Order Number 9131689

The weighted likelihood bootstrap and an algorithm for prepivoting

Newton, Michael Abbott, Ph.D.

University of Washington, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

The Weighted Likelihood Bootstrap and an Algorithm for Prepivoting

by

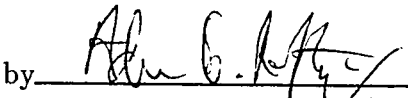
Michael A. Newton

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

1991

Approved by 
(Chairperson of Supervisory Committee)

Program Authorized
to Offer Degree Statistics

Date June 11 1991

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Michael Newton

Date June 11- 1991

University of Washington

Abstract

The Weighted Likelihood Bootstrap and an Algorithm for Prepivoting

by Michael A. Newton

Chairperson of Supervisory Committee: *Professor Adrian E. Raftery*
Department of Statistics

The method of bootstrapping, which has transformed the theory and practice of frequentist statistical inference, is applicable within the Bayesian paradigm. Rather than simulating data that might have been observed, this Bayesian extension, called the weighted likelihood bootstrap, involves simulating parameters corresponding to distributions that might have generated the observed data. The weighted likelihood bootstrap is an extension of earlier work by D. Rubin (Annals of Statistics, 1981) from purely nonparametric models into semi and fully parametric models for data. The resulting simulation, which is viewed as simply a Monte Carlo approximation to a posterior distribution of interest, has desirable asymptotic properties. This simulation method produces easily generated samples from a posterior under an effective prior which can be identified either exactly or approximately in certain models. The simulation is straightforward, requiring only an algorithm for maximum likelihood estimation. It is also closely related to frequentist bootstrapping procedures. The weighted likelihood bootstrap is applied to a wide variety of statistical models.

The prepivoting procedure is studied in a general modeling framework and an efficient Monte Carlo algorithm, called bootstrap recycling, is introduced. This algorithm is shown to be simulation consistent; that is, it produces a closer closer approximation to the right answer as the amount of computing resources gets large. This new algorithm, which is an alternative to the iterated bootstrap, is applied to the likelihood ratio test of a sparse contingency table, and to the construction of likelihood based confidence sets in a complex stochastic model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My view of statistics in particular, and of scientific thinking in general, has enjoyed its most profound development during the last five years while I have studied in the Statistics Department here at the University of Washington. I have been fortunate to work on interesting problems with impressive thinkers in an environment skillfully maintained to nurture good statistical research. I am indebted to many people for making my life and work here so enjoyable and for ensuring that this project was brought fruition.

With no exceptions, the staff members of the department, Cheryl Cronk, Daniel Ijiomah, Lorie Lucky, and Kristin Sprague, have done their utmost to help me in every way. I am also grateful to the computer people Bill Dunlap, Jim Flanagan, Alice Kelly, and Phil Neal for helping me with the numerous problems encountered in completing my computer–dependent dissertation.

Many graduate students have influenced my work, perhaps Charlie Geyer and Nhu Le being the most notable. Both have contributed a great deal to my understanding of prepivoting, the iterated bootstrap, and statistics generally.

The great diversity of skills and interests of the faculty has made studying statistics as a student simultaneously frustrating and challenging. While the poles of asymptotic theory, computing theory, data analysis, and so on provide many opportunities for graduate students, they also make it difficult to assess the quality and relevance of ones own work. My supervisors have guided me well in this regard. Jon Wellner has shown me the importance of clarity and sound theoretical reasoning. My understanding of bootstrap asymptotics, for example, owes much to his seminar series of Spring quarter, 1990. Ever concerned with the basic precepts of statistical inference, Peter Guttorp has taught me much about inference and about stochastic processes. My best training in data analysis, computation, and stochastic modeling came through

our joint work with Jan Abkowitz and Michael Linenberger of the Department of Hematology. David Mason, with his unlimited enthusiasm for tackling hard technical problems, has also taught me about the bootstrap and about writing papers. Don Percival taught me spectral analysis and has helped me many times in my association with the Applied Physics Laboratory.

The vast majority of my thesis work has grown out of regular discussions with my advisor Adrian Raftery. I am indebted to him for his many good ideas, his constant encouragement, and his expert guidance.

My research was funded in largest part by the Applied Physics Laboratory at the University of Washington. Partial funding came from ONR contract M00014-81-K-0095 P00010.

To my family:
Marianna,
Mom, Dad,
Elizabeth, Dennis, Emily, Jessica,
Christopher, Sandra, Laura Jane, and Andrew.

Chapter 1

THE WEIGHTED LIKELIHOOD BOOTSTRAP

1.1 Introduction

Much recent research has focused on the computational problems of Bayesian inference, and while advances such as Laplace approximations (Tierney and Kadane, 1986; Tierney, Kass and Kadane, 1989) and Gibbs sampling (Gelfand and Smith, 1990) are crucial, there is undoubtedly scope for further investigation. In this chapter we introduce a bootstrap procedure for approximate simulation of posterior distributions. Although this procedure, called the *weighted likelihood bootstrap*—hereafter WLB—is quite generally applicable, it may be most useful in problems where other methods are difficult to apply. For example, in many regression models the full conditional distributions used by the Gibbs sampler are not easily available, and in other models the Hessians required by the Laplace method may be irksome to calculate. In some problems, such as those to which the EM algorithm or iteratively reweighted least squares are commonly applied, evaluating the likelihood function may be tedious or impossible, even though there are simple ways to find the maximum likelihood estimator; in such cases the WLB allows us to approximate the posterior distribution without ever evaluating the likelihood function.

The WLB starts with the simulation of a large number of weight vectors having a distribution determined by the data analyst. Each weight vector is associated with a weighted likelihood function which deviates somewhat from the actual likelihood function. The empirical distribution formed by the maximizers of these weighted likelihoods is used to approximate the posterior distribution of the parameter. Consequently, approximate Bayesian inference by simulation is straightforward in models where maximum likelihood estimation is feasible. The distribution of the weight vectors determines the accuracy of the approximation, and although many choices are possible we concentrate on Dirichlet distributed weights in this paper. The WLB

procedure is described in detail in Section 1.2.

The WLB provides a solution that is exact up to an unknown *effective prior.* This function is not a prior in the usual sense of the term. It does not represent prior ignorance or knowledge of the unknown parameter, and in fact it may depend on the sample size and the data themselves. The effective prior is merely the function which modifies the likelihood to produce the distribution which we sample by the WLB. This prior depends heavily on the chosen distribution of weights, and one hope is that in a given problem, we can find a weight distribution leading to a flat effective prior. In certain special models the effective prior can be identified (without simulation), and sometimes we can make an educated guess at this prior (see the study of discrete data models in Section 1.4). For the canonical choice of weight distribution, an asymptotic approximation to this effective prior can be worked out in a special case (see Section 1.6). A data-dependent modification of the flat weight distribution can be derived from maximum entropy arguments (see Section 1.6). Indeed, further research on the WLB may focus on the choice of weight distribution yielding a prescribed effective prior.

Asymptotic validity of the WLB is guaranteed in a class of sufficiently smooth models where maximum likelihood estimates are computed as roots of a likelihood equation. Roughly, the WLB, in its canonical formulation, captures the mean and variance of the posterior distribution, and sometimes captures the skewness. Precise results are stated in Section 1.3 while a full development of the theory is given in Chapter 3.

The WLB yields exact Bayesian inference in multinomial and Markov chain models when the parameter space is unconstrained, in the sense that a data-independent effective prior (actually the conjugate prior) can be identified. In a general multinomial (or Markov chain) model the WLB amounts to sampling probability vectors from an unconstrained posterior and then projecting them onto the model of interest. This projection induces a distribution on the parameter space which is close to a posterior distribution in a precise technical sense. Discrete data models are studied closely in Section 1.4.

The great attraction of the WLB is its ease of application. Code to compute maximum likelihood estimates can be used immediately to estimate the entire posterior distribution; actual calculation of the likelihood function is not necessarily required.

Thus the WLB can be used in conjunction with the EM algorithm (Dempster *et al.*, 1977) or iteratively reweighted least squares (Green, 1984). No specialized code, like that required to run a Gibbs sampler or to calculate the Laplace approximation, is needed, although the resulting simulation is not exact. Implementation of the WLB using some standard optimization methods is studied in Section 1.5, and Chapter 2 contains a number of examples of its use. The WLB approximation is quite good for moderate to large sample sizes.

The WLB may also be of interest for frequentist inference because it yields an approximation to the likelihood function, thus allowing approximate non-Bayesian likelihood-based inference. By comparison with, for example, the nonparametric bootstrap of Efron (1979), the WLB avoids the problem of whether to resample cases or residuals in linear regression, and it gives satisfactory results in logistic regression, where the Efron bootstrap estimate of variance is always infinite—see Section 1.4.3. It also provides solutions to the prediction and calibration problems (Sections 2.4 and 2.5) that avoid the underestimation of uncertainty that occurs in most frequentist methods because of not taking account of uncertainty about model parameters (Aitchison and Dunsmore, 1975).

The WLB is called a *bootstrap* procedure because of its connections with the classical bootstrap of Efron (1979) and also because it generalizes the less well known Bayesian bootstrap of Rubin (1981). In fact, all of these procedures can be viewed as weighted likelihood bootstraps where we vary the weight distribution and the form of the likelihood. In connecting the classical bootstraps with our procedure, this general view uses the *empirical likelihood* of Owen (1988). In addition, there is a close connection to the bootstrap method of Künsch (1989) for time-series, and so Bayesian inference for non-Gaussian, nonlinear, or state-space time series models can be addressed with the WLB. These connections are expounded in Section 1.7, where we also show a simple comparison between the WLB and Efron's bootstrap.

Interestingly, a fundamentally different type of bootstrap procedure is suggested by weighting a partial likelihood. Chapter 4 gives a preliminary study of this special bootstrap for the proportional hazards model of Cox (1972). Consideration of a bootstrap procedure as a method of randomly weighting observations has spawned some interesting theoretical investigations into such random weighting methods. Mason and Newton (1990) study consistency properties of bootstrapped means for the class

of exchangeable weights. Præstgaard (1991) studies a randomly weighted empirical process and proves general results comparable to those of Giné and Zinn (1990) for the regular bootstrap. Haeusler, Mason and Newton (1991) give a survey.

1.2 The method

Suppose data $X_1^n := (X_1, X_2, \cdots, X_n)$ are modeled by a family of distributions having densities f_θ with respect to some dominating measure on the sample space. In the notation, we suppress the dependence of these densities upon n. The parameter space Θ indexes the family and each f_θ determines a joint distribution for the random vector X_1^n. The likelihood function of θ is defined as

$$L_n(\theta) := f_\theta(x_1, x_2, \cdots, x_n)$$

where $x_1, x_2, \cdots, x_n$ are the realized data. Factoring the joint density, we can write

$$L_n(\theta) = \prod_{i=1}^{n} f_{\theta,i}\left(x_i | x_1^{i-1}\right). \tag{1.1}$$

For $i \geq 2$, $f_{\theta,i}\left(x_i | x_1^{i-1}\right)$ is the conditional density of X_i given $X_1^{i-1} = x_1^{i-1}$ evaluated at x_i. The first factor in equation (1.1) is the marginal density of X_1 at x_1. Of course, different factorizations of the joint density are possible when the data are modeled as dependent. In time series models, the time order suggests a natural factorization.

The likelihood function is formed by contributions from each data point x_i. In a sense, these contributions are equal because the power to which the density at x_i is raised is the same for each i. Although perhaps a strange sense of equality, there is much to be gained by varying this contribution of each observation to the likelihood. Our development hinges on the construction of Dirichlet distributed random vectors.

Recall that if $Y_1, Y_2, \cdots, Y_m$ are independent Gamma random variables with shape parameters $\alpha_1, \alpha_2, \cdots, \alpha_m$ and a common scale parameter, and $S_m := \sum_i Y_i$, then the vector $S_m^{-1}(Y_1, Y_2, \cdots, Y_m)$ is said to have a $m-$dimensional Dirichlet distribution with parameters $\alpha_1, \cdots, \alpha_m$. Notationally,

$$\frac{1}{S_m}(Y_1, Y_2, \cdots, Y_m) \sim \text{Dirichlet}_m(\alpha_1, \alpha_2, \cdots, \alpha_m).$$

When all the parameters α_j are equal to 1, the vector has a uniform Dirichlet distribution. Some useful properties of Dirichlet random vectors are reviewed in the first section of Appendix A.

The *weighted likelihood function* is defined as

$$\tilde{L}_n(\theta) := \prod_{i=1}^{n} \left[f_\theta \left(x_i | x_1^{i-1} \right) \right]^{w_{n,i}} \tag{1.2}$$

where $w_n := (w_{n,1}, w_{n,2}, \cdots, w_{n,n})$ is n times an $n-$dimensional Dirichlet random vector which is independent of the data. Whereas the likelihood $L_n(\theta)$ is a fixed function after the data are observed, the weighted likelihood $\tilde{L}_n(\theta)$ has randomness induced by the Dirichlet weights. Unless stated otherwise, we choose $\alpha_i = 1$ for all i because this case is the direct generalization of Rubin's Bayesian bootstrap.

We are interested in the parameter value that maximizes the weighted likelihood function and so we denote by $\tilde{\theta}_n$ any parameter value satisfying

$$\tilde{L}_n(\tilde{\theta}_n) \geq \tilde{L}_n(\theta) \qquad \text{for all } \theta \in \Theta.$$

Our thesis is that the conditional distribution of $\tilde{\theta}_n$ (given the data) often provides a good approximation to a posterior distribution of θ. Although this conditional distribution usually eludes analytic calculation, it can be estimated by simulation whenever maximum likelihood estimation is feasible. This simulation amounts to repeatedly sampling Dirichlet weight vectors and then maximizing $\tilde{L}_n(\theta)$ for each such vector.

Consider the following, much-studied linkage example from genetics (Rao, 1973; Dempster *et al.*, 1976; Tanner and Wong, 1989). Independently for $i = 1, 2, \cdots, n = 197$, we observe multinomial observations X_i on four classes with success probabilities $p_j(\theta)$ given by

$$(p_1, p_2, p_3, p_4) = \frac{1}{4}(2 + \theta, 1 - \theta, 1 - \theta, \theta)$$

for $\theta \in [0, 1]$. Observed cell counts are $(125, 18, 20, 34)$. From the definition in equation (1.2), we have

$$\tilde{L}_n(\theta) \propto \left((1 + \theta)^{\gamma_1} (1 - \theta)^{\gamma_2 + \gamma_3} \theta^{\gamma_4} \right)^n$$

where $(\gamma_1, \cdots, \gamma_4)$ is a collapsed version of the original Dirichlet weight vector, namely

$$\gamma_j = \frac{1}{n}\sum_{i=1}^{n} w_{n,i} 1[X_i \text{ is in class } j]. \tag{1.3}$$

Furthermore, the point $\tilde{\theta}_n$ maximizing $\tilde{L}_n$ is

$$\tilde{\theta}_n = -\frac{1}{2}(\gamma_2 + \gamma_3 - 2\gamma_1 + 1) + \frac{1}{2}\sqrt{(\gamma_2 + \gamma_3 - 2\gamma_1 + 1)^2 + 8\gamma_4}\,.$$

To perform the WLB, we repeatedly generate scaled Dirichlet vectors w_n and compute $\tilde{\theta}_n$. Simplifying matters, note that by the collapsing property of Dirichlet vectors (see (A.5) of Appendix A)

$$(\gamma_1, \gamma_2, \gamma_3, \gamma_4) \sim \text{Dirichlet}_4(125, 18, 20, 34)\,,$$

and so the simulation involves repeatedly generating vectors of length 4 instead of length $n = 197$. Figure 1.1 compares a histogram based on 5000 draws of $\tilde{\theta}$'s to the likelihood function for θ. The approximation is quite good.

Of course, many computational methods can be brought to bear on this relatively simple model. Fortunately, this simplicity allows us to gain insight into the workings of the WLB procedure, designed primarily for more complex models. We provide some intuition for the WLB in Section 1.4 where the linkage example is studied more closely in the context of models for discrete data.

1.3 Asymptotic justification

Often, the maximizer of the weighted likelihood can be computed by solving the weighted likelihood equations

$$\frac{\partial \log \tilde{L}_n(\theta)}{\partial \theta_k} = 0 \quad k = 1, 2, \cdots, K,$$

where K is the dimension of the parameter space. Three asymptotic results justifying the WLB in such cases have been proven. We restrict attention to models for independent and identically distributed data, and we require certain Cramér-like smoothness conditions to hold. Proofs and a detailed account of sufficient regularity conditions are provided in Chapter 3. Throughout this section, we assume that the

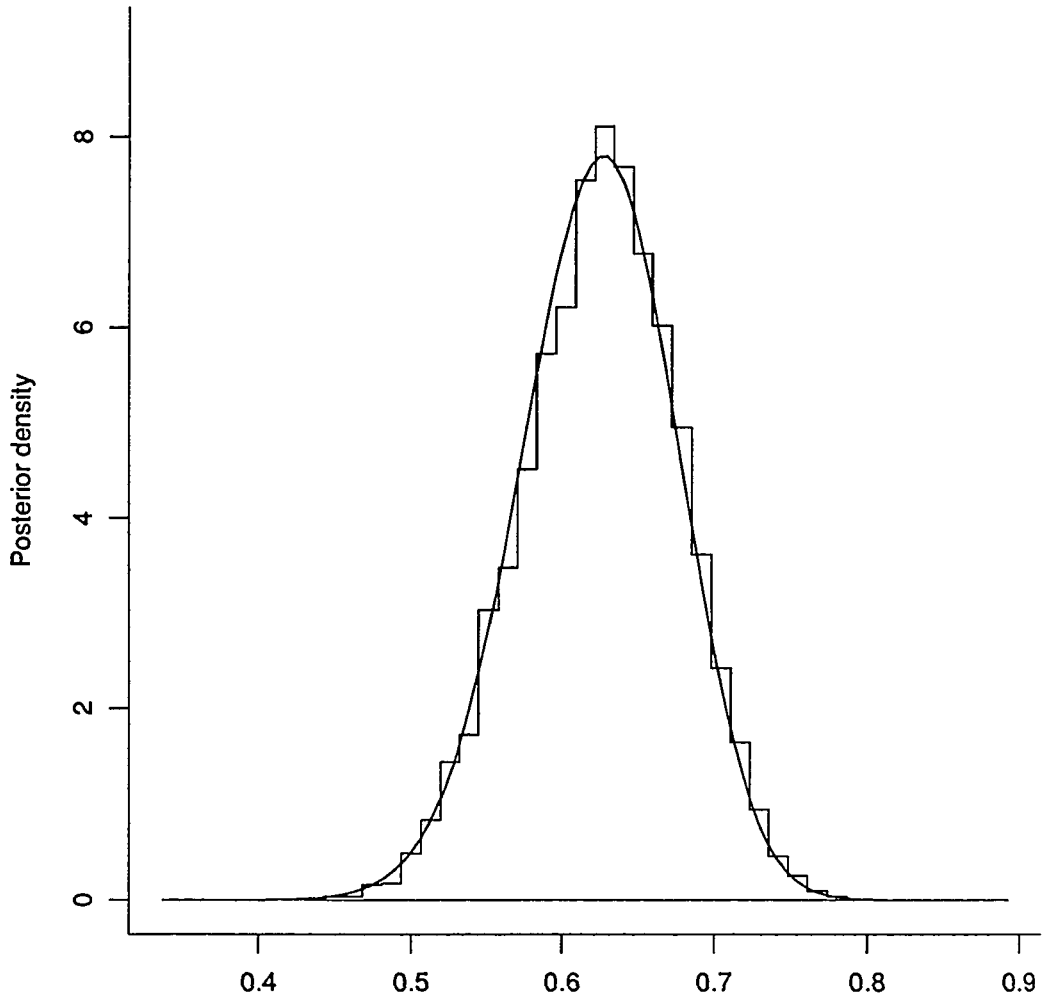


Figure 1.1: The solid curve is the likelihood function for θ from the linkage example described in Section 1.2. The histogram summarizes 5000 draws from a WLB simulation.

model is correctly specified; i.e. data are sampled from some f_{θ_0} and that the weights have a uniform Dirichlet distribution (all $\alpha_i = 1$). Letting $\hat{\theta}_n$ denote the maximum likelihood estimator, we have the following results.

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

$$P\left(|\tilde{\theta}_n - \hat{\theta}_n| > \epsilon \,|\, X_1^n\right) \to 0$$

along almost every sample path.

Theorem 2 *As $n \to \infty$, and for every Borel set $A \subset \Theta \subset \mathbb{R}^K$,*

$$P\left(\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \in A \,|\, X_1^n\right) \to P(\, Z \in A\,)$$

along almost every sample path. Here, Z is a normal random vector with mean 0 and covariance matrix equal to $I(\theta_0)^{-1}$, the inverse Fisher information.

These two results describe the first order behavior of the conditional distribution of $\tilde{\theta}_n$. Of course it is standard theory (e.g. Hartigan 1983) that under somewhat different regularity conditions, the posterior probability

$$P\left(\sqrt{n}(\theta - \hat{\theta}_n) \in A \,|\, X_1^n\right)$$

converges to the same limit as in Theorem 2 (usually in probability along sample paths). Therefore in relatively smooth models, the simulated distribution of $\tilde{\theta}_n$ is the same – at least up to first order – as the posterior distribution of interest. (The reader will also notice that, unconditionally, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to the same normal limit.)

