **54**, 323–333.
Rosenkrantz, S. (1992) The Bayes factor for model evaluation in a hierarchical Poisson model for area counts. *PhD Dissertation*. Department of Biostatistics, University of Washington, Seattle.
Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, ch. 5. London: Chapman and Hall.
Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.
Tanner, M. A. (1991) *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. New York: Springer.
Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Statist. Ass.*, **82**, 528–550.
Terrell, G. R. (1990) The maximal smoothing principle in density estimation. *J. Am. Statist. Ass.*, **85**, 470–477.
Tierney, L. (1991) Markov chains for exploring posterior distributions. *Technical Report 560*. School of Statistics, University of Minnesota, Minneapolis.
Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–86.
Tierney, L., Kass, R. E. and Kadane, J. B. (1989) Approximate marginal densities of non-linear functions. *Biometrika*, **76**, 425–433; correction, **78** (1991), 233.
Torrie, G. M. and Valleau, J. P. (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.*, **23**, 187–199.
Tu, D. (1986) On the asymptotic expansions relating to the randomly weighted statistics of minimum contrast estimators. *Technical Report*. Institute of Systems Science, Academia Sinica, Beijing.
———(1992) Approximating the distribution of a general standardized functional statistic with that of jackknife pseudo values. In *Exploring the Limits of Bootstrap* (eds R. LePage and L. Billard), pp. 279–306. New York: Wiley.
Tu, D. and Zheng, Z. (1987) On the Edgeworth's exansions of random weighting method. *Chin. J. Appl. Probab. Statist.*, **3**, 340–347.
———(1991) Random weighting: another approach to approximate the unknown distributions of pivotal quantities. *J. Combin. Inform. Syst. Sci.*, **16**, 249–270.
Voter, A. F. (1985) A Monte Carlo method for determining free-energy differences and transition state theory rate constants. *J. Chem. Phys.*, **82**, 1890–1899.
Weng, C. S. (1989) On a second-order property of the Bayesian bootstrap. *Ann. Statist.*, **17**, 705–710.
West, M. (1992) Modelling with mixtures. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.
Zheng, Z. (1987) Random weighting methods. *Acta Math. Appl. Sin.*, **10**, 247–253.
Zheng, Z. and Tu, D. (1988) Random weighting method in regression models. *Sci. Sin.* A, **31**, 1442–1459.