While the first order result is encouraging, it is, in a sense, also minimal. If it were not satisfied, then a simple normal approximation would be better. (The WLB, however, does not require that the Fisher information be known.) Although one might determine the actual order of the approximation by studying Edgeworth expansions, as in Hall (1988) or Weng (1989). we do not attempt this here. We do, however, provide one theoretical result suggesting that the WLB approximation is better than first order. If θ is a one dimensional parameter and if certain smoothness conditions hold on the model then:

Theorem 3 *Along almost every sample path, as* $n \to \infty$,

$$n^2 E\left((\tilde{\theta}_n - \hat{\theta}_n)^3 \mid X_1^n \right) \to h(\theta_0)/\left(I(\theta_0)\right)^3$$

where

$$h(\theta) = 2E\left(\frac{\partial \log f_\theta(X)}{\partial \theta}\right)^3.$$

The relevance of this result becomes apparent when we consider the skewness of a posterior distribution for θ. From results in Hartigan (1983) for example, it can be shown that

$$n^2 E((\theta - \hat{\theta}_n)^3 \mid X_1^n) \to g(\theta_0)/(I(\theta_0))^3,$$

where

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

Indeed, one can characterize the models for which $g(\theta) = h(\theta)$. This class includes the exponential families in which θ is the mean value parameter (see Theorem A.8 of Appendix A).

The asymptotic results stated here and proved in Chapter 3 are restricted to models for independent and identically distributed random variables. Certainly one of the attractions of the WLB, however, is that it can be applied equally well to models for dependent or nonidentically distributed data. We expect that a proof of first order correctness of the WLB for certain dependent models can be derived. In fact, such a proof will undoubtedly use asymptotic theory on the maximum likelihood estimator for such models (see Hall and Heyde, 1984). (Theorem 2 is proved using methods similar to those used traditionally to prove asymptotic normality of the maximum likelihood estimator as in Serfling, 1980, for example.)

1.4 Discrete data models

For the first application of the WLB, we consider the important special class of discrete data models. A critical observation is that the WLB allows exact Bayesian

inference for *unconstrained* multinomial and Markov chain models, in the sense that a data-independent effective prior can be identified. We show how to flatten this effective prior by modifying the distribution of the Dirichlet weights. For general discrete data models, the WLB amounts to drawing probability vectors γ from the *big* model and projecting them onto points in the smaller model of interest by maximizing the weighted likelihood. The actual posterior distribution on probability vectors would be sampled by rejecting all but those probability vectors γ that land in (or close to) the small model. Tanner and Wong (1987) call such a rejection procedure the Dirichlet sampling process. By contrast, the WLB involves no rejection and is thus very efficient. Every sampled probability vector is associated with a point in the model. The objective is to sample raw probability vectors in such a way as the distribution induced on the small model is a posterior under an effective prior that can be identified. To clarify these points, we examine in some detail a simple trinomial model, a logistic regression model and a first-passage-time distribution of a Markov chain.

1.4.1 Unconstrained multinomials

Consider a random sample $X_1, \cdots, X_n$ of single independent multinomial observations on k classes with probability vector $\theta = (\theta_1, \cdots, \theta_k)$. Notationally,

$$X_i \sim_{iid} \text{Mult}_k(1, \theta)\,.$$

A full, or unconstrained model for X_i is the simplex $\mathcal{S}_k$ of all possible probability vectors in $\mathbb{R}^k$. Upon observing the $n-$sample, the likelihood and weighted likelihood functions are

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

Here, y_j is the number of X_i landing in the j^{th} class. Similarly, the vector $\gamma = (\gamma_1, \cdots, \gamma_k)$ is formed from the original weights $w_{n,i}$ (see equation (1.2)) by adding the $w_{n,i}$ which fall in the same class:

$$\gamma_j = \frac{1}{n} \sum_{i=1}^{k} w_{n,i} 1[X_i \text{ in class } j]\,.$$

It is a well known property of Dirichlet random vectors that γ has a Dirichlet distribution with parameters $\beta = (\beta_1, \cdots, \beta_k)$ where

$$\beta_j = \sum_{i=1}^{k} \alpha_i 1[X_i \text{ in class } j] .$$

Therefore its probability density (or more precisely the density of the first $k-1$ components) is proportional to

$$\prod_{j=1}^{k} \gamma_j^{\beta_j - 1} 1[\beta_j > 0]$$

for $\gamma \in \mathcal{S}_k$. For this *unconstrained* model (θ can be anywhere in $\mathcal{S}_k$) the parameter value $\tilde{\theta}$ which maximizes $\tilde{L}$ is simply $\tilde{\theta} = \gamma$. This means that the distribution of simulated vectors $\tilde{\theta}$ is the same as the posterior distribution of θ under the prior

$$\pi_0(\theta) \propto \prod_{j=1}^{k} \theta_j^{\beta_j - y_j - 1}$$

where y_j is the number of X_i landing in class j. Thus the effective prior is conjugate. Note that in the canonical case where $\alpha_i = 1$, the effective prior is improper and is proportional to the square of Jeffreys' prior:

$$\pi_0(\theta) \propto \prod_{j=1}^{k} \theta_j^{-1} . \tag{1.4}$$

This simulation is the same as Rubin's (1981) Bayesian bootstrap described in Section 1.8.2.

In this special case where the conditional distribution of $\tilde{\theta}$ is known exactly, we can study what happens when the Dirichlet weights $w_{n,i}$ are slightly modified. Let $j(i)$ be the class into which the observation X_i falls. If the weights are distributed

$$(w_{n,1}, w_{n,2}, \cdots, w_{n,n}) \sim \text{Dirichlet}_n \left(1 + 1/y_{j(1)}, 1 + 1/y_{j(2)}, \cdots, 1 + 1/y_{j(n)}\right)$$

then simulated $\tilde{\theta}$'s have the same distribution, given the data, as θ does under a *uniform* prior. This simple example shows how modifying the distribution of the weights is like modifying the prior distribution. Of course the effect of this modification, like the effect of any reasonable prior in a parametric model, becomes negligible as the sample size n increases.

1.4.2 Constrained multinomials

Although all multinomial probability vectors are constrained to sum to one, it is often the case that models of interest put further constraints on these probabilities. Let us reconsider the first example from Section 1.2. In this so called linkage example, four counts are observed from n independent multinomial random variables. Table 1.1 shows four sets of possible counts; these data sets were also analyzed by Tanner and Wong (1987).

Table 1.1: Four examples of linkage data; one per row

y_1	y_2	y_3	y_4	$n = \sum y_i$
125	18	20	34	197
13	2	2	3	20
14	0	1	5	20
3	2	2	3	10

Theory on the recombination of genes restricts the probability vector p to a one dimensional subset of $\mathcal{S}_4$

$$\mathcal{P}_\Theta = \{p \in \mathcal{S}_4 : \; p = \frac{1}{4}(2+\theta, 1-\theta, 1-\theta, \theta), \;\; \theta \in [0,1]\}\,,$$

where θ^2 is a *recombination fraction* determining the degree of linkage between two factors. Here, we think of the data as trinomial with counts $(y_1, y_2 + y_3, y_4)$ and probability vector

$$p = \frac{1}{4}\,(2+\theta, 2(1-\theta), \theta) \in \mathcal{P}_\Theta \subset \mathcal{S}_3\,.$$

To estimate the posterior distribution of θ, we use the WLB. The weighted likelihood function is

$$\tilde{L}(\theta) \propto (2+\theta)^{n\gamma_1}(1-\theta)^{n\gamma_2}\theta^{n\gamma_3}\,.$$

The vector γ is the same as in the last section (with $k = 3$), having density proportional to

$$\gamma_1^{y_1-1}\gamma_2^{y_2+y_3-1}\gamma_3^{y_4-1} \tag{1.5}$$

when all $\alpha_i = 1$. Figure 1.2 shows contour lines of the probability density of γ on $\mathcal{S}_3$ for each of the data sets given in Table 1.1. The line defining $\mathcal{P}_\Theta$ is also drawn.

The parameter value $\tilde{\theta}$ which maximizes $\tilde{L}$ can be computed in several ways. We can either modify the EM algorithm (see Section 1.5.2), or we can note the exact solution

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

Now the WLB simulation proceeds as follows. We repeatedly generate vectors γ according to a Dirichlet distribution (equation (1.5)) and then we compute $\tilde{\theta}$. In effect, when computing $\tilde{\theta}$, we are finding the point p in the model $\mathcal{P}_\Theta$ which is closest to the point γ according to the measure

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

This is the Kullback-Leibler information number relating p and γ. Figure 1.3 shows how this projection from $\mathcal{S}_3$ into $\mathcal{P}_\Theta$ happens.

In Figure 1.4, a histogram from 5000 simulated $\tilde{\theta}$'s is compared with two posterior distributions for each of the data sets in Table 1.1. One of the posteriors used in the comparison is simply the likelihood function, which is equal to the posterior under a uniform prior. The other posterior comes from a special prior

$$\pi_0(\theta) \propto \{(2+\theta)(1-\theta)\theta\}^{-1}\,.$$

Note that this prior is equal to the restriction of equation (1.4) to probability vectors $p \in \mathcal{S}_3$ which respect the model (i.e. $p = p(\theta)$). The approximation is reasonably good and improves with increasing sample size. We have been able to accurately guess the effective prior here by considering the restriction of the prior in the full model. One reason for success, is that the model is *linear*. In next section we study a logistic regression model which is *curved* relative to the full model of probabilities.

The case shown in Figure 1.4.3 is somewhat difficult 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 provides a close approximation to the posterior under the prior $\pi_0(\theta)$, and is somewhat different from the likelihood function.

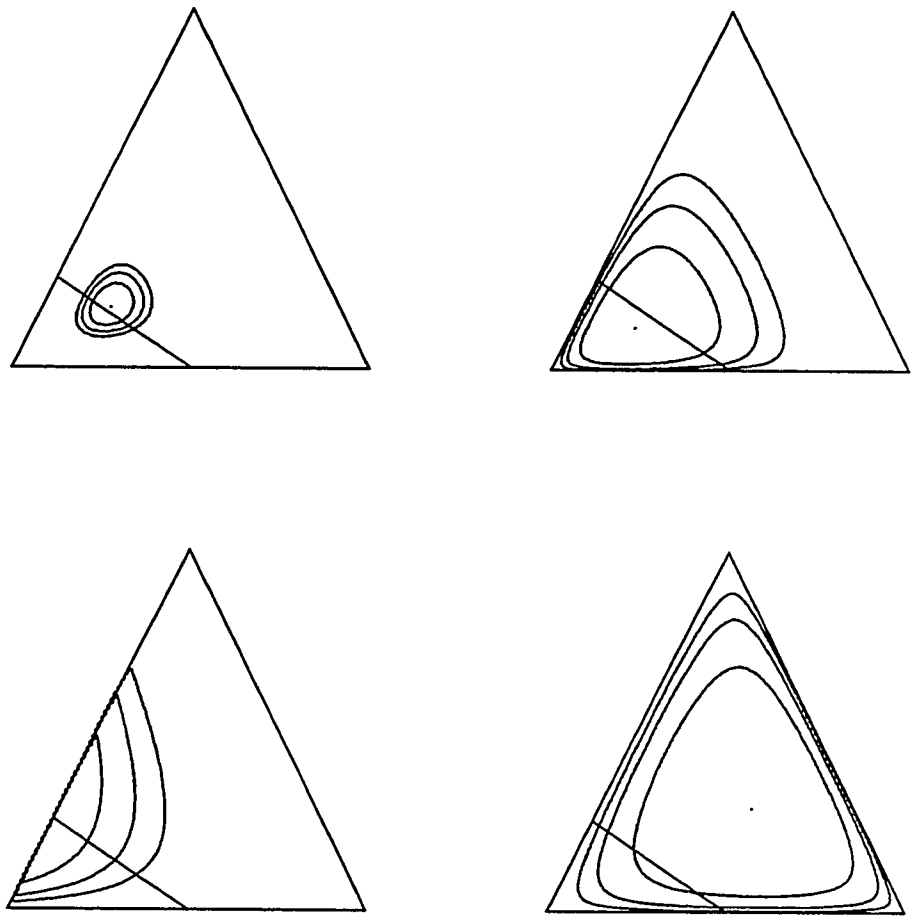


Figure 1.2: This plot shows the posterior Dirichlet densities on the full model in each of the 4 data sets from table (1.1). Trinomial probability vectors which satisfy the model constraints lie on the line segment of negative slope. The contour lines are at $0, -1, -2$, and -3 units of log density.

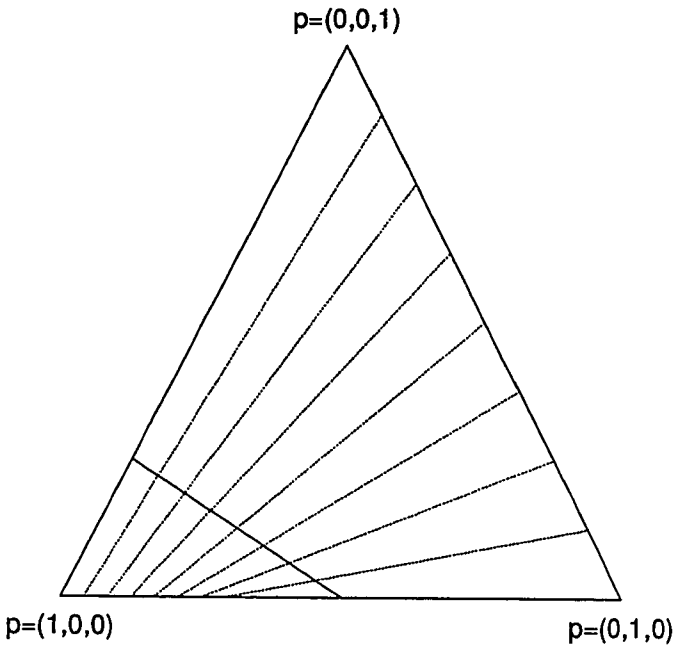


Figure 1.3: The simplex $\mathcal{S}_3$ shown here represents all possible trinomial probability vectors. The line of negative slope is the linkage model; the set of probability vectors satisfying certain constraints. In a WLB simulation, points are sampled from $\mathcal{S}_3$ and then projected down into the model. The dashed lines show how this projection happens. All the points on the same dashed line are projected onto the same point in the model.

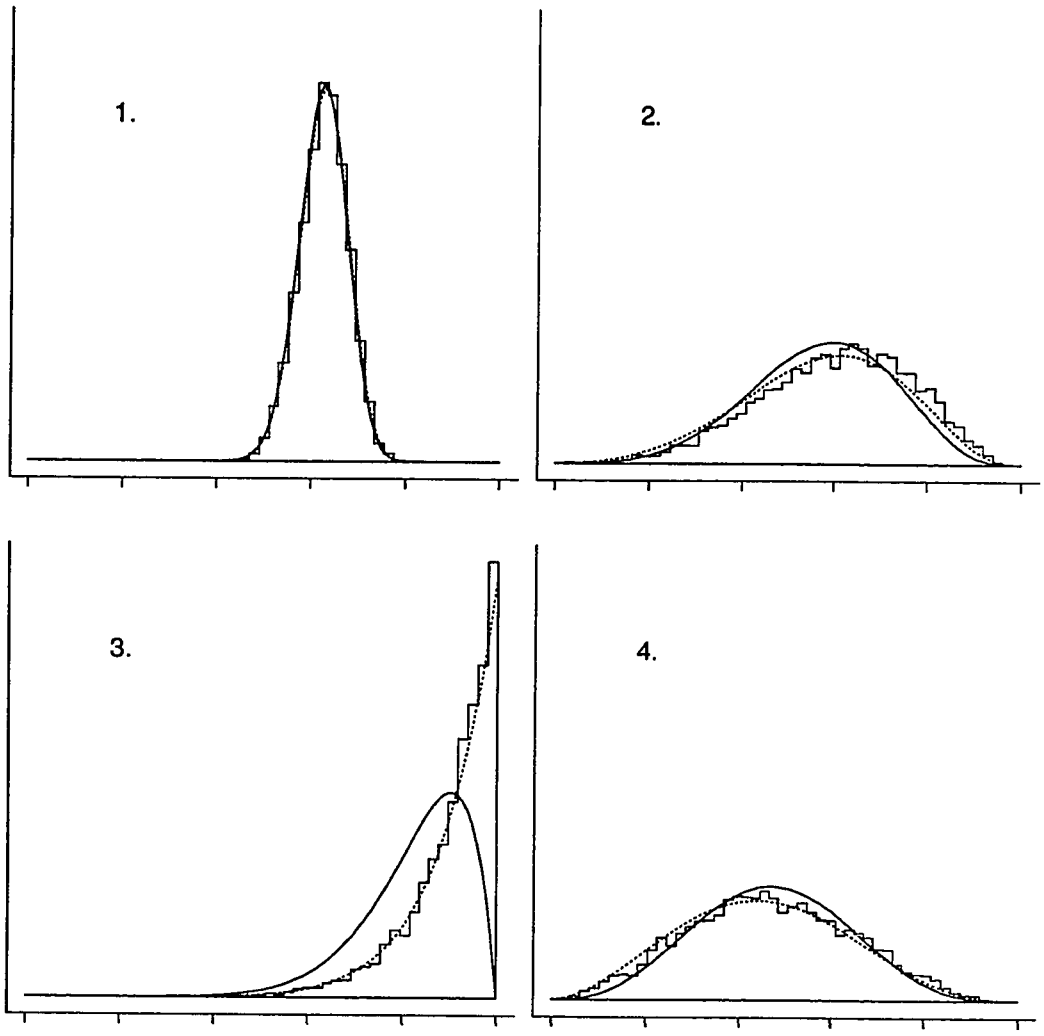


Figure 1.4: A histogram from the WLB simulation is compared with two posterior distributions for each of the four data sets in Table 1.1. Solid curves show the likelihood functions while dashed curves show the posteriors under the prior $p(\theta) \propto 1/\{\theta(2+\theta)(1-\theta)\}$. The histograms are based on 5000 draws.

The prior in the WLB simulation can be changed by modifying the parameters of the distribution of $w_{n,i}$. As noted earlier, this so called effective prior is not necessarily the prior of choice in any given problem. It is rather the prior which leads to a posterior distribution exactly the same as the distribution of simulated $\tilde{\theta}$'s. That is, the likelihood times the effective prior equals the conditional density of $\tilde{\theta}$. In Figure 1.5, a histogram of 5000 simulated $\tilde{\theta}$'s is compared with the likelihood function as in Figure 1.4. The difference between Figure 1.5 and 1.4 is in the distribution of the weights γ used. In Figure 1.4 we have

$$\gamma = (\gamma_1, \gamma_2, \gamma_3) \sim \text{Dirichlet}_3(y_1, y_2 + y_3, y_4)\,,$$

while in Figure 1.5,

$$\gamma = (\gamma_1, \gamma_2, \gamma_3) \sim \text{Dirichlet}_3(y_1 + 1, y_2 + y_3 + 1, y_4 + 1)\,.$$

With this second choice of weights, the effective prior is effectively uniform.

1.4.3 Logistic regression

As we have shown in several simple examples, the WLB allows approximate simulation from a posterior distribution. Not surprisingly, the effective prior associated with such a posterior distribution depends on how we parameterize the model. A uniform prior in one parameterization is modified by the Jacobian of a nonlinear transformation to become a nonuniform prior. This is the case for the logistic regression model described below. To complicate matters slightly in this example, the transformation maps a two dimensional manifold in $\mathbf{R}^5$ into $\mathbf{R}^2$. Therefore, the standard change-of-variables formula is inapplicable. Using the theory of calculus on manifolds, however, we can calculate a Jacobian for this problem. When modified by this Jacobian, the WLB samples give an estimated posterior which is very close to the posterior under a uniform prior on the parameters of interest.

We consider the following, purely illustrative, example from Pearson and Hartley (1976, p.8). At each of five distances from a target, a charge is detonated and it is recorded whether or not the target has been perforated. This experiment is repeated 16 times. Table 1.2 shows the results. In studying the chance that a target is

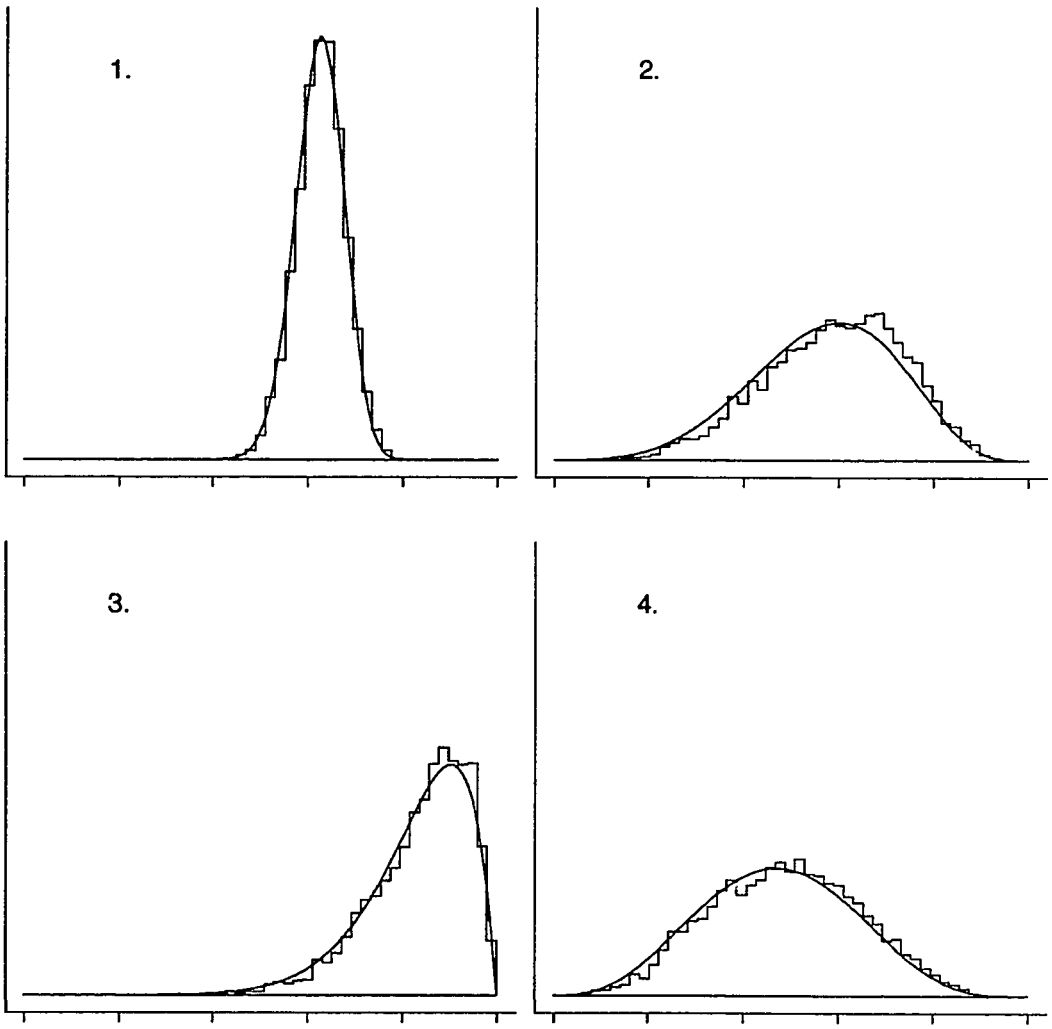


Figure 1.5: This plot is similar to Figure 1.4 except that modified weights are used in the WLB simulation. Modifying the weights makes the resulting histograms closer to the likelihood functions which they are estimating.

perforated, one might consider the standard logistic regression model

$$\log\left(\frac{p_i}{1-p_i}\right) = \theta_0 + \theta_1 d_j$$

where p_j is the chance of a perforation in the target when the charge is detonated at distance d_j. In such a logistic regression model, we may be interested in the posterior distribution of the regression parameters $\theta = (\theta_0, \theta_1)$.

Table 1.2: Data on the perforations caused by detonated charges at different distances

distance of charge	53	49	45	41	37
coded distance d_j	0	1	2	3	4
number of detonations m_j	16	16	16	16	16
number of perforations y_j	0	9	9	12	16

Following the prescription given in Section 1.2, the weighted likelihood function for θ becomes

$$\tilde{L}(\theta) \propto \prod_{j=1}^{5} p_j^{n\gamma_j}(1-p_j)^{n\bar{\gamma}_j}$$

where, when all $\alpha_i = 1$

$$\gamma = (\gamma_1, \bar{\gamma}_1, \gamma_2, \bar{\gamma}_2, \cdots, \bar{\gamma}_5) \sim \text{Dirichlet}_{10}(y_1, m_1 - y_1, y_2, m_2 - y_2, \cdots, m_5 - y_5)\,.$$

To compute $\tilde{\theta}$ which maximizes $\tilde{L}$, we can use a standard algorithm for computing maximum likelihood estimates. Simply treat the $n\gamma_j$ like counts y_j and the $n\bar{\gamma}_j$ like $m_j - y_j$. (This simple treatment of weights like data works in exponential family models, but not in general. We discuss maximization of $\tilde{L}$ in more detail in Section 1.5.) The upper panel in Figure 1.6 shows a kernel density estimate based on 3000 draws in a WLB simulation. A Gaussian kernel density estimate is used by appealing to the principle of maximal smoothing (Terrel 1990). The covariance matrix of the kernel is thus

$$\left(\frac{5^4}{2^7 3m}\right)^{1/3}$$

times the covariance matrix of the m WLB samples.

If we want to approximate the posterior density for θ under a uniform prior, then we must consider the fact that θ is a nonlinear transformation of probabilities. Even if we can simulate the posterior density of the probabilities under a uniform prior, the induced distribution for θ is no longer proportional to the likelihood function. We try to account for this nonlinear transformation by modifying our kernel density estimate. Considering the earlier results, we might suspect that our WLB simulation is drawing from a uniform prior on the probability scale. For each set of weights γ there is a vector $p_\gamma = (p_1, \cdots, p_5)$ which *respects* the model and for which the weighted likelihood is largest, i.e. $p_\gamma = p(\tilde{\theta})$. These probabilities live on a two-dimensional manifold in $\mathbb{R}^5$. In running our simulation, we are sampling from some density on this manifold. Here we must be careful, because the density is with respect to the natural measure on the manifold (Hausdorff measure) rather than Lebesgue measure, but it is being sampled nonetheless (see Billingsley, 1986, Section 19, especially Theorem 19.3). Now there is a one-to-one transformation between this manifold and $\mathbb{R}^2$ where θ lives. If this sampling density on the manifold is proportional to the likelihood, then the induced density on $\mathbb{R}^2$ is proportional to the likelihood times the Jacobian

$$J(\theta) = \left(|u|^2|v|^2 - |uv|^2\right)^{1/2} .$$

Here u and v are vectors of length 5 representing the derivatives of the map from the manifold into $\mathbb{R}^2$:

$$u_j(\theta) = p_j(\theta)(1 - p_j(\theta))$$
$$v_j(\theta) = d_j u_j .$$

The lower panel of Figure 1.6 shows the contour lines of $1/J(\theta)$. To get a posterior for θ under a uniform prior, we divide our kernel density estimate by $J(\theta)$. Figure 1.7 shows two summaries of this modified joint distribution. The upper panel compares estimates of the marginal posterior density for the intercept parameter θ_0 while the lower panel does the same for the slope parameter θ_1. The dashed lines result from modifying the kernel density estimate of the WLB. The solid lines show the gold standard: the Gibbs sampler solution. The WLB gives a very good approximation in this example. In computing the Gibbs sampler solutions, we ran a single Markov chain over the parameter space. There is a burn in of 10 steps, and we ran the

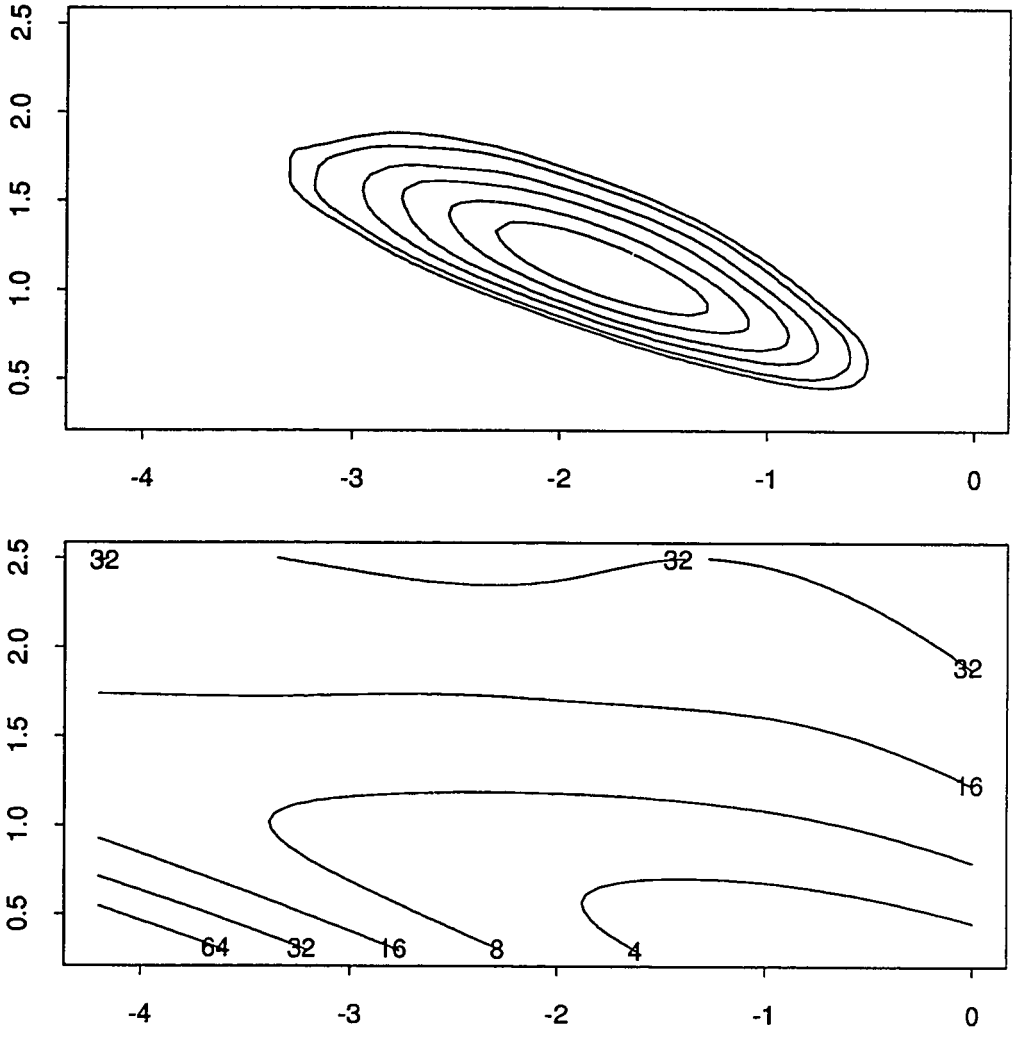


Figure 1.6: The top panel shows a contour plot of a kernel density estimate from 3000 WLB samples. This is our initial estimate of the joint posterior density of two logistic regression parameters (Section 1.4.3). Contour lines determine probability regions of levels 50, 75, 90, 95, 99 and 99.5 percent. The lower panel shows contour lines of $1/J(\theta)$ where $J(\theta)$ is the Jacobian of the transformation from probability space into the sample space.

chain for a further 5,000 iterations, sampling every 10th value. The marginal densities were determined by averaging conditional densities as described in Gelfand and Smith (1990). Although no simple form is available for the conditional distributions involved, we use a discrete approximation at each step in the chain.

It is interesting to compare these results with Efron's bootstrap. Consider a logistic regression model in which we view the Bernoulli observations and their corresponding covariates as being randomly sampled from some distribution (as opposed to having fixed covariates). The bootstrap distribution of the MLE has infinite variance for such models (Breslow, 1987). To see why, note that in the space of covariates, we can find a hyperplane which splits the data into two parts in such a way that some 0's are on one side and some 1's are on the other side. By sampling with replacement from the data, we can imagine a bootstrap sample for which we observe only 0's on one side of this hyperplane and only 1's on the other side. For such a bootstrap sample, there is no θ on the interior of the parameter space which can maximize the likelihood (in essence, the MLE must give probability 0 to the points on the side of the hyperplane where all 0's have been observed). As the chance of such a split goes to zero very quickly, the asymptotic validity of the bootstrap is ensured (see Lee, 1990). Unfortunately, for any fixed data set, the practitioner will inevitably run into points at infinity in his or her bootstrap simulation. This is not the case for the WLB, essentially because points never get hard 0 weights.

1.4.4 Finite state Markov chains

In this section, we see how the WLB works when there is simple dependence structure in the data. Specifically, we consider a $K-$state Markov chain $\{X_t : t = 0, 1, \cdots, n\}$ which starts at some known state x_0. Table 1.3 summarizes a single realization of a five-state chain of length 100. Letting $\theta = (\theta_{j,k})$ be the transition matrix, the weighted likelihood function for θ reduces from a product over the 100 time points to a product over the 25 states

$$\tilde{L}(\theta) = \prod_{j=1}^{5} \prod_{k=1}^{5} \theta_{j,k}^{n\gamma_{j,k}}.$$

As for the multinomial model, the weights $\gamma_{j,k}$ are derived from the initial weights $w_{n,i}$ by aggregation over common transitions. With the minimal constraints on the

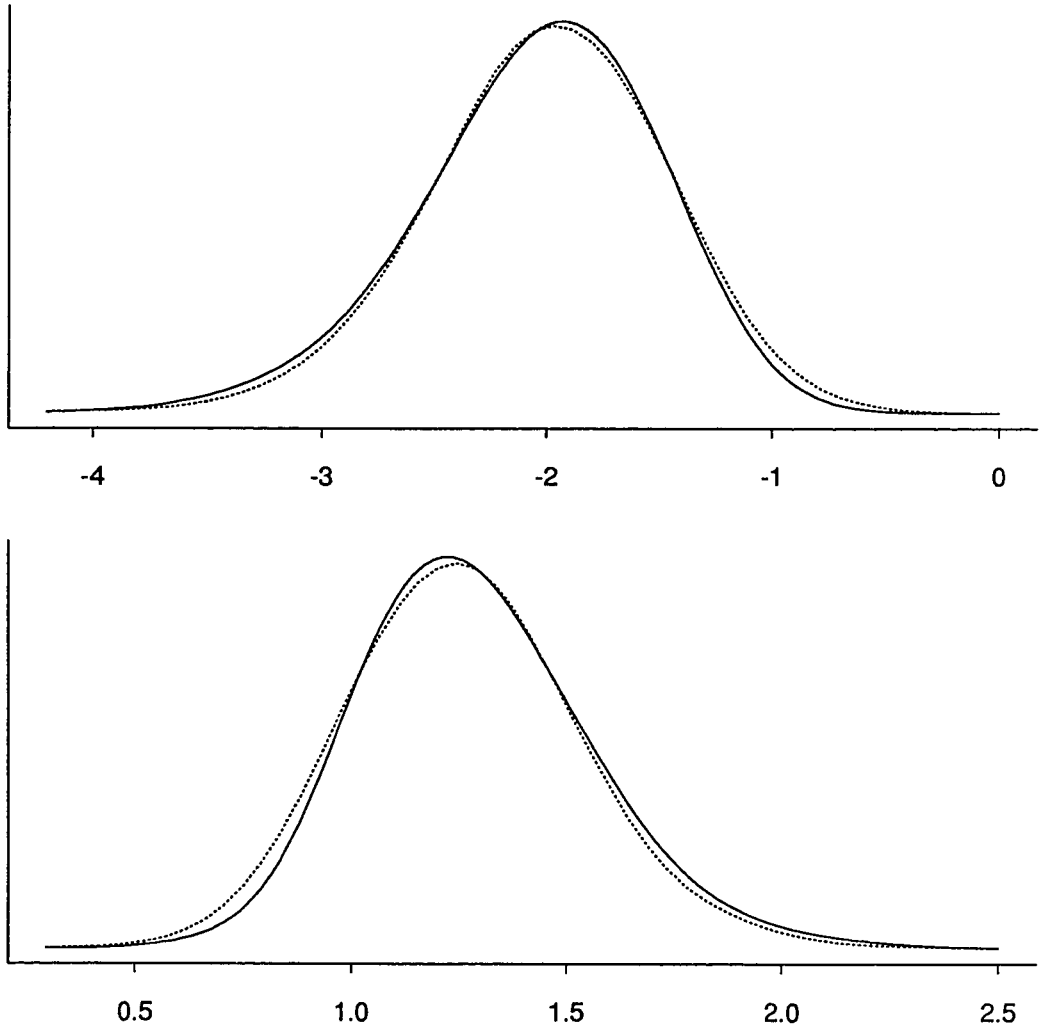


Figure 1.7: Estimates of the marginal posterior density for each parameter in the logistic regression model of Section 1.4.3 are shown above. Solid lines result from Gibbs sampling, while dashed lines result from the WLB. To calculate these dashed lines, the bivariate density estimate (upper plot in Figure 1.6) is multiplied by $1/J(\theta)$ (lower plot in Figure 1.6) and then the result is marginalized.

transition probabilities, the matrix $\tilde{\theta}$ maximizing $\tilde{L}$ has entries

$$\tilde{\theta}_{j,k} = \frac{\gamma_{j,k}}{\sum_k \gamma_{j,k}}.$$

In fact, by various properties of the Dirichlet distribution (see (A.4) and (A.5) of Appendix A), the density of $\tilde{\theta}$ (or strictly the density of the first 4 columns) is proportional to the likelihood function times the prior

$$\prod_{j=1}^{5}\prod_{k=1}^{5}\theta_{j,k}^{-1}$$

when all $\alpha_i = 1$. By slightly modifying the α_i like in Section 1.4.1, we can change the distribution of $\gamma_{j,k}$ to that the effective prior is flat on the space of transition matrices.

Table 1.3: Transitions of a length 101 realization of a 5 state Markov chain

		State at time t					total
		1	2	3	4	5	
state	1	3	5	6	6	3	23
at	2	15	1	2	4	3	25
time	3	0	10	1	4	3	18
$t-1$	4	2	3	10	2	3	20
	5	2	6	0	4	2	14

Simulation of the posterior is important if we want to do inference on some complicated function of the transition matrix. Consider, for example, Bayesian inference about a first-passage-time distribution. Such a distribution, say for passage from state 1 to state 5, is based on probabilities of the form

$$\eta_t = P(X_t = 5,\, X_{t-1} \neq 5, \cdots, X_2 \neq 5 \,|\, X_1 = 1) \tag{1.6}$$

for $t = 2, 3, \cdots$. Clearly, $\eta_2 = \theta_{1,5}$. It is well known that recursive updating gives an expression for η_t for $t > 2$. To see this, construct a sequence of vectors u_t in $\mathbb{R}^5$

defined by

$$u_t(j) = \begin{cases} \theta_{1,j} & \text{if } t = 2 \\ \sum_{k=1}^{4} \theta_{k,j} u_{t-1}(k) & \text{if } t > 2. \end{cases} \tag{1.7}$$

As $\eta_t = u_t(5)$, simulation of transition matrices θ leads to simulated probabilities η_t through the recursive relation above.

The upper plot of Figure 1.8 shows the maximum likelihood estimate of the $1-5$-first passage time distribution based on data in Table 1.3. This estimate is computed by first computing the MLE of the transition matrix and then using that in equations (1.6) and (1.7).

The results of a WLB simulation of the first passage time distribution are shown in the lower panel of Figure 1.8. Based on the data in Table 1.3, we simulated transition matrices $\tilde{\theta}$ from their posterior under a uniform prior. For each of these 500 simulated transition matrices, we determined the median and upper quartile of the first passage time distribution using equations (1.6) and (1.7). The resulting marginal posterior mass functions of these parameters are shown in the lower panel of Figure 1.8.

1.5 Maximizing the weighted likelihood function

Standard methods for computing maximum likelihood estimates (MLE's) can often be used to maximize a weighted likelihood function. The upshot of this in practice is that computer code for calculating MLE's can be invoked, unchanged, to perform the WLB simulation. Two such methods are iteratively reweighted least squares (IRLS) (Green, 1984) and the EM algorithm (Dempster *et al.*, 1977). Note that explicit calculation of the likelihood function itself is not required, so that in problems solved by IRLS and the EM algorithm we can use the WLB to approximate the posterior distribution even if we are unable to evaluate the likelihood function.

1.5.1 Iteratively reweighted least squares

Consider, as in Section 1.3, a weighted likelihood function $\tilde{L}$ (or its logarithm $\tilde{l}$) which is maximized by solving the (vector) *weighted likelihood equation*

$$\frac{\partial \tilde{l}}{\partial \theta}(\theta) = 0 \tag{1.8}$$

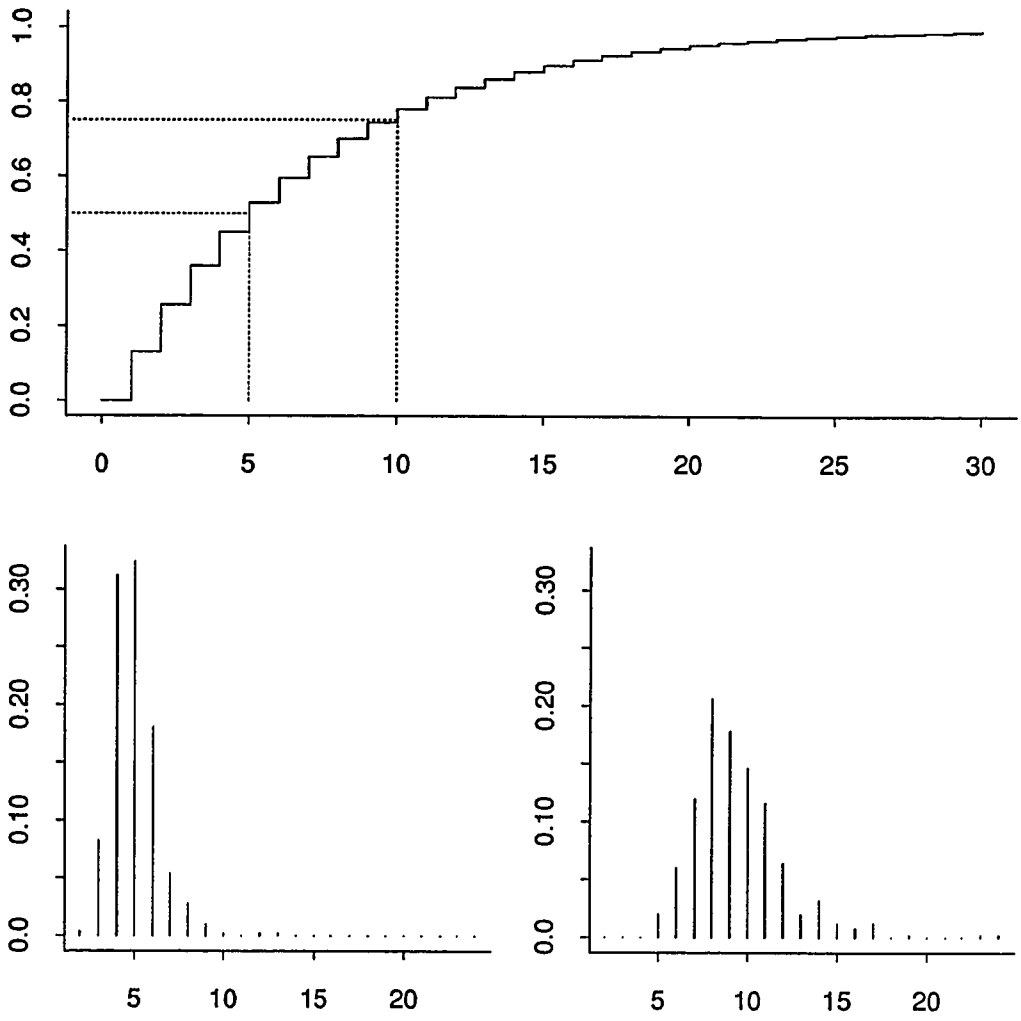


Figure 1.8: The upper graph shows an estimate of a first passage time distribution for the observed Markov chain summarized in Table 1.3. The distribution is for passage from state 1 to state 5 and is based on the maximum likelihood estimate of the transition probabilities. The median and upper quartile of this distribution are indicated with dashed lines. The lower plots are marginal posterior mass functions of the median and upper quartile of this first passage time distribution. They are based on 500 WLB samples.

for $\tilde{\theta} \in \mathbf{R}^p$. There is a close connection between a solution of equation (1.8) and the IRLS solution to the corresponding likelihood equation

$$\frac{\partial l}{\partial \theta}(\theta) = 0\,. \tag{1.9}$$

Here l is the logarithm of the likelihood function and equation (1.9) is solved by the MLE $\hat{\theta} \in \mathbf{R}^p$.

In the general formulation described in Green (1984), the loglikelihood l is viewed as a function of an $n-$vector of predictors $\eta = (\eta_1, \eta_2, \cdots, \eta_n)^T$. 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 $n \times p$ matrix $(\partial \eta/\partial \theta)$, the likelihood equation (1.9) becomes simply

$$D^T u = 0\,. \tag{1.10}$$

Now to study the IRLS solution of the weighted problem (1.8), suppose that the densities from equation (1.1) have the form

$$f_{\theta,i}(x_i | x_1^{i-1}) = \psi_i(\eta_i, x_i) \tag{1.11}$$

where for each i, ψ_i is a fixed, known function determined by the model. The predictor η_i, as well as being functionally dependent on the parameter θ may depend on fixed covariates or the *past*, x_1^{i-1}, but not on x_i itself. Of course, a host of models satisfy equation (1.11) including generalized linear models (McCullagh and Nelder, 1989) and the autoregressive time series models discussed in Section 2.5.

In contrast to equation (1.10), the weighted likelihood equation (1.8) can be written as follows for models described by equation (1.11):

$$\frac{\partial \tilde{l}}{\partial \theta} = D^T W u = 0\,. \tag{1.12}$$

Here W is an $n \times n$ diagonal matrix whose diagonal entries are the scaled Dirichlet weights $w_{n,i}$ defining the weighted likelihood function, (see equation (1.2)).

Following Green (1984), the iterative Newton-Raphson solution to the ordinary likelihood equation (1.10) is to first evaluate u, D, and the second derivatives of l

at an initial guess $\hat{\theta}_0$. Then an updated guess $\hat{\theta}_1$ is computed by solving the linear system

$$-\frac{\partial^2 l}{\partial\theta\theta^T}(\hat{\theta}_1 - \hat{\theta}_0) = D^T u\,. \tag{1.13}$$

Iteration continues until convergence. In the standard Fisher scoring or IRLS solution, on the other hand, the matrix $(\partial^2 l/\partial\theta\theta^T)$ in (1.13) is replaced by an approximation $D^T A D$ where A is the expectation (under the current parameter value) of the $n \times n$ matrix $(\partial^2 l/\partial\eta\eta^T)$. This approximation is derived from the expansion

$$\frac{\partial^2 l}{\partial\theta\theta^T} = D^T \frac{\partial^2 l}{\partial\eta\eta^T} D + \sum_{i=1}^{n} \frac{\partial l}{\partial\eta_i}\frac{\partial^2\eta_i}{\partial\theta\theta^T}\,, \tag{1.14}$$

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 $\hat{\theta}_0$ and then solving the linear system

$$D^T A D(\hat{\theta}_1 - \hat{\theta}_0) = D^T u \tag{1.15}$$

for $\hat{\theta}_1$. Again, iteration continues until convergence. We must assume that D is of full rank p and A is positive definite to ensure a unique solution at each iteration. By noting that equation (1.15) defines the normal equations for a regression problem, we can compute $\hat{\theta}_1$ by regressing $A^{-1}u + D\hat{\theta}_0$ on D with weight matrix A. That is

$$\hat{\theta}_1 = (D^T A D)^{-1} D^T A(A^{-1}u + D\hat{\theta}_0)\,. \tag{1.16}$$

Hence the name IRLS.

Apparently, then, a solution $\tilde{\theta}$ to the weighted likelihood equation (1.8) is in hand if we can derive some approximation to the matrix $(\partial^2\tilde{l}/\partial\theta\theta^T)$. By analogy with (1.14), we have

$$\frac{\partial^2\tilde{l}}{\partial\theta\theta^T} = D^T W \frac{\partial^2 l}{\partial\eta\eta^T} D + \sum_{i=1}^{n} w_{n,i}\frac{\partial l}{\partial\eta_i}\frac{\partial^2\eta_i}{\partial\theta\theta^T}\,. \tag{1.17}$$

Now applying the same approximation as in Fisher scoring, we get the following iterative algorithm to compute $\tilde{\theta}$. Evaluate u, D, and A at an initial guess $\tilde{\theta}_0$, and then solve

$$D^T W A D(\tilde{\theta}_1 - \tilde{\theta}_0) = D^T W u\,. \tag{1.18}$$

As an IRLS algorithm, we compute $\tilde{\theta}_1$ by regressing $A^{-1}u + D\tilde{\theta}_0$ on D with weight matrix WA. The two algorithms – one for solving the likelihood equation, and one for solving the weighted likelihood equation – use the same components u, D, and A. The only difference is that the weight matrix in the IRLS algorithm to compute $\tilde{\theta}$ is WA instead of A.

The WLB simulation involves repeatedly generating weight matrices W and then performing the IRLS algorithm described above. From the form of the estimating equations, we get a convenient simplification of this simulation. The weight matrix W can have a random sample of Gamma's $Y_1, \cdots, Y_n$ (of Section 1.2) on its diagonal instead of scaled Dirichlets.

1.5.2 EM algorithm

For some models, including exponential families, the weighted likelihood function is proportional to a likelihood function given by some modified sufficient statistics. If the EM algorithm applies to the original problem, then it can be used in an obvious way to maximize the weighted likelihood.

In general, the weighted likelihood function is not proportional to a likelihood function for any modified data set. Nonetheless, the essential trick underlying the EM algorithm can be used to maximize the weighted likelihood function. To see how this works, recall that the EM algorithm is an iterative procedure for finding local maxima of $l(\theta)$, the log likelihood. The algorithm is based on a partition of $l(\theta)$ as

$$l(\theta) = Q(\theta, \theta^{(p)}) - H(\theta, \theta^{(p)}) \tag{1.19}$$

where $\theta^{(p)}$ is any point in Θ considered to be the p^{th} iterate of the algorithm. The function $Q(\theta, \theta^{(p)})$ is maximized by $\theta^{(p+1)}$ for fixed $\theta^{(p)}$, while the function H satisfies

$$H(\theta, \theta^{(p)}) \leq H(\theta^{(p)}, \theta^{(p)}). \tag{1.20}$$

It follows from equations (1.19) and (1.20) that the sequence of iterates $\theta^{(1)}, \theta^{(2)}, \cdots$ has non-decreasing likelihood. The partition in (1.19) is introduced because $Q(\theta, \theta^{(p)})$ is often much easier to maximize than $l(\theta)$. In the standard EM algorithm, Q is taken to be a conditional expectation of a *complete data* log likelihood. Calculation of Q is the standard E step, while maximization of Q is the M step.

In models where the EM algorithm is used, we can often partition the log weighted likelihood function

$$\tilde{l}(\theta) = \tilde{Q}(\theta, \theta^{(p)}) - \tilde{H}(\theta, \theta^{(p)}) \tag{1.21}$$

by analogy to (1.19). The main difference is that $\tilde{Q}$ may not be a conditional expectation in the way that Q is. The algorithm nonetheless finds local maxima of $\tilde{l}$. This is because the function $\tilde{H}$ is constructed to satisfy the crucial relationship

$$\tilde{H}(\theta, \theta^{(p)}) \leq \tilde{H}(\theta^{(p)}, \theta^{(p)}) . \tag{1.22}$$

As an example, we construct an EM algorithm for maximizing the weighted likelihood function in a finite mixture model. In such a model, the data $X_1, \cdots, X_n$ have a density

$$f_\theta(x) = \sum_{j=1}^{K} \pi_j f_j(x) \tag{1.23}$$

where $\pi = (\pi_1, \cdots, \pi_K)^T$ is a probability vector and $f_1, \cdots, f_k$ are densities on the sample space. The parameter θ determines π and all the f_j. The likelihood from a sample of size n is

$$L(\theta) = \prod_{i=1}^{n} \left(\sum_{j=1}^{K} \pi_j f_j(x_i) \right) \tag{1.24}$$

while the weighted likelihood is

$$\tilde{L}(\theta) = \prod_{i=1}^{n} \left(\sum_{j=1}^{K} \pi_j f_j(x_i) \right)^{w_{n,i}} . \tag{1.25}$$

In deriving an algorithm to maximize the likelihood (1.24), we construct a *complete data* likelihood function

$$L_c(\theta) = \prod_{i=1}^{n} \left(\sum_{j=1}^{K} z_{i,j} \pi_j f_j(x_i) \right) \tag{1.26}$$

where

$$z_i = (z_{i,1}, \cdots, z_{i,K}) \sim_{iid} \text{Mult}_K(1, \pi) .$$

The function L_c is the likelihood function in the hypothetical situation where pairs (X_i, Z_i) are observed. The interpretation, of course, is that Z_i indicates which population X_i is being sampled from. Following the prescription of Dempster *et al.* (1977), we construct

$$\begin{aligned} Q(\theta, \theta^{(p)}) &= E\left(\log L_c(\theta) | X_1^n, \theta^{(p)}\right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{K} \tau_{j,i}^{(p)} \log(\pi_j f_j(x_i)) \end{aligned}$$

where

$$\begin{aligned} \tau_{j,i}^{(p)} &= E\left(z_{i,j} | X_1^n, \theta^{(p)}\right) \\ &= \frac{\pi_j^{(p)} f_j^{(p)}(x_i)}{\sum_{k=1}^{K} \pi_k^{(p)} f_k^{(p)}(x_i)}. \end{aligned}$$

The algorithm for maximizing $L(\theta)$ proceeds as follows. At the p^{th} iteration, compute $Q(\theta, \theta^{(p)})$ and then let $\theta^{(p+1)}$ be the value of θ maximizing this function. Continue until convergence. Note that $Q(\theta, \theta^{(p)})$ is much easier to maximize than $\log L(\theta)$ because the logarithm of a sum has been converted into the sum of logarithms.

In deriving an iterative algorithm to maximize $\tilde{l} = \log \tilde{L}$, we construct the function

$$\tilde{Q}(\theta, \theta^{(p)}) = \sum_{i=1}^{n} w_{n,i} \sum_{j=1}^{K} \tau_{j,i}^{(p)} \log(\pi_j f_j(x_i)). \tag{1.27}$$

To satisfy (1.21), we define

$$\begin{aligned} \tilde{H}(\theta, \theta^{(p)}) &= \tilde{Q}(\theta, \theta^{(p)}) - \tilde{l}(\theta) \\ &= \sum_{i=1}^{n} w_{n,i} \sum_{j=1}^{K} \tau_{j,i}^{(p)} \log \tau_{j,i}. \end{aligned}$$

By a consequence of Jensen's inequality (see Rao, 1973, equation *1e.6.1*),

$$\tilde{H}(\theta, \theta^{(p)}) \leq \tilde{H}(\theta^{(p)}, \theta^{(p)}).$$

Therefore, in the p^{th} stage of the algorithm, construct $\tilde{Q}(\theta, \theta^{(p)})$ and then let $\theta^{(p+1)}$ be the value of θ maximizing this function. This gives a sequence of iterates which have

nondecreasing weighted likelihood. Convergence properties of the algorithm are likely to be similar to those of the standard EM algorithm. See Dempster *et al.* (1977), and Wu (1983). Note that the function $\tilde{Q}$ described here is not a conditional expectation of a weighted *complete data* likelihood.

1.6 Effective prior

We saw in the multinomial example of Section 1.4.1 that the WLB simulation is equivalent to sampling the posterior distribution induced by a conjugate prior, which equals the square of Jeffreys prior when all $\alpha_i = 1$. In general, the conditional distribution of $\tilde{\theta}$ is not exactly equal to a posterior distribution induced by any fixed, data-independent prior, as we saw in Section 1.4. As another example, consider data $X_1, \cdots, X_n$ having a Poisson distribution with mean θ. It is easily shown that $\tilde{\theta}_n = n^{-1} \sum_i w_{n,i} X_i$, and so the support of its distribution is the range of the observed data. That the likelihood in this case is positive on $(0, \infty)$ means that the *effective prior* is also supported by the range of the data. Consequently, this effective prior is not a prior density in the usual sense – it is simply the function $\pi_{0,n}$ which multiplies the likelihood to give the density of $\tilde{\theta}$. This section contains two more facts about the effective prior. The first is an asymptotic approximation is a special class of one parameter models. The second is an idea on how to modify a uniform Dirichlet weight distribution.

1.6.1 A limiting effective prior

In this section we study properties of the effective prior in certain smooth, one-parameter models. We assume that the weights have a uniform Dirichlet distribution. Using an asymptotic expansion for the moments of a posterior we derive an equation which the effective prior has to satisfy approximately. This leads to the result that for exponential family models parameterized by the mean, the effective prior is approximately equal to the square of the Jeffreys prior.

To proceed, suppose that data $X_1, X_2, \cdots, X_n$ are modeled as a random sample from some density $f_\theta(x)$ where θ is a real parameter. Let $\psi_i(\theta) = \log f_\theta(X_i)$ and define the scaled log-likelihood $l_n(\theta) = (1/n) \sum \psi_i(\theta)$. Sufficient smoothness conditions on $\psi_i(\theta)$ must be assumed to ensure that Taylor series expansions are valid. Let $\hat{\theta}$

be the maximum likelihood estimator of θ and denote k^{th} derivatives by bracketed superscripts. Thus $l_n^{(1)}(\hat{\theta}) = 0$. Assuming smoothness conditions on the effective prior $\pi_{0,n}$, the following expression for the expectation of $\tilde{\theta}$ can be derived following Lindley (1980), for example.

$$E(\tilde{\theta}|X_1^n) = \hat{\theta} - \frac{1}{nl_n^{(2)}(\hat{\theta})}\left(\frac{\pi_{0,n}^{(1)}(\hat{\theta})}{\pi_{0,n}(\hat{\theta})} - \frac{l_n^{(3)}(\hat{\theta})}{2l_n^{(2)}(\hat{\theta})}\right) + O\left(\frac{1}{n^2}\right) \ a.s.. \tag{1.28}$$

This expansion is derived in detail in Appendix A. Smoothness of the underlying model ensures that

$$l_n^{(2)}(\hat{\theta}) = -I(\theta_0) + o(1) \ a.s.$$
$$l_n^{(3)}(\hat{\theta}) = K(\theta_0) + o(1) \ a.s.$$

where θ_0 is the *true* parameter value generating the data, $I(\theta)$ is the Fisher information, and

$$K(\theta) = E\psi_1^{(3)}(\theta)\,.$$

If the sequence of effective priors $\pi_{0,n}$ and their derivatives $\pi_{0,n}^{(1)}$ have almost sure limits π_0 and $\pi_0^{(1)}$, then it follows from (1.28) that

$$E(\tilde{\theta}|X_1^n) = \hat{\theta} + \frac{1}{nI(\theta_0)}\left(\frac{\pi_0^{(1)}(\theta_0)}{\pi_0(\theta_0)} + \frac{K(\theta_0)}{2I(\theta_0)}\right) + o\left(\frac{1}{n}\right) \ a.s.. \tag{1.29}$$

Now consider a model in which $\hat{\theta}$ is the sample average and $\tilde{\theta}$ is a weighted average. We then have

$$E(\tilde{\theta}|X_1^n) = \hat{\theta}\,. \tag{1.30}$$

The only way that equations (1.29) and (1.30) can both be true is if the limiting effective prior π_0 satisfies the differential equation

$$\frac{\pi_0^{(1)}(\theta)}{\pi_0(\theta)} + \frac{K(\theta)}{2I(\theta)} = 0\,. \tag{1.31}$$

It turns out (see Theorem A.8 of Appendix A) that for exponential family models parameterized by $\theta = E(X)$, (so that the sample average of the natural sufficient statistic is the MLE),

$$K(\theta) = -2I^{(1)}(\theta). \tag{1.32}$$

Therefore in such models, the limiting effective prior, if it exists, is proportional to the square of the Jeffreys prior:

$$\pi_0(\theta) \propto I(\theta).$$

To see whether or not this is a good approximation, we compared the posterior under the limiting effective prior to the standard WLB for a small set of Poisson data. The data are counts of fossils of the extinct mammal *Litolestes notissimus* in thirty squares of a grid (example 3.1 from Olkin *et al.* 1980). The counts are 0 (16 times), 1 (9 times), 2 (3 times), 3 once, and 4 once. As we see from Figure 1.9, the effective prior is approximated better by the limiting effective prior $p(\theta) \propto \theta^{-1}$ than by the flat prior $p(\theta) \propto 1$. As the hanging boxplots indicate, there is still a fair bit of error in both of these approximations.

1.6.2 A non-uniform weight distribution

The Dirichlet weights w_n used in the WLB might be referred to as *vanilla* weights when all $\alpha_i = 1$. In particular, they do not use any information about the model or the data. By modifying the α_i we may be able to reduce the error of the WLB simulation; that is flatten the effective prior. Our development in this section hinges on the notion of maximum entropy (see Jaynes, 1968, for example).

Consider the special case where $\tilde{\theta} = n^{-1}\sum w_{n,i}x_i$ and suppose we know some properties of the posterior distribution, such as the first two moments, namely

$$E(\theta|x_1^n) = c_1$$
$$E(\theta^2|x_1^n) = c_2.$$

We can design our weights so that $\tilde{\theta}$ satisfies the same constraints. Note that

$$E(\tilde{\theta}|x_1^n) = \frac{\sum \alpha_i x_i}{\alpha.}$$
$$E(\tilde{\theta}^2|x_1^n) = \frac{\{(\sum \alpha_i x_i)^2 + \sum \alpha_i x_i^2\}}{\alpha.(\alpha. + 1)}$$

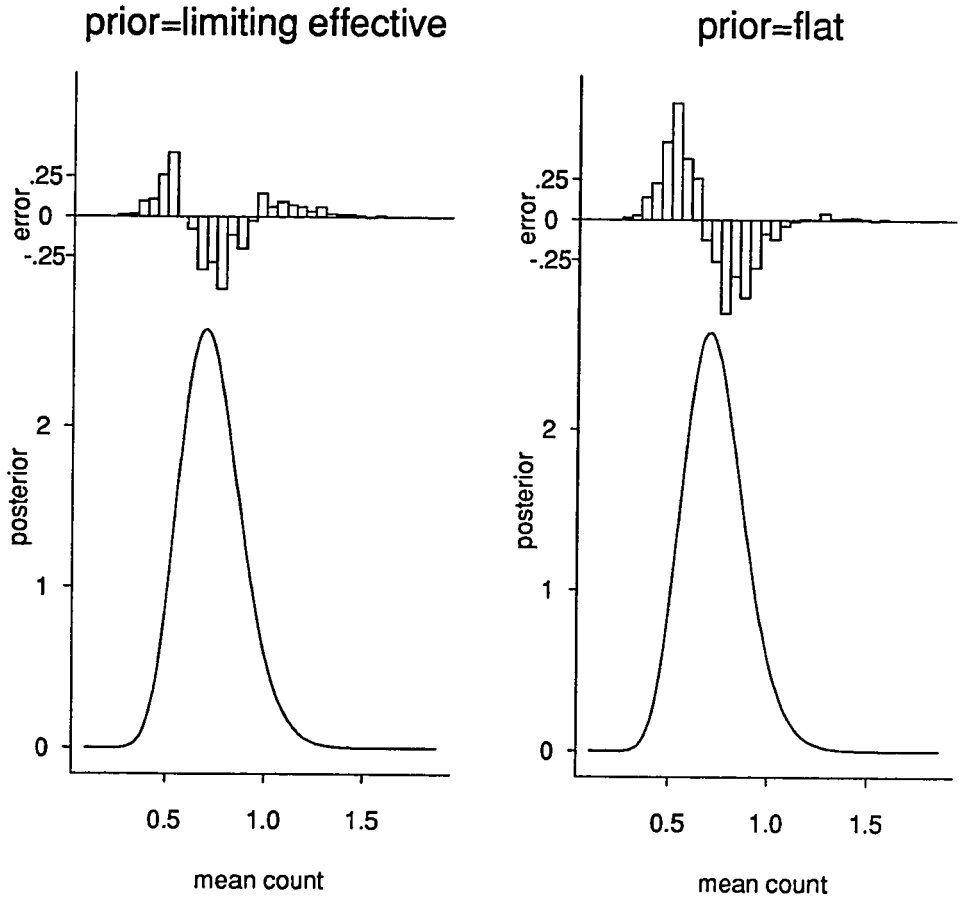


Figure 1.9: These plots compare the distribution of 5000 simulated $\tilde{\theta}$'s with two posterior distributions for the parameter of the fossil count example. The posterior density on the left side comes from the limiting effective prior $p(\theta) \propto \theta^{-1}$. The one on the right comes from a flat prior. The hanging boxplots above each density show the absolute error in the WLB approximation. Plotted is the difference between the WLB density estimate and the average posterior density in each bin.

where $\alpha_\cdot = \sum \alpha_i$. If we have $\alpha_\cdot = n$, then the collection of $\beta_i = \alpha_i/n$ forms a probability distribution on the data satisfying the constraints

$$\sum \beta_i x_i = c_1 := k_1$$
$$\sum \beta_i x_i^2 = (n+1)c_2 - nc_1^2 := k_2 .$$

Of course, many vectors $(\beta_1, \cdots, \beta_n)$ satisfy these constraints but only one is *as uniform as possible*, and that one maximizes the entropy $-\sum \beta_i \log \beta_i$. This vector has entries

$$\beta_i = \frac{e^{\lambda_1 x_i + \lambda_2 x_i^2}}{Z(\lambda_1, \lambda_2)}$$

where $Z(\lambda_1, \lambda_2)$ is the partition function ensuring that the β_i sum to one. The parameters λ_1, λ_2 are the solution of

$$\frac{\partial \log Z}{\partial \lambda_i}(\lambda_1, \lambda_2) = k_i \quad i = 1, 2 .$$

Figure 1.10 shows this idea applied to the following small sample of 24 failure times modeled as exponential:

3	5	5	13	14	15	22	22	23	30	36	39
44	46	50	72	79	88	97	102	139	188	197	210

(from Proschan, 1963, aircraft 7914, and later in Cox and Lewis, 1966). The WLB using the modified weights provides a better approximation than using the vanilla weights.

1.7 Connections to other bootstraps

We now consider the connections between this weighted likelihood method and the bootstrap methods of Efron (1979), Rubin (1981), and Künsch (1989). These connections can be clarified by considering specific models and alternative weighting schemes.

For instance, suppose that we restrict attention to models for iid data, and we construct multinomial weights as follows: let $\xi_1, \xi_2, \cdots, \xi_n$ be a random sample of

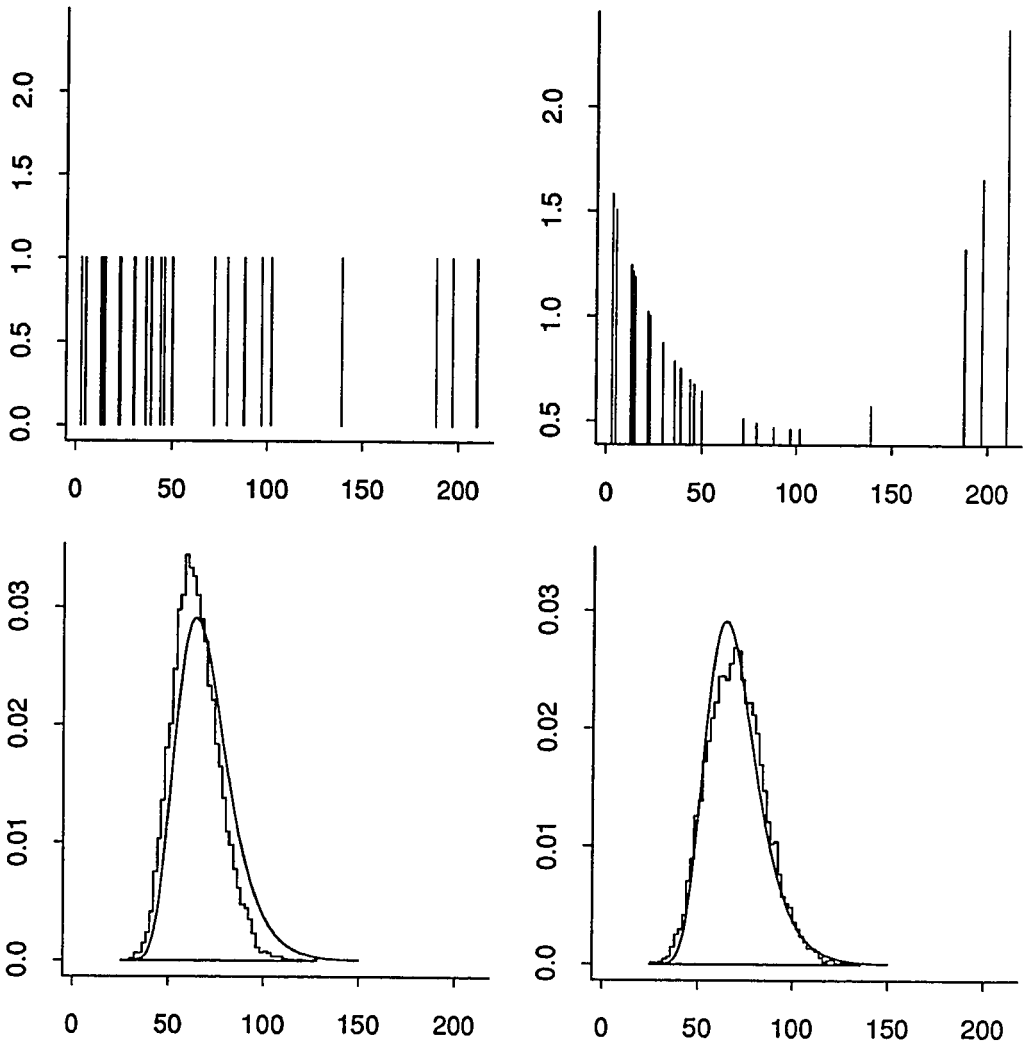


Figure 1.10: The lower panel shows the results of the WLB using the uniform weights (left) and using modified weights (right). The upper panel shows the parameter vector $(\alpha_1, \cdots, \alpha_n)$ used in each simulation. The parameters on the right are chosen so that the resulting distribution has the correct mean and variance while being as uniform as possible. This is accomplished by minimizing $\sum \alpha_i \log \alpha_i$ subject to $\sum \alpha_i = n$, $\sum \alpha_i x_i = nk_1$, and $\sum \alpha_i x_i^2 = nk_2$ for constants $k_1 = 69.9$, $k_2 = 10627$. The data are 24 failure times of air conditioners (Proschan, 1963), and solid lines show the likelihood under an exponential model. Histograms are based on 10000 draws.

uniform $(0,1)$ variables and put $m_n = (m_{n,1}, m_{n,2}, \cdots, m_{n,n})$ where

$$m_{n,i} = \sum_{k=1}^{n} 1\left[(i-1) < n\xi_k \leq i\right] .$$

With these multinomial weights replacing the scaled Dirichlet weights, the weighted likelihood function in equation (1.2) becomes

$$\begin{aligned} \tilde{L}_n(\theta) &= \prod_{i=1}^{n} [f_\theta(x_i)]^{m_{n,i}} \\ &= \prod_{i=1}^{n} f_\theta(x_i^*), \end{aligned}$$

where $x_1^*, \cdots, x_n^*$ is a bootstrap sample, namely a sample of size n with replacement from the original data $x_1, \cdots, x_n$. Thus with multinomial weights and the iid assumption, the WLB amounts to *nonparametrically* bootstrapping the maximum likelihood estimator. Indeed, the multinomial weights are quite similar to the Dirichlet weights as they have the same mean and almost the same covariance matrix (see Rubin 1981). We concentrate on Dirichlet weights because of their Bayesian justification as shown in Section 1.4.1.

In fact, we can view Efron's nonparametric bootstrap as a WLB of any statistic which is a function of the empirical distribution function. To see this, and to see how the weighted likelihood method generalizes Rubin's Bayesian bootstrap, we need the concept of empirical likelihood from Owen (1988, 1990).

1.7.1 Empirical Likelihood

Let $X_1, X_2, \cdots X_n \sim_{iid} F_0$ and suppose that $Z_1, Z_2, \cdots, Z_m$ are the distinct values of the observed X_i's. Consider the family of distributions that are absolutely continuous with respect to the empirical distribution function of the data:

$$\mathcal{F}_n = \{F : F << \mathsf{F}_n\} .$$

Here, F_n puts mass m_j/n on Z_j with m_j the number of X_i's equal to Z_j. Each $F \in \mathcal{F}_n$ is supported on $(Z_1, Z_2, \cdots, Z_m)$, and is associated with a vector of probability masses $(f_1, f_2, \cdots, f_m)$. The empirical likelihood function is defined to be

$$L_{emp}(F) = \begin{cases} \prod_{j=1}^{m} f_j^{m_j} & \text{if } F \in \mathcal{F}_n \\ 0 & \text{otherwise} \end{cases}$$

It can be shown, for example, that F_n maximizes L_{emp} over all distributions on the sample space.

1.7.2 Weighted Empirical Likelihood

Treating L_{emp} as an ordinary likelihood function, we can define a weighted empirical likelihood by analogy with (1.2). Let $(w_{n,1}, w_{n,2}, \cdots, w_{n,n})$ be an appropriate set of weights, either scaled Dirichlet or multinomial, and put

$$\gamma_j = \sum_{i=1}^{n} w_{n,i} 1[X_i = Z_j] .$$

Define the weighted empirical likelihood $\tilde{L}_{emp}$ as

$$\tilde{L}_{emp}(F) = \begin{cases} \prod_{j=1}^{m} f_j^{\gamma_j} & \text{if } F \in \mathcal{F}_n \\ 0 & \text{otherwise} \end{cases}$$

This is the obvious definition since for $F \in \mathcal{F}_n$,

$$\tilde{L}_{emp}(F) = \prod_{i=1}^{n} f_{j_i}^{w_{n,i}}$$

where j_i is the index such that $X_i = Z_{j_i}$.

By a standard result (see Rao, 1973, p. 58), the point $\tilde{F}$ maximizing $\tilde{L}_{emp}$ puts mass γ_j/n at Z_j. For multinomial weights, $\tilde{F}$ is precisely the empirical distribution function of a nonparametric bootstrap sample. For scaled uniform Dirichlet weights, $\tilde{F}$ is precisely the distribution function sampled in the Bayesian bootstrap simulation of Rubin (1981).

Inference by the WLB and the Bayesian bootstrap may be the same even for parametric likelihood when applied to particular parameters. Consider the Poisson mean, for example. If θ is the Poisson mean, then $\tilde{\theta}_n$ is $\sum_i w_{n,i} X_i/n$. Both the Bayesian bootstrap and the WLB involve simulating uniform Dirichlet weight vectors and recomputing $\tilde{\theta}_n$. Thus if inference is restricted to the mean, then the WLB and the Bayesian bootstrap are equivalent. However, the WLB is viewed as an approximation to parametric inference rather than valid nonparametric inference. Suppose, for example, that we want to know the posterior distribution of the parameter $\phi = P(X > x)$. If x is larger than the all the data points, then the Bayesian

bootstrap posterior for ϕ puts point mass at 0. The WLB uses the structure of the assumed model and so views ϕ as a fixed function of θ. Simulating an approximate posterior for ϕ by the WLB amounts to simulating $\tilde{\theta}_n$'s and each time computing $\tilde{\phi}$. This estimated posterior does not have point mass at 0.

1.7.3 Blockwise Bootstrap

A special case of Künsch's (1989) bootstrap for stationary observations can also be viewed as a WLB. To see this, first suppose that we model the sequence $X_1, \cdots, X_n$ by assuming a Markov property

$$f_\theta\left(x_i|x_1^{i-1}\right) = f_\theta\left(x_i|x_{i-p+1}^{i-1}\right) \quad p \geq 2\,.$$

Suppose further that we condition on the initial $p-1$ observations so that (1.2) becomes

$$\tilde{L}_n(\theta) = \prod_{i=p}^{n}\left[f_\theta\left(x_i|x_{i-p+1}^{i-1}\right)\right]^{w_{n-p+1,i}}\,.$$

Then $\tilde{\theta}_n$ satisfies

$$\sum_{i=p}^{n} w_{n-p+1,i}\psi(x_{i-p+1}, \cdots, x_i; \tilde{\theta}_n) = 0 \tag{1.33}$$

for a score function ψ.

Now turning to Künsch's (1989) *blockwise* bootstrap: The first step of Künsch's procedure is to construct, for some p, overlapping blocks $Y_i = (X_i, X_{i+1}, \cdots, X_{i+p-1})$ for $i = 1, 2, \cdots, k = n-p+1$. Statistics that are functions of the empirical distribution of these Y_i can be bootstrapped. For example, $\hat{\theta}_n$ satisfying

$$\sum_{i=1}^{k}\psi(Y_i; \hat{\theta}_n) = 0$$

is such a statistic. While Künsch's bootstrap involves a second level of blocking, we consider here the special case where this second level block size is one. In this case, the blockwise bootstrap amounts to Efron's bootstrap applied to the initial blocks $Y_1, \cdots, Y_k$. Therefore, the bootstrapped statistic $\hat{\theta}_n^*$ satisfies

$$\sum_{i=1}^{k} m_{k,i}\psi(Y_i; \hat{\theta}_n^*) = 0$$

for multinomial weights $m_{k,i}$ as above establishing the connection through (1.33).

The WLB is not restricted to Markov models like the one above. For instance, using recursive updating, we can compute factors in the likelihood of the standard state-space model

$$Y_t = FX_t + \epsilon_t$$
$$X_t = GX_{t-1} + \delta_t,$$

which is not Markovian for the observed $\{y_t\}$. We may also readily apply the WLB to inference about long-memory time series models such as the fractional differencing model (Hosking, 1981; Haslett and Raftery, 1989).

1.7.4 A Comparison with Efron's Bootstrap

We did a small simulation experiment to compare the WLB with Efron's bootstrap. In particular, we are interested in multiple linear regression models where the regressors are considered fixed. For such models, the data are not iid, and so Efron's nonparametric bootstrap does not apply. However in practice, the simulation proceeds by resampling residuals from the fitted model. These bootstrap residuals are then tagged on to the fits to give bootstrap data. Of course, if the regressors are not considered fixed, then the standard bootstrap algorithm applies. We are also interested in how the dimension of the model affects the different bootstrap procedures.

Our experiment is as follows. Firstly, we generate one parameter vector θ of length p_{max} and one design matrix x of size $n \times p_{max}$ where $n = 15$ and $p_{max} = 10$. The parameter vector is drawn from a multivariate normal and the design matrix has one column of ones and then 9 columns of realized uniform $(0,1)$ random variables. The design matrix and the parameter vector are fixed throughout the experiment. For each p between 2 and p_{max}, we generate 500 regression samples of size n. That is, with x_p the first p columns of x and θ_p the first p elements of θ, we simulate 500 Y's from the model

$$Y = x_p\theta_p + \epsilon \qquad \epsilon \sim N_n(0, \sigma^2 I_n)$$

where $\sigma^2 = .04$. For each of these simulated data sets we construct two confidence sets; one by the WLB and one by Efron's bootstrap. The confidence sets are for a

parameter representing the predicted value of the response at a new covariate value x_{new} (in the center of the design space). We thus have confidence intervals for a one dimensional parameter even though p varies. The confidence intervals are constructed by the percentile method at each of three nominal levels using 1000 bootstrap (or WL bootstrap) samples. The results are summarized in Figure (1.11).

On all accounts, the confidence intervals constructed by the WL bootstrap have better coverage. Not surprisingly, they also are longer on average. Neither confidence interval gets very close to the nominal coverage, and although many modifications of the basic interval are possible, none are tried here. Of course it would be ridiculous to fit a 10 parameter model to 15 observations, nonetheless this experiment does give us some insight into the dependence of a bootstrap on the dimension of the model. Another interesting fact lies in the nature of the bootstrapping. For Efron's bootstrap one must decide between sampling the raw observations or the residuals. No such decision is involved in the WL bootstrap, and the result is actually somewhere in between these two positions. The simulated $\tilde{\theta}$'s take the form of weighted least squares estimates,

$$\tilde{\theta}_n = (X^t W X)^{-1} X^t W Y$$

where X is the design matrix, Y is the vector of responses and W is a diagonal matrix having a uniform Dirichlet vector on the diagonal.

1.8 Discussion

We have introduced a bootstrap-like procedure for simulating, at least approximately, from the posterior distribution of a parameter. In models for discrete data, this weighted likelihood bootstrap works quite well, and it is very easy to implement. More examples are studied in the next chapter. The WLB generalizes Rubin's (1981) Bayesian bootstrap from strictly nonparametric models to parametric ones. The WLB is different from the so-called parametric bootstrap (Efron, 1982), although both procedures use the structure of the assumed model. Of course, the parametric bootstrap is used to estimate sampling distributions, not posterior distributions. For dependent data models, the WLB can be shown to have some features in common with the blockwise bootstrap of Künsch. We gain some intuition for the WLB by comparing it with the Dirichlet sampling process (Tanner and Wong, 1987).

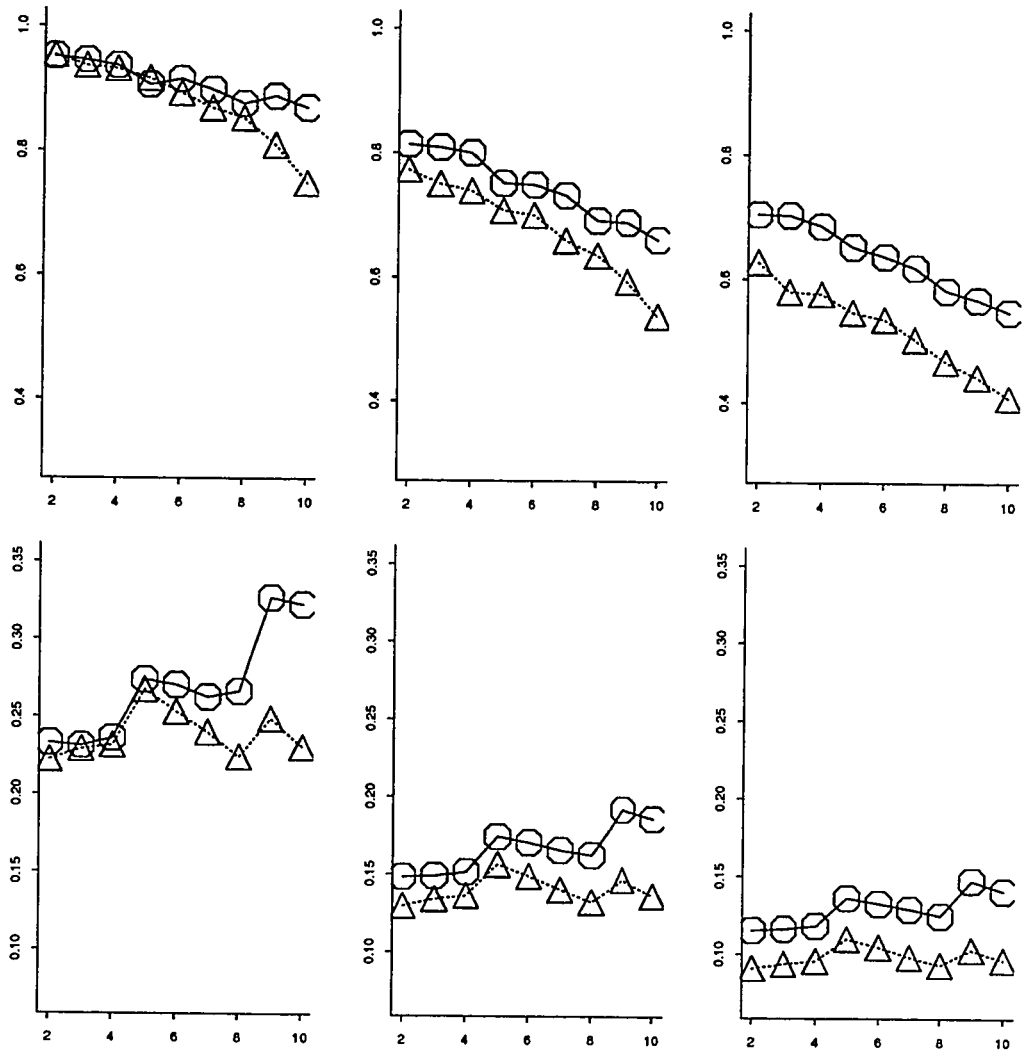


Figure 1.11: This plot summarizes a small simulation experiment to compare the WLB with Efron's bootstrap. The model is multiple linear regression, with fixed regressors, as described in Section 1.7.4. Each of the 3 columns above corresponds to a different nominal level of the confidence interval for x_{new}: 99, 90, and 80 percent. The top row shows observed coverage probabilities and the bottom row shows average interval lengths. The horizontal axis in each plot shows the dimension of the regression model; from 2 to 10 parameters. Circles correspond to the WLB and triangles to Efron's bootstrap.

Other bootstrap methods based on random weights have been proposed (Zheng and Tu, 1988 and references therein; Mason and Newton, 1991), but these involve weighting observations in a random sample rather than factors in a more general likelihood, and aim to evaluate sampling distributions rather than posterior distributions. Boos and Monahan (1986) have studied the use of the Efron (1979) bootstrap to approximate a posterior distribution through the sampling distribution of a pivot; the methods discussed here, by comparison, apply in the absence of pivotal quantities.

Perhaps the most interesting areas for further research are questions about the effective prior and the distribution of the weights. For a small class of models, we can roughly determine what prior distribution the WLB corresponds to. It would be important if by some simple modification of the weights, the effective prior became roughly uniform. We found such a modification for the linkage example (Section 1.4.2). We computed an effective prior for the logistic regression model (Section 1.4.3) by supposing a uniform effective prior on one scale and then determining the Jacobian of a nonlinear transformation. No general recipe yet exists for finding this prior. The maximum entropy argument from Section 1.9.2 gives us parameters of the weight distribution which are as uniform as possible while satisfying certain constraints. Without knowledge of the posterior, it is not clear what constraints we can put on the weights. However, perhaps by constraining the mean (or mode) and variance from the normal approximation, we can pick up skewness through the WLB.

One potentially useful application of the WLB is as the source of initial samples for importance sampling. A nagging problem with importance sampling is that to obtain any reasonable level of efficiency, the importance sampling function must be close to the density of interest. Results presented here suggest that the WLB is simulating a density quite close to the posterior density of interest, and so corresponding importance sampling weights, being density ratios, may not stray too far from one. Of course, we must be able to evaluate the density of $\tilde{\theta}$ which we can do by constructing a kernel density estimate for example. In the same way, the WLB can provide initial samples for the sampling importance-resampling (SIR) algorithm of Rubin (1987, 1988). (The SIR algorithm has been called a weighted bootstrap by Smith and Gelfand (1990), not to be confused with the WLB.) Adaptive importance sampling (Evans 1991) might be useful in this regard. In this setup, there is a whole family of importance sampling functions. We sample initially at one, make a com-

parison, and then find a new member of the class which is closer to the density of interest. For the WLB, the family of importance sampling distributions is the family of Dirichlet distributed weights. We may be able to search through this family to give us a good weight distribution.

Chapter 2

APPLYING THE WEIGHTED LIKELIHOOD BOOTSTRAP

This chapter shows the weighted likelihood bootstrap applied to a number of models common in statistical analysis.

2.1 Nonlinear Regression and Turkey Feed

Figure (2.1) shows data recording the four–week weight of turkeys fed various percentages of one of two different sources of a hormone (from Noll *et al.* 1984, and then Weisberg, 1985). The following nonlinear regression model was contemplated to determine the difference, if any, between the two sources:

$$y_i = \theta_1 + \theta_2(1 - \exp\{-\theta_3 z_{1,i} - \theta_4 z_{2,i}\}) + \epsilon_i \tag{2.1}$$

where ϵ_i form a random sample of mean 0 normal random variables with variance θ_5. The covariates $z_{1,i}$ and $z_{2,i}$ record the dose of each hormone as a percentage of the diet. The experiment is designed and the covariate vectors are orthogonal.

The WLB is readily applied here by using IRLS as described in Section 1.5.1. To see how well the procedure approximates the posterior distribution of the parameters, we compare the output to the results of running a Gibbs sampler on the problem. The Gibbs sampler is quite an attractive tool as it gives, in principle, an arbitrarily good approximation to the posterior in this example. To run the sampler though, we must be able to sample from the full conditional distributions of each parameter. Such a distribution, written

$$[\theta_i | \theta_j \; j \neq i, \text{ data}]$$

is the conditional posterior distribution of θ_i given that the other parameters are fixed. The full conditional distributions of θ_1 and θ_2 and θ_5 have simple expressions:

$$[\theta_1 | \theta_j \; j \neq 1, \text{ data}] = \text{Normal}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \eta_i), \frac{1}{n}\theta_5\right)$$

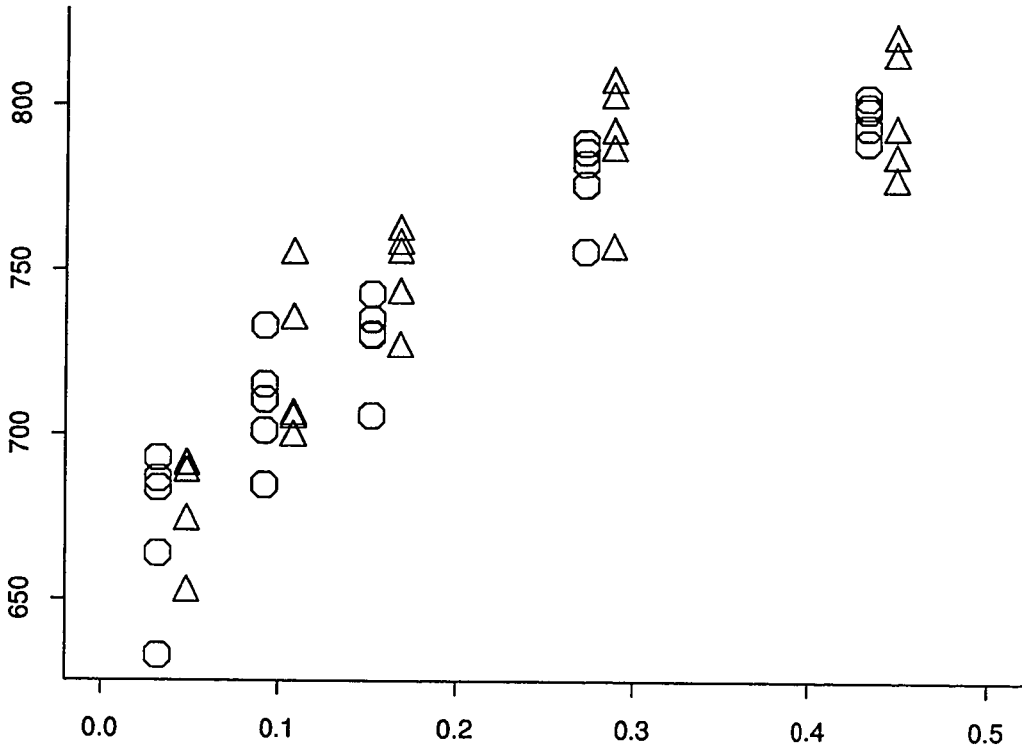


Figure 2.1: Data on an experiment to study a hormone on the growth of young turkeys is shown above. The x-axis records the level of a hormone in the diet (as % of diet) of young turkeys. The y-axis shows the weight, in grams, of the turkeys after four weeks. There are two treatments, indicated by circles and triangles, since the hormone can enter the diet in one of two ways. The pictured data are actually simulated but share the same summary statistics as reported (only summaries are reported by Noll et al., 1984).

$$[\theta_2|\theta_j\ j \neq 2,\ \text{data}] = \text{Normal}\left(\frac{\sum_{i=1}^n (y_i - \theta_1)\gamma_i}{\sum_{i=1}^n \gamma_i^2},\ \sum_{i=1}^n \gamma_i^2\right)$$

$$[(\theta_5)^{-1}|\theta_j\ j \neq 5,\ \text{data}] = \text{Gamma}\left(\frac{1}{2}\sum_{i=1}^n (y_i - \mu_i)^2,\ \frac{n}{2}\right)$$

where

$$\gamma_i = \exp\{-\theta_3 z_{1,i} - \theta_4 z_{2,i}\}$$
$$\eta_i = \theta_2(1 - \gamma_i)$$
$$\mu_i = \theta_1 + \eta_i\,.$$

Nonlinearity of the model hinges on the parameters θ_3 and θ_4 which do not have simple full conditional distributions. These conditional densities are proportional to the likelihood function, but that is all we know. The important fact is that individually, the full conditional densities are one dimensional, and so a numerical approximation should be relatively simple. Code is developed to sample a discrete approximation of these two conditional posteriors. This discrete approximation is formed by chopping up the support of the full conditional distribution of the parameter (either θ_3 or θ_4). We construct a grid of 400 points which covers that portion of the support contained within 8 standard deviations either side of the maximum likelihood estimate. The standard deviations are simply computed from the inverse Fisher information matrix. On this grid, we evaluate the likelihood function. After normalization, we then have a discrete approximation to the conditional posterior density of that parameter (θ_3, or θ_4). Note that every time we have to sample a conditional distribution of θ_3 or θ_4, the discrete approximation has to be recalculated because it depends on the values of the other parameters which are continually changing.

Figure (2.2) compares the WLB and the Gibbs sampler for this nonlinear regression model. For each of the four regression parameters in equation (2.1), estimates of the marginal posterior density are plotted. Both estimates (WLB and Gibbs sampler) are determined by applying a Gaussian kernel density estimator to 1000 simulated parameter values. In the Gibbs sampler algorithm, we save every 25'th value in a single Markov chain after a 'burn in' of 100. This chain is formed by randomly choosing an index from 1 to 5 and then updating the margin having that index by drawing from the associated full conditional distribution. This is a slightly different procedure

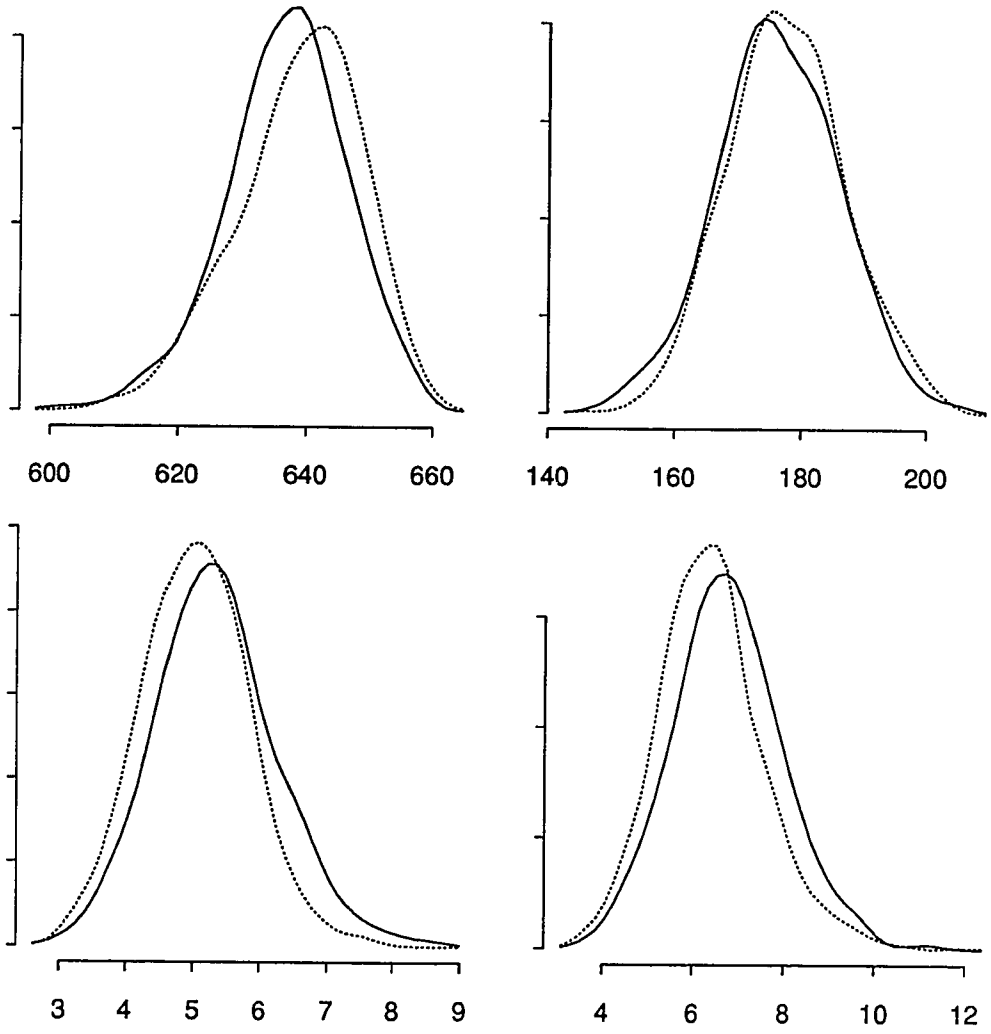


Figure 2.2: A comparison of the WLB and the Gibbs sampler under a flat prior is shown above. The model is the nonlinear regression described in Section 2.1. The four plots above show the estimated marginal posterior densities for each of the four regression parameters. The solid lines represent the Gibbs sampler while the dashed lines show the WL bootstrap solution. These curves are Gaussian kernel density estimates based on samples of size 1000. As suggested by the maximum smoothing principle of Terrel (1990), the standard deviation of the kernel is $3/\left(35N\sqrt{4\pi}\right)^{1/5}$ times the standard deviation of the sample ($N = 1000$).

than the standard Gibbs sampler where the coordinates are sampled in turn, and one step of the chain is accomplished after one complete scan of the coordinates.

Although not exact, the WL bootstrap appears to be doing reasonably well. The data shown in Figure (2.1) are not real data, but rather are simulated data sharing the same summary statistics as the published data. The problem is that only summary data from this experiment on turkey growth are published. We performed the above analysis on several other simulated data sets and got similar results.

2.2 Normal mixtures and classification

Although statistical inference for mixture models has a long history (Everitt, 1985), difficult technical problems remain. Consider, for example, the use of mixture models for classification (McLachlan and Basford, 1988). It is not at all obvious how one assesses the uncertainty of a classification. Simulation methods, in particular the WLB, offer potential solutions.

Consider, as an example, $n = 200$ bivariate observations modeled as a mixture of two bivariate normal populations. Each observation has density

$$\begin{aligned} f_\theta(x) &= \pi_1 g_1(x) + (1-\pi_1) g_2(x) \\ &= \pi_1 \frac{1}{2\pi|\Sigma_1|} \exp\left\{-\frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1)\right\} \\ &\quad + (1-\pi_1)\frac{1}{2\pi|\Sigma_2|} \exp\left\{-\frac{1}{2}(x-\mu_2)^t \Sigma_2^{-1}(x-\mu_2)\right\} \qquad x \in \mathbb{R}^2 . \end{aligned}$$

The parameter θ here is formed from the mixing probability π_1, the two mean vectors μ_1, μ_2, and the two covariance matrices Σ_1 and Σ_2, and hence lives in $\mathbb{R}^{11}$. The upper left panel of Figure 2.3 shows 200 points simulated from this model for a particular choice of θ. Using the EM algorithm, this model is fit to these data, yielding the estimated density whose contour lines are shown in the upper right panel of Figure 2.3. Plotted in the lower panel of Figure 2.3 are the data, now classified into two groups by the estimated *dividing curve*: the locus of points $x \in \mathbb{R}^2$ where

$$\hat{\tau}(x) = \frac{\hat{\pi}_1 \hat{g}_1(x)}{\hat{\pi}_1 \hat{g}_1(x) + (1-\hat{\pi}_1)\hat{g}_2(x)} = \frac{1}{2} .$$

Here, $\hat{g}_j$ indicates the estimated component density. (Note the actual likelihood function is unbounded in this example. If, however, we force the covariance matrices

to have determinant larger than some ϵ, then the root of the likelihood equation having highest likelihood can be viewed as the MLE.)

Figure 2.4 summarize an application of the WLB to the classification method described above. Essentially, we are bootstrapping the dividing curves to get some idea of the uncertainty in the classification.

2.3 Spectral analysis

An important feature of the WLB is its applicability to dependent data models. From equation (1.2) we see that many different weighted likelihood functions could be defined for a given dependent data model (one for each different factorization of the joint density function). For time-series models, the natural ordering leads to a canonical weighted likelihood function.

The upper left plot of Figure 2.5 shows a simulated realization of a Gaussian AR(3) time series model with parameter $\phi = (.13, -.15, .20)$ and error variance $\sigma^2 = 1$. The spectrum for such a process is

$$h(\omega;\phi) = \frac{\sigma^2}{|1 - \sum_{k=1}^{3} \phi_k \exp(-2\pi i \omega k)|^2}$$
$$= \frac{\sigma^2}{1 - 2\sum_{k=1}^{3} \phi_k \cos(2\pi\omega k) + \left(\sum_{k=1}^{3} \phi_k \cos(2\pi\omega k)\right)^2 + \left(\sum_{k=1}^{3} \phi_k \sin(2\pi\omega k)\right)^2}$$

for frequencies $0 \leq \omega \leq 1/2$. The upper right plot of Figure 2.5 shows this estimated spectrum. Inference about some aspect of the spectrum is often desired. Suppose for example, we are interested in estimating the posterior distribution of the parameter θ which equals the fraction of power from frequencies no greater than 0.2 cycles per unit time. This parameter is a ratio of integrals of the spectral density, and so inference by analytical methods may be difficult.

The WLB is easily carried out here. For simplicity, we condition on the first three observed X_i's so that $\tilde{\phi}$ can be computed by regressing the present on the past. For each simulated $\tilde{\phi}$, we compute $\tilde{\theta}$ by numerically integrating $h(\omega;\tilde{\phi})$. The histogram in Figure 2.5 summarizes the 500 simulated $\tilde{\theta}$'s. We see immediately that quite a bit of uncertainty is left in our knowledge of the power fraction θ.

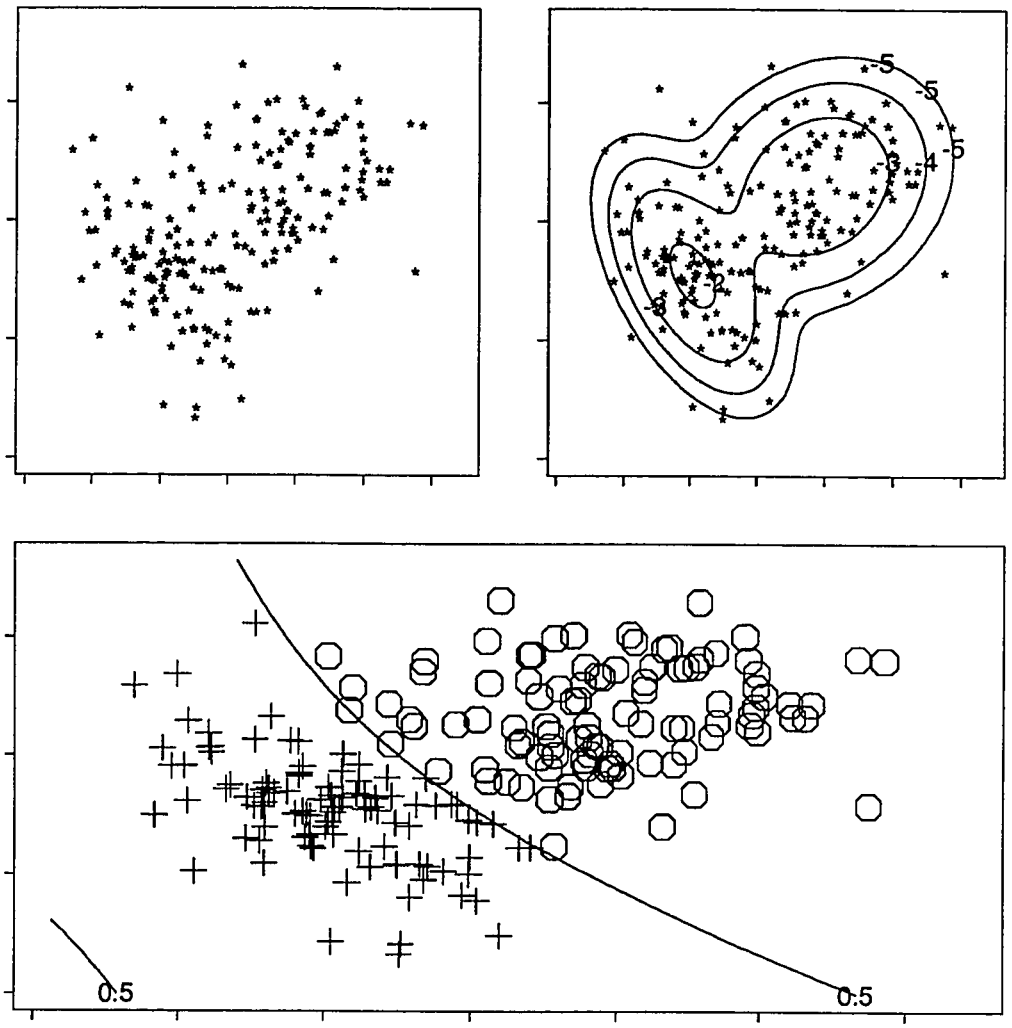


Figure 2.3: The upper left plot shows 200 simulated observations from a mixture of two bivariate normal distributions. To its right, a plot shows the same data overlaid with a contour map of the estimated density (estimated by maximum likelihood). The lower plot shows the corresponding maximum likelihood classification. The solid line, which indicates that the posterior probability of group membership equals .5, divides the data into two groups, denoted by circles and crosses.

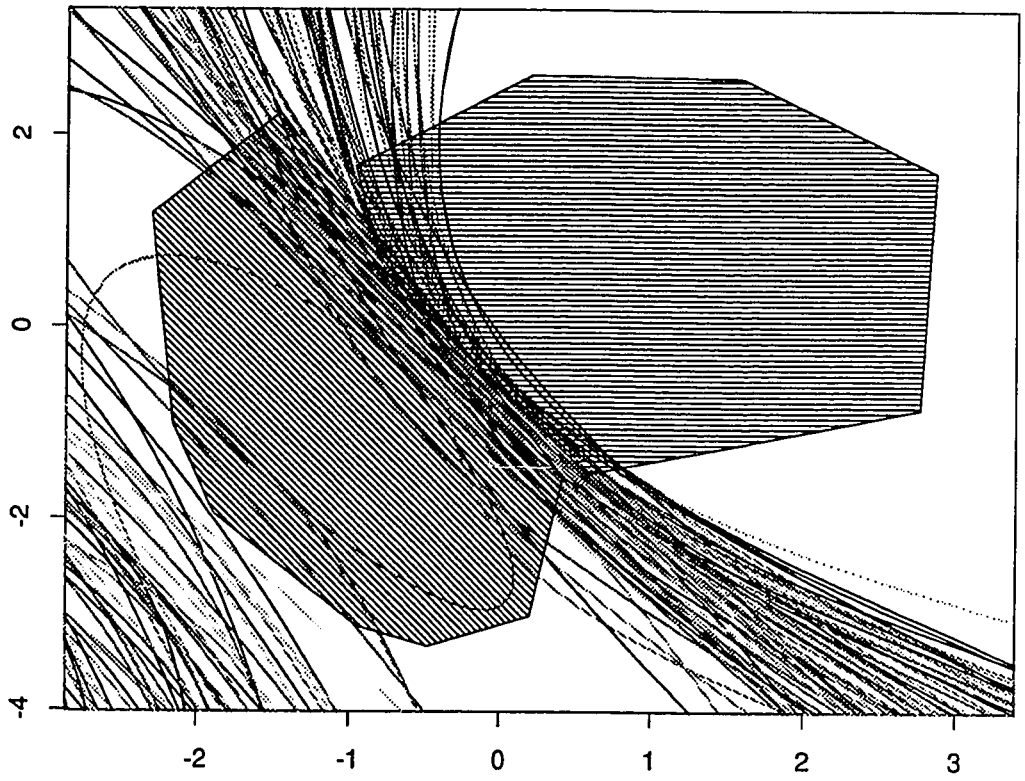


Figure 2.4: This rather dramatic picture is an attempt to assess the uncertainty in the classification shown on the lower plot of Figure 2.3. The shaded regions are the convex hulls of the data classified by the maximum likelihood classification. The collection of curves represent a sampling of 85 dividing curves from their posterior distribution. This sampling is done by the WLB, by first sampling parameters $\tilde{\theta}$ and then computing the corresponding dividing curves. For each curve shown, the estimate used was the most likely root of the likelihood equation.

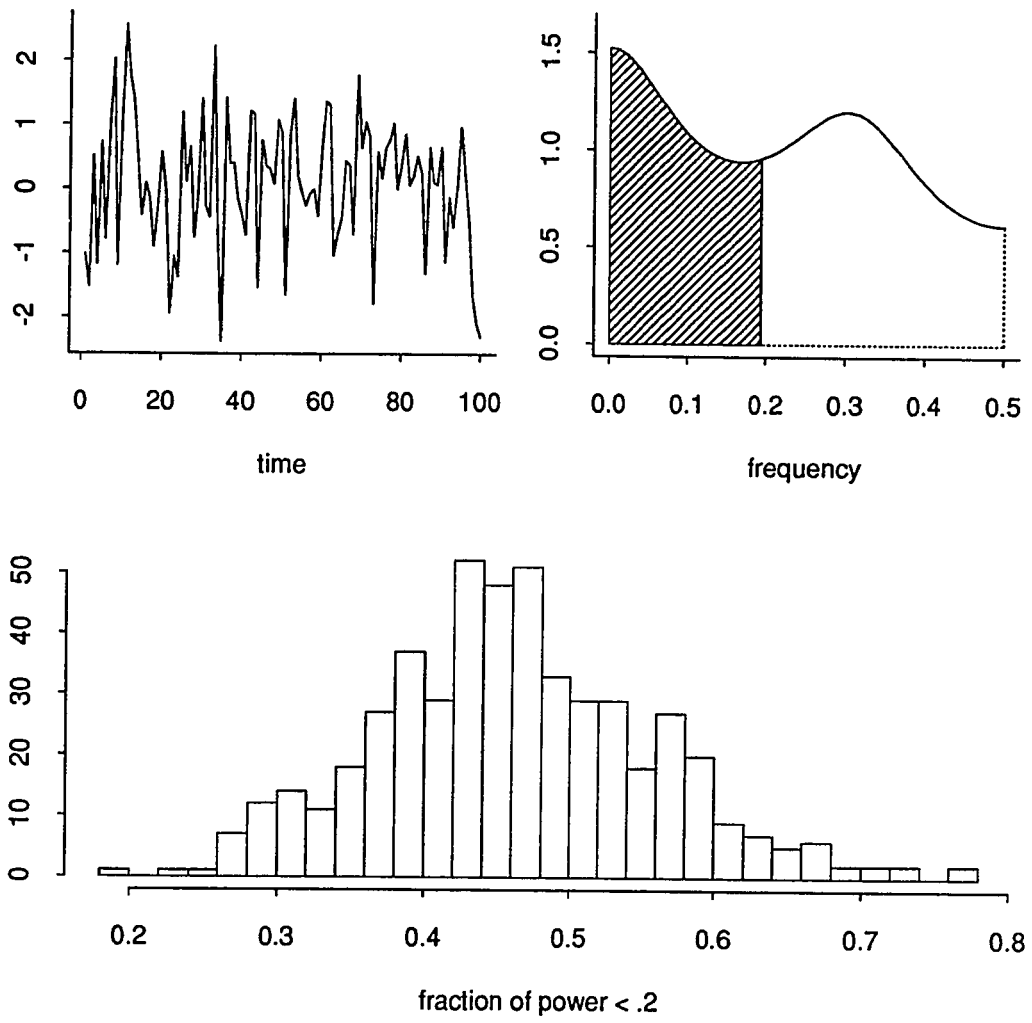


Figure 2.5: These plots summarize a simulation experiment using the WLB with spectral analysis methods. The upper left plot shows a simulated Gaussian $AR(3)$ time series of length 100 (from parameters $\phi = (.13, -.15, .20)$. The upper right plot shows the estimated spectral density function for this series. The shaded part indicates the amount of power at low frequencies (less than .2 cycles per time unit). In this example, 45% of the power is estimated to be at these low frequencies. Results of running WLB on the power fraction parameter are shown in the histogram. See Section 2.3.

2.4 *A bimodal posterior*

We construct an artificial example to show how the WLB works when the posterior distribution is bimodal. Data X_i are modeled as a random sample from a normal distribution with mean θ and variance θ^2. Figure 2.6 shows the normalized likelihood and histogram of 1000 samples of the WLB. Of course the normal approximation fails completely in this example. The WLB approximation is reasonable. Note that if such a posterior distribution were to exist in two or more dimensions, the Gibbs sampler would work because the induced Markov chain would not be irreducible.

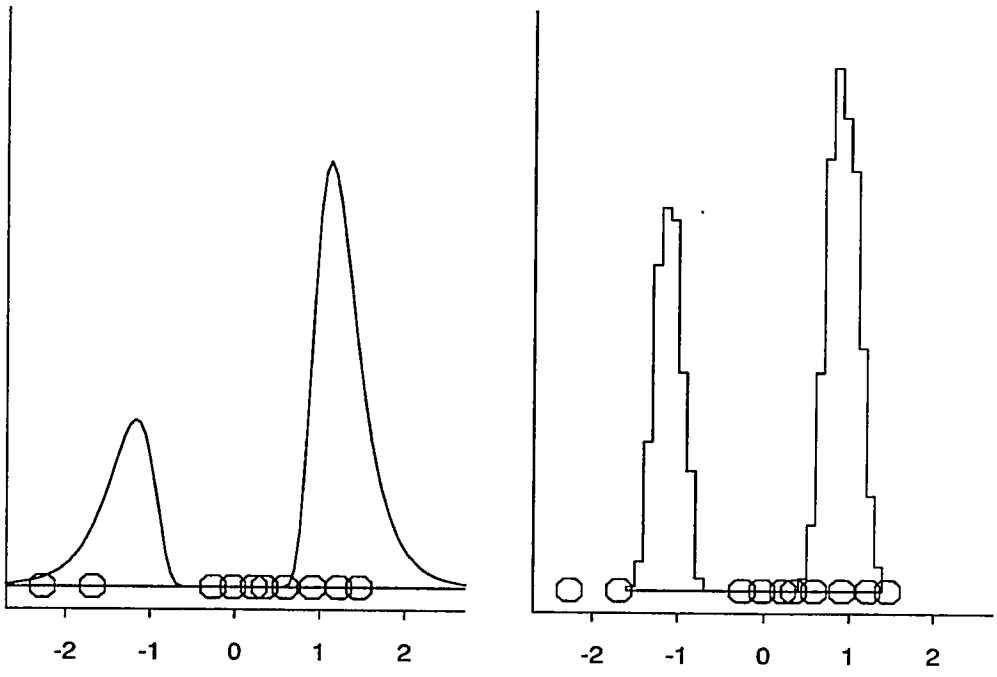


Figure 2.6: The figure on the left shows the normalized likelihood function for θ given the data (circles) when the model is normal with mean θ and variance θ^2. On the right is a histogram of 1000 draws from a WLB.

2.5 *Prediction*

Since the WLB permits approximate simulation from a posterior, it allows all sorts of inference questions to be addressed. For example, Bayesian predictive distributions

are typically mixtures over the posterior distribution of a parameter. Therefore, simulation from such predictive distributions can be done in two steps. First, we simulate parameter values using the WLB. Second, for each parameter value from the first stage we simulate a prediction according to the model. An alternative here is to average conditional predictive distributions.

Very often, predictive distributions are required in time series analysis. Specifically, suppose data X_t are modeled by

$$X_t = f_{\theta,p}(X_1^{t-1}, \epsilon_1^{t-1}) + \epsilon_t$$

where ϵ_t is a Gaussian white noise process, and $f_{\theta,p}$ is a fixed function given θ and p. Typically, p determines how many past values of the observation and noise process are used to predict X_t, and θ is a real parameter vector of length determined by p. In Box-Jenkins terminology, p describes the model specification; for example $p = (0, 1, 2)$ in the class of ARIMA models specifies a second order moving average model after one differencing. Conditioning on the model specification (i.e. p and the form of f), a predictive density for X_{t+s} given X_1^t can be written as a mixture

$$P(X_{t+s}|X_1^t) = \int\int P(X_{t+1}^{t+s}|X_1^t, \theta)\, dX_{t+1}^{t+s-1}\, d\theta\,. \tag{2.2}$$

We can simulate from this predictive distribution in two steps. First, we generate $\tilde{\theta}$'s by a WLB. For each $\tilde{\theta}$, we use the model structure to simulate future values of the series X_{t+1}^{t+s}. The resulting X_{t+s}'s are approximately draws from the desired predictive distribution.

The advantage of a Bayesian approach to the prediction problem is that uncertainty about the model parameters gets propagated into our conclusions. Failure to take account of this uncertainty typically leads to underestimation of the uncertainty about the quantity being predicted (Aitchison and Dunsmore, 1975). In decision-making problems, this leads to bias in favor of decisions that are favored by more certain information and of overly risky courses of action (Hodges, 1989). In principle, uncertainty about the model specification can also be considered.

Economic time series are often modeled using Box-Jenkins methods. Two typical series is shown in Figures 2.7 and 2.8 (from Nelson and Plosser, 1982). The data are post-World-War II yearly summaries, and consequently form fairly short series. Figure 2.7 shows the first-differenced series of log of U.S. money stock – a

measure of the amount of money in the economy. Figure 2.8 shows yearly average U.S. unemployment rate.

Prediction is often of interest, but in the Box-Jenkins formalism, there is no obvious non-Bayesian way to take into account uncertainty about parameters of the fitted model. Prediction intervals designed to have some nominal coverage are typically too short because model uncertainty and parameter uncertainty are not incorporated into the predictions. The WLB provides a simple way around this problem by allowing approximate simulation of the Bayesian predictive distributions. Figures 2.7 and 2.8 also shows the results of running the WLB. This illustrates the fact that the standard Box-Jenkins method can substantially underestimate the prediction variance.

The Bayesian solution to the calibration problem also involves a mixture over a posterior distribution similar to equation (2.2).

2.6 Calibration

Many authors have studied the problem of calibration. A detailed account is given in Aitchison and Dunsmore (1975). In a recent paper, Racine-Poon (1988) uses various approximations to compute the Bayesian solution to a particular nonlinear calibration problem. To briefly review the calibration problem, recall that the data come in two groups D_c, the calibration data, and D_p the prediction data. The calibration data D_c consist of pairs (Y_i, X_i). In Racine-Poon's case, X_i is the concentration of an agrochemical in soil, and Y_i is the weight of nasturtium plants in a pot containing that soil. Weight is modeled as a function of agrochemical concentration using the relationship

$$Y_i = f(X_i, \theta) + \epsilon_i$$

where ϵ_i are independent normal random errors. In this example, f has the form

$$f(x, \theta) = \theta_1/(1 + \exp(\theta_2 + \theta_3 \log x))$$

for $x > 0$ and θ_1 for $x = 0$. Unlike for a regression model, interest here is not directly on the parameter θ determining the effect of agrochemical concentration on weight. Rather, for the calibration problem, there is a set D_p of weights Y_i that correspond to soil of some unknown agrochemical concentration. Of interest is the value η of this unknown concentration.

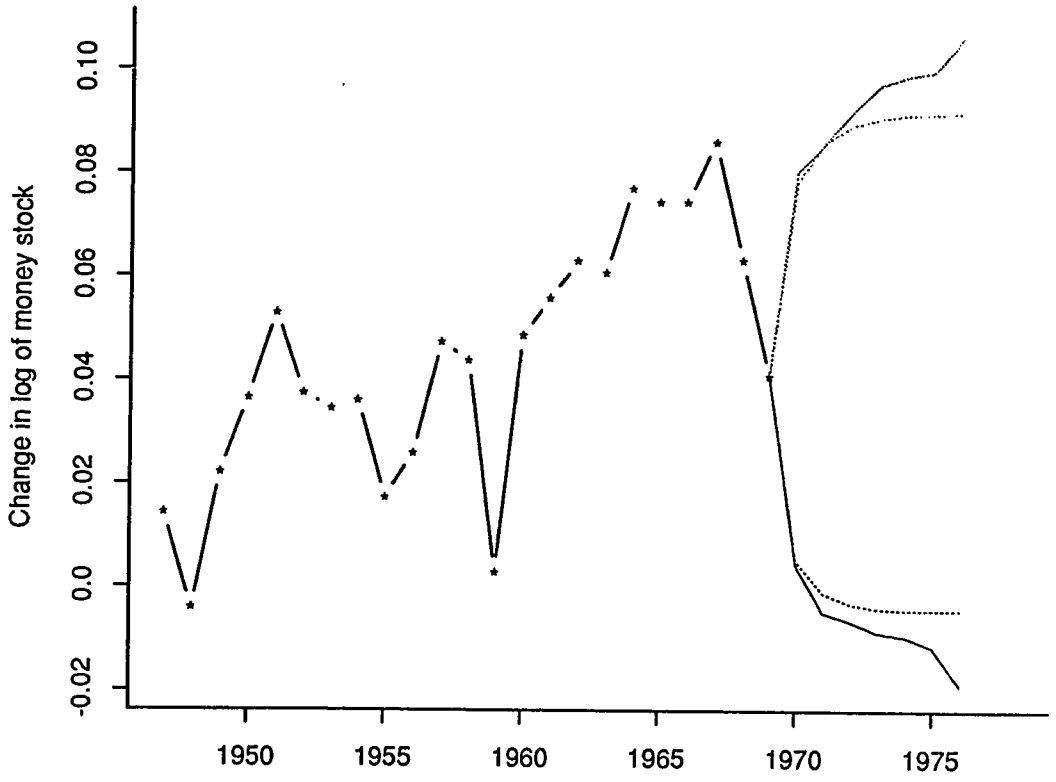


Figure 2.7: Prediction intervals for an economic time series: Shown above is the annual change in the log of money stock from 1947 to 1970 with two sets of 95% prediction intervals. The solid line is produced by the WLB, and the dotted line is based on Box-Jenkins methodology. In performing the WLB, we first simulate $AR(1)$ parameters by weighted regression. Error variances are subsequently sampled from their posterior, and then future data sequences are sampled. The solid lines represent upper and lower quantiles of 1000 simulated futures at each of 7 time points. The AR parameter is estimated to be .65 and the prediction variance is estimated at .00035. Diagnostics, including a check of the prediction residuals and a Portmanteau test 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.

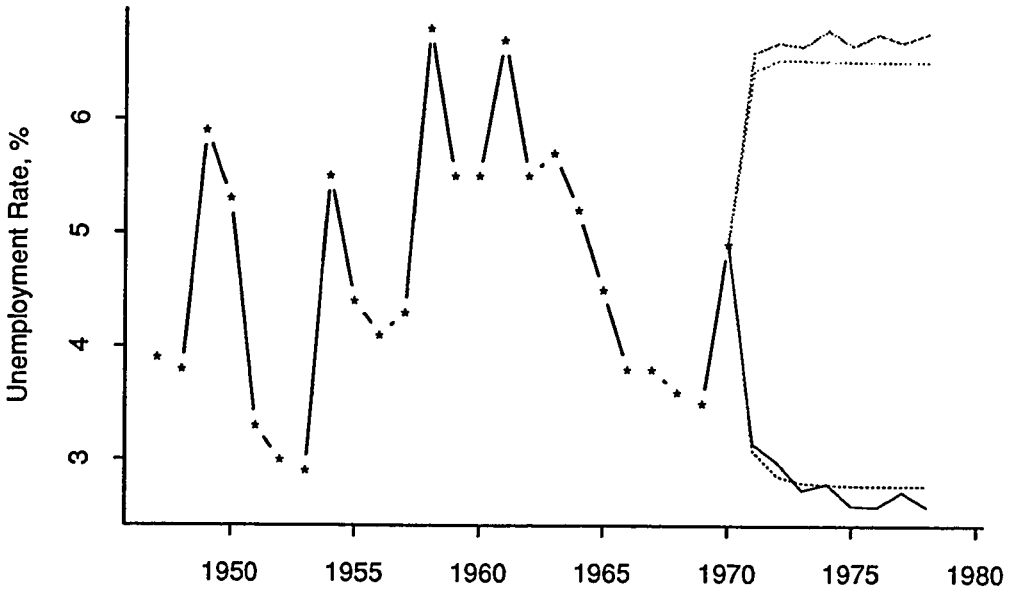


Figure 2.8: Prediction intervals for an economic time series: Shown above are the U.S. yearly average unemployment rates from 1947 to 1970. The bands after 1970 are two sets of 90% prediction intervals. The solid lines are produced by the WL bootstrap, and the dotted lines are based on Box-Jenkins methodology. In performing the WL bootstrap, 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 solid lines represent upper and lower quantiles of 1000 simulated futures at each of 8 time points. The AR parameter is estimated at .44, and the prediction variance is estimated at 1.04. Diagnostics, including a check of the prediction residuals and a Portmanteau test suggest that the $AR(1)$ model is adequate.

The Bayesian solution of the calibration problem amounts to constructing a posterior distribution for η. As for the prediction problem, this posterior distribution takes the form of a mixture:

$$P(\eta|D_c, D_p) = \int P(\eta|D_p, \theta)\, P(\theta|D_c)\, d\theta\,.$$

It turns out in this example that the conditional posterior density of η given D_p and θ has a simple analytic form (under a flat prior for η over a finite range). Therefore, we can approximate the marginal posterior for η by first sampling $\tilde{\theta}$'s from $P(\theta|D_c)$ (using the WL bootstrap) and then averaging all the $P(\eta|D_p, \tilde{\theta})$ at each point on a grid of η values. Code used to compute the maximum likelihood estimator of θ is used to compute $\tilde{\theta}$. The method is an example of that described in Section 1.5.2. Figure (2.9) summarizes these calculations.

Note that in the approximation used by Racine-Poon, the posterior for θ is assumed to be normal with mean $(897.3, -.6273, 1.378)$ and covariance matrix

$$\begin{pmatrix} 196 & .997 & -.810 \\ .997 & .013 & -.009 \\ -.810 & -.009 & .014 \end{pmatrix}.$$

Alternatively, using the WLB, the simulated $\tilde{\theta}$'s have mean $(898.6, -.6173, 1.352)$ and covariance matrix

$$\begin{pmatrix} 174 & 1.09 & -.875 \\ 1.09 & .015 & -.010 \\ -.875 & -.010 & .012 \end{pmatrix}.$$

Comparing determinants, we see that the WLB distribution of $\tilde{\theta}$ is more concentrated than the normal approximation used in Racine-Poon (1988). We do not know the right answer here, as both solutions are approximate.

2.7 Model Selection

A common problem in statistics is how to choose one sub-model from a collection comprising a single large model. The Bayesian solution to this problem is to base the choice on the posterior probability of the submodels. These posterior probabilities are

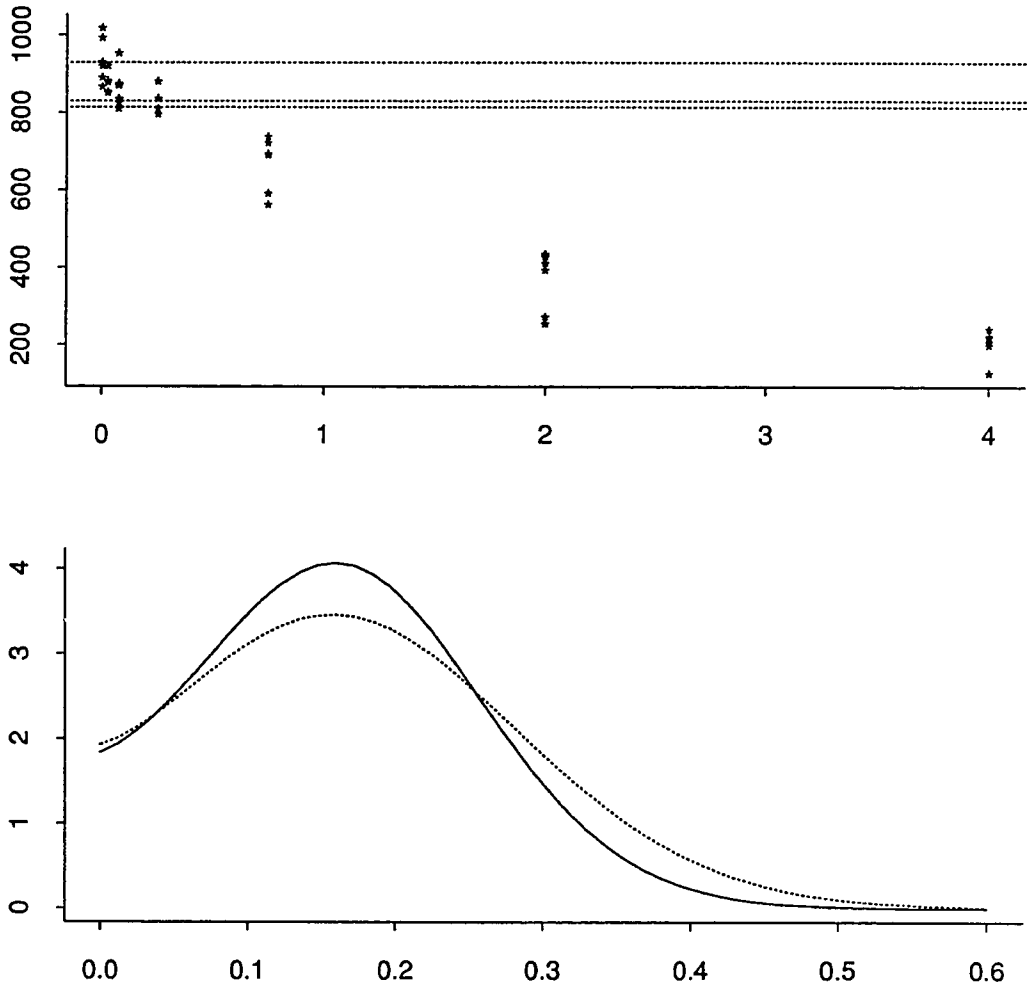


Figure 2.9: Results of a nonlinear calibration experiment on nasturtium: The upper graph summarizes the data. Calibration data are plotted as 42 pairs (y_i, x_i); x_i is the concentration of an agrochemical in the soil, and y_i is the weight in milligrams of the nasturtium plants in the i^{th} pot. The three horizontal dashed lines indicate data from the prediction experiment. Three weights are observed at an unknown concentration η. The lower graph shows two approximations of the posterior density of η. The solid line is computed using the WLB while the dashed line is computed by the method used in Racine-Poon (1988).

computed from prior probabilities and *marginal likelihoods*–the probability of the data given the sub-model. They are *marginal* because the parameter has been integrated out over the sub-model. Such marginal likelihoods are equivalent to prior predictive probabilites of the data, and may be difficult to calculate being integrals over the parameter space. The WLB can sometimes be used to approximate the marginal likelihood of a sub-model, which is written

$$p(x \mid M_j) = \prod_{i=1}^{n} p(x_i \mid x_1^{i-1}, M_j) \tag{2.3}$$

where $M_j \subset \Theta$ is the j^{th} sub-model. Each factor on the right-hand side of equation (2.3) may be evaluated using the results on prediction in Section 2.5, as follows. First, generate a sample from the predictive distribution of x_i given x_1^{i-1} under M_j, as described in Section 2.5. Then evaluate the predictive density at the value x_i only, using, for example, a kernel density estimate. The resulting predictive densities are then multiplied together to obtain the marginal likelihood (2.3). Equation (2.3) also underlies the "prequential" approach to inference of Dawid (1984).

Chapter 3

ASYMPTOTICS FOR THE WLB

3.1 Introduction

In this chapter, the asymptotic theory outlined earlier in Section 1.3 is fully developed. The theory is restricted to the iid case; that is, we consider independent random variables $X_1, X_2, \cdots, X_n, \cdots$, each a mapping

$$X_i : \Omega_{dat} \to \mathbb{R}^m \qquad m \geq 1$$

and each defined on the probability space

$$(\Omega_{dat}, \mathcal{A}_{dat}, P_{\theta_0})$$

where the common probability measure P_{θ_0} is one element of a finite-dimensional model

$$\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^K\} \qquad K \geq 1.$$

We suppose that each P_θ has a density f_θ with respect to a σ–finite measure μ on $(\Omega_{dat}, \mathcal{A}_{dat})$.

The weighted likelihood function for θ depends on the data $X_1, X_2, \cdots, X_n$ and an independent vector of weights

$$(w_{n,1}, w_{n,2}, \cdots, w_{n,n}) = \frac{n}{\sum_j Y_j}(Y_1, Y_2, \cdots, Y_n)$$

where, formally, each Y_i mapping Ω_{wt} into $\mathbb{R}^+$ is defined on the probability space

$$(\Omega_{wt}, \mathcal{A}_{wt}, P_{wt}).$$

Specifically, P_{wt} induces a Gamma distribution on $\mathbb{R}^+$. We restrict attention to uniform Dirichlet weights in this chapter, and so P_{wt} defines an exponential distribution.

Combining the data and the weights, the weighted likelihood function is

$$\tilde{L}_n(\theta) = \prod_{i=1}^{n} [f_\theta(X_i)]^{w_{n,i}} .$$

In the last two chapters, we examined the weighted likelihood function for particular models, comparing the conditional distribution (given the data) of the parameter $\tilde{\theta}$ maximizing $\tilde{L}_n(\cdot)$ to a posterior density of θ. For a given data set, the exact conditional distribution of $\tilde{\theta}$ can be approximated arbitrarily well by simulation. However, general optimality results can be stated only by appealing to asymptotic conditional distributions.

Fortunately, theoretical ground has already been broken by researchers studying classical likelihood theory. Wheras the theory of weighted likelihood describes the stochastic properties of the weighted likelihood function given the data, much of classical likelihood theory is concerned with the stochastic properties of the ordinary likelihood function before the data are observed. The similarities between the objects under study in the two theories imply that the classical theory provides a template for my work on weighted likelihood. The two theories are distinct in the sense that results from one do not imply results from the other, but the methods of proof tend to be parallel.

3.2 Preliminaries

3.2.1 The nature of conditional probabilities

There are two basic probability spaces at work in bootstrapping problems: one governing the data and one governing the weights. The WLB uses Dirichlet distributed weights, although Efron's original bootstrap is based on multinomial weights. In fact, many weight distributions are possible (see Haeusler, Mason, and Newton, 1991). In studying the asymptotic properties of the WLB, it makes sense to view the first probability space as an infinite product space

$$(\Omega_1, \mathcal{A}_1, P_1) := \left(\Omega_{dat}^{\infty}, \mathcal{A}_{dat}^{\infty}, P_{\theta_0}^{\infty}\right)$$

and similarly the second probability space

$$(\Omega_2, \mathcal{A}_2, P_2) := (\Omega_{wt}^{\infty}, \mathcal{A}_{wt}^{\infty}, P_{wt}^{\infty}) .$$

A single point $\omega_1 \in \Omega_1$ determines an infinite sequence of data, and a single $\omega_2 \in \Omega_2$ determines an infinite triangular array of real weights. Although each X_i is originally defined as a mapping from the margin Ω_{dat}, we may also view it as a coordinate mapping from the product space Ω_1, and similarly for Y_i. When considering various convergence statements, it is useful to have the data and weights all defined on the same probability space. Having data independent of weights, this single probability space is the product measurable space

$$(\Omega, \mathcal{A}) = (\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2)$$

endowed with the product measure $P = P_1 \times P_2$. Viewed this way, each X_i or Y_i is a function of $\omega = (\omega_1, \omega_2) \in \Omega$. A single $w \in \Omega$ determines a realization of data and weights.

The very essence of bootstrapping lies in the consideration of conditional probabilities given data. Formally, such a conditional probability is, for each $A \in \mathcal{A}$, a function

$$P(A|\sigma(X_1, X_2, \cdots, X_n))(\omega) \qquad \omega \in \Omega$$

which is measurable with respect to

$$\sigma(X_1, X_2, \cdots, X_n) \subset \mathcal{A}$$

and which satisfies the relation

$$\int_B P(A|\sigma(X_1, X_2, \cdots, X_n))\, dP = P(A \cap B) \qquad \text{for all } B \in \sigma(X_1, \cdots, X_n).$$

As defined, this conditional probability is a function of $\omega = (\omega_1, \omega_2)$ and indeed may not be a probability distribution on $(\Omega, \mathcal{A})$ for a given fixed ω. Functions satisfying this definition but differing on a set of P measure zero are called versions of the conditional probability. The following lemma sures up our intuition about conditional probabilities for bootstrapping.

Lemma 1 (of the unconscious bootstrapper) *There exists on $(\Omega, \mathcal{A})$ a regular conditional distribution given $\sigma(X_1, \cdots, X_n)$ which depends only on the data. That is, there exists a function $P(A|x_1, x_2, \cdots, x_n)$ which is a probability measure on $(\Omega, \mathcal{A})$ for each $(x_1, \cdots, x_n) = (X_1(\omega), \cdots, X_n(\omega))$ and which is a version of $P(A|\sigma(X_1, \cdots, X_n))$ for each fixed $A \in \mathcal{A}$.*

PROOF. The result is a consequence of well known results from measure theory. For instance, Theorem 1.5, page 603 of Doob (1953), or Theorem 20.1 of Billingsley (1986), or Theorem 9.1.2 of Chung (1974) imply that $P(A|\sigma(X_1, \cdots, X_n))$ must equal a measurable function of $X_1, \cdots, X_n$. That this conditional probability is regular is also a standard, but nontrivial result–see Theorem 2 page 217 of Chow and Teicher (1988) for instance. □

This lemma is useful because it allows us to view conditional probabilities (given data) involving $\tilde{L}_n$ as random variables on $(\Omega_1, \mathcal{A}_1, P_1)$. It seems, moreover, that this result is taken for granted in statistical discussions of bootstrapping–hence the title. Throughout this chapter, $P(\cdot|X_1^n)$ is the notation used for this conditional distribution.

3.2.2 Conditional convergence

Let random variables (or vectors) $U, V_1, V_2, V_3, \cdots$, be defined on the product space $(\Omega, \mathcal{A})$. We say V_n converges in conditional probability $a.s.[P_1]$ to U if for all $\epsilon > 0$

$$P\left(\| V_n - U \| > \epsilon \mid X_1^n\right) \to 0 \qquad a.s.[P_1] \tag{3.1}$$

as $n \to \infty$. The distance $\| \cdot \|$ is ordinary Euclidean distance. The notation $a.s.[P_1]$ is read *almost surely under* P_1 and means for P_1 almost every infinite sample sequence $X_1, X_2, \cdots$. For brevity, this convergence is denoted

$$V_n \to_{c.p.} U \qquad a.s.[P_1].$$

The following convergence properties are easily verified.

Lemma 2 *Consider two sequences* $\{Z_n\}$, *and* $\{U_n\}$ *and two other random variables* Z, U *all defined on the product space* $(\Omega, \mathcal{A})$. *If*

$$Z_n \to_{c.p.} Z \quad a.s.[P_1] \quad \text{and} \quad U_n \to_{c.p.} U \quad a.s.[P_1]$$

then

$$Z_n U_n \to_{c.p.} ZU \quad a.s.[P_1] \quad \text{and} \quad Z_n + U_n \to_{c.p.} Z + U \quad a.s.[P_1].$$

PROOF. For each fixed infinite sequence of data, the results follow from properties of convergence in probability. □

The following conditional weak law is useful.

Lemma 3 *Let the data X_i and the exponentials Y_i be defined as before. If g is a real-valued, measurable function such that $E_{\theta_0}|g(X_i)| < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i g(X_i) \rightarrow_{c.p.} E_{\theta_0} g(X_i) \quad a.s.[P_1].$$

PROOF. For $[P_1]$-almost every sequence $\{z_i = g(X_i(\omega_1))\}$, the strong law of large numbers gives both

$$\frac{1}{n}\sum_{i=1}^{n} z_i \rightarrow E_{\theta_0}\, g(X_i) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} |z_i| \rightarrow E_{\theta_0}\, |g(X_i)| < \infty$$

and Lemma 14 of Appendix A gives

$$\frac{1}{n} \max_{1 \le i \le n} |z_i| \rightarrow 0.$$

The conditions for Theorem 13 of Appendix A are satisfied with $a_{n,i} = z_i$. □

In fact, the Y_i can be any iid sequence with mean 1 and finite variance σ^2 and the result still holds.

3.3 Conditional Consistency

Consistency of an estimator has a natural analog in a theory of weighted likelihood. Suppose that $\{\bar{\theta}_n\}$ is a sequence of random variables on the product space $(\Omega, \mathcal{A})$. It is called *conditionally consistent* if

$$\bar{\theta}_n \rightarrow_{c.p.} \theta_0 \qquad a.s.[P_1].$$

Several facts are noteworthy. If $\{\hat{\theta}_n\}$ is any strongly consistent estimator of θ_0, then it is conditionally consistent in the trivial sense that the conditional probability in (3.1) equals 0 for large enough n. (If $n > N(\epsilon, \omega_1)$, then $|\hat{\theta}_n - \theta_0| > \epsilon$ is an impossible event.) On the other hand, a conditionally consistent sequence is generally not

consistent in the usual sense because it is governed by a different probability measure than the one generating the data. Other forms of conditional consistency are suggested by this definition. For instance, one could have convergence in $[P_1]$ probability replacing $a.s.[P_1]$. For either of these, there is an analogue of strong consistency: conditional almost sure convergence, either almost surely or in probability with respect to $[P_1]$. The definition above is sufficient for the results of this chapter, although a stronger sense of conditional consistency is used in Chapter 4. To study conditional consistency of the $\tilde{\theta}_n$, we first need some regularity conditions on the model.

Regularity Conditions 1

C_1 *Identifiability: For any* $\theta_1 \neq \theta_2$ *both in* Θ, *there exists a set* $A \in \mathcal{A}_{dat}$ *such that*

$$P_{\theta_1}(A) \neq P_{\theta_2}(A)\,.$$

C_2 *For the* true *density* f_{θ_0} *and another density* f_{θ_1} *in the model,*

$$\phi = \int f_{\theta_0}(x) \log\left(\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}\right) d\mu(x) > -\infty\,.$$

The next theorem asserts that conditionally upon the data, the weighted likelihood function tends to achieve its maximum at the truth. It is kind of a conditional consistency result about the entire weighted likelihood function.

Theorem 4 *Let* θ_1 *be a point in* Θ *not equal to* θ_0. *Let* $E_n \in \mathcal{A}_2$ *be defined by*

$$E_n(\omega_1) = \left\{\omega_2 \in \Omega_2 \;\; : \tilde{L}_n(\theta_0) > \tilde{L}_n(\theta_1)\right\}$$

for $\omega_1 \in \Omega_1$. *Under conditions* C_1 *and* C_2, *as* $n \to \infty$

$$P(\; E_n \mid X_1^n \;) \to 1 \;\; a.s.[P_1]\,.$$

PROOF. Note that $E_n(\omega_1) \in \mathcal{A}_2$ because the weighted likelihood is a measurable function. Also, $\tilde{L}_n(\theta_0) > 0$ with P_1 probability one (for any fixed set of weights), and so

$$E_n(\omega_1) = \left\{\omega_2 \in \Omega_2 \;\; : \sum_{i=1}^{n} w_{n,i} \log(f_{\theta_0}(X_i)) > \sum_{i=1}^{n} w_{n,i} \log(f_{\theta_1}(X_i))\right\} \qquad (3.2)$$

where the right hand side may equal $-\infty$, but the left hand side is bigger than $-\infty$ with P_1 probability 1. Thus, $a.s.[P_1]$,

$$E_n(\omega_1) = \left\{ \omega_2 \in \Omega_2 \ : \frac{1}{n} \sum_{i=1}^{n} Y_i V_i < 0 \right\}$$

where Y_i are the exponential variables determing the weights and

$$V_i := \log \left(\frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)} \right)$$

when this is well defined, and an arbitrary number otherwise.

By Jensen's inequality,

$$\phi = E_{\theta_0} V_i \leq \log E_{\theta_0} \frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)} = \log 1 = 0$$

where ϕ is as defined in condition C_2 and is a finite negative number. The identifiability condition C_1 insures that the above inequality is strict. Equality can happen only if $f_{\theta_1}(X_i) = f_{\theta_0}(X_i)$ $a.s.[P_1]$. By the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} V_i \to \phi \qquad a.s.[P_1]$$

and so by Lemma 3,

$$\frac{1}{n} \sum_{i=1}^{n} Y_i V_i \to_{c.p.} \phi < 0 \qquad a.s.[P_1] \, .$$

Since this average is converging in the above sense to a negative number, the conditional probability of the event E_n must go to 1 $a.s.[P_1]$. □

For a relatively small class of models, the following conditional consistency result is immediate from Theorem 4.

Corollary 1 *Under conditions C_1 and C_2, if Θ is finite, then $\{\tilde{\theta}_n\}$ exists, is conditionally consistent, and is unique in the following sense. If $\{\bar{\theta}_n\}$ also maximizes the weighted likelihood, then*

$$P(\, \tilde{\theta}_n = \bar{\theta}_n \mid X_1^n \,) \to 1 \quad a.s.[P_1] \, .$$

PROOF. Existence is clear because we can always find at least one $l_j = \tilde{L}_n(\theta_j)$ which is at least as large as all the other l_i's in a finite set. To prove consistency, let $\{\tilde{\theta}_n\}$ be a sequence of maximizers of $\{\tilde{L}_n\}$. For each n

$$\begin{aligned} P\left(\tilde{\theta}_n = \theta_0 \mid X_1^n\right) &= P\left(\tilde{L}_n(\theta_0) \geq \tilde{L}_n(\theta_k) \quad \forall \theta_k \neq \theta_0 \mid X_1^n\right) \\ &\geq P\left(\tilde{L}_n(\theta_0) > \tilde{L}_n(\theta_k) \quad \forall \theta_k \neq \theta_0 \mid X_1^n\right) \\ &= P\left(\bigcap_k [\tilde{L}_n(\theta_0) > \tilde{L}_n(\theta_k)] \mid X_1^n\right). \end{aligned}$$

The conditional probability of this last intersection must converge to 1 $a.s.[P_1]$ because, by Theorem 4, the conditional probability of each of the component events converges to 1 in this sense. Uniqueness in conditional probability follows immediately. □

More conditions are required to extend the conditional consistency result to a larger class of models. Taking a cue from classical theory, we note that it is precisely here in the parallel theory of consistency of MLE's that methods of proof diverge along two distinct paths. The path opened by Wald (1949) and followed notably by Keifer and Wolfowitz (1956) and Perlman (1972), assumes compactness of Θ. Using the Heine-Borel Theorem, the problem is brought down to considering finite Θ, and hence the importance of Theorem 4. Other strong assumptions are also required. A second and completely different line of proof was given by Cramér (1946) for $\Theta \subset \mathbb{R}$, (see Lehmann 1983 for the general result). By assuming certain smoothness properties of the model, Cramér used a Taylor expansion of the log likelihood to establish consistency of roots of the likelihood equation. This method is mathematically less demanding, but suffers from the weakness of its conclusions. Loosely, it says that a consistent sequence of roots of the likelihood equation exists. If the likelihood is differentiable and attains its maximum in the interior of Θ, then the MLE $\hat{\theta}_n$ is indeed a root of the likelihood equation. However, Cramér's theorem says nothing about the MLE itself.

These considerations notwithstanding, we follow the Cramér method of proof to establish a conditionally consistent sequence of roots of the weighted likelihood equation. For the general case of $\Theta \in \mathbb{R}^K$, we extend a proof of Foutz (1977) to get stronger conclusions.

Regularity Conditions 2

C_3 $\Theta \subset \mathbf{R}^K$ *contains an open ball* B *containing* θ_0.

C_4 *For each* $\theta \in B$, *all the first partial derivatives of* $\log f_\theta(x)$ *with respect to the components of* θ *exist and are continuous for almost all* x.

Note that if $\tilde{\theta}$ is not on the boundary of Θ, and if the log weighted likelihood is differentiable (condition C_4), then $\tilde{\theta}$ must be a root of the *weighted likelihood equations*

$$\frac{\partial \log \tilde{L}_n}{\partial \theta_k}(\theta) = 0 \quad \text{for each} \quad k \tag{3.3}$$

As the next theorem establishes in the one parameter case, a conditionally consistent sequence of roots of equations (3.3) exists.

Theorem 5 *If* $K = 1$ *and conditions* C_1, C_2, C_3, C_4 *hold on the family* $\mathcal{P}_\Theta$, *then there exists a sequence* $\{\check{\theta}_n\}$ *such that*

$$P\left(\frac{\partial \log \tilde{L}_n}{\partial \theta}(\check{\theta}) = 0 \,\middle|\, X_1^n \right) \to 1 \quad a.s.[P_1]$$

and moreover, this sequence of roots of the weighted likelihood equation is conditionally consistent.

PROOF. Choose $\epsilon > 0$ small enough so that both $\theta_1 = \theta_0 - \epsilon/2$ and $\theta_2 = \theta_0 + \epsilon/2$ are contained in B (This is possible by condition C_3). Now apply Theorem 4 twice, using θ_1 and θ_2 as the points in Θ not equal to θ_0. Let $N \subset \Omega_1$ denote the null set where the convergence in Theorem 4 fails to happen upon these two applications. Fix $\omega_1 \in \Omega_1 \setminus N$, and consider a sequence of events E_n each in $\mathcal{A}_2$ defined by

$$E_n(\omega_1) = \{\omega_2 \in \Omega_2 \,:\, \tilde{l}_n(\theta_0) > \tilde{l}_n(\theta_0 - \epsilon/2) \text{ and } \tilde{l}_n(\theta_0) > \tilde{l}_n(\theta_0 + \epsilon/2)\}$$

where $\tilde{l}_n(\cdot)$ is the natural logarithm of the weighted likelihood $\tilde{L}_n(\cdot)$. Furthermore, consider two sequences defined by

$$A_n(\omega_1) = \{\omega_2 \in \Omega_2 \,:\, \tilde{l}_n(\theta_0) > \tilde{l}_n(\theta_1)\}$$

and

$$B_n(\omega_1) = \{\omega_2 \in \Omega_2 \ : \tilde{l}_n(\theta_0) > \tilde{l}_n(\theta_2)\}$$

By Theorem 4,

$$P(\, A_n(\omega_1) \,|\, X_1^n \,)(\omega_1) \to 1 \ \text{ and } \ P(\, B_n(\omega_1) \,|\, X_1^n \,)(\omega_1) \to 1$$

as $n \to \infty$. Therefore, $P(\, E_n \,|\, X_1^n \,)$ also converges $a.s.[P_1]$ to 1 as $n \to \infty$.

Take a point $\omega_2 \in E_n(\omega_1)$ This corresponds to one possible set of weights in the weighted likelihood function. Without loss of generallity, suppose that $\tilde{l}_n(\theta_1) \geq \tilde{l}_n(\theta_2)$ and consider a point θ^* defined by

$$\theta^* = \inf \left\{ \gamma > \theta_0 \ : \ \tilde{l}_n(\gamma) = \tilde{l}_n(\theta_1) \right\} .$$

Continuity of $\tilde{l}_n$ and the fact that $\tilde{l}_n(\theta_0) > \tilde{l}_n(\theta_1)$ insures that θ^* exists and satisfies

$$\theta_0 < \theta^* \leq \theta_2 \, .$$

Now by the Mean Value Theorem, there exists at least one point in (θ_1, θ^*) where the derivative of the weighted log likelihood equals 0. All such points are roots of the weighted likelihood equation. Consider the nonempty set of such critical points

$$\Gamma = \{ \, \gamma \in (\theta_1, \theta^*) \ : \ \tilde{l}_n(\gamma) = 0 \, \} \ .$$

Since $\tilde{l}_n(\theta_0) > \tilde{l}_n(\theta_1)$, at least one member of Γ corresponds to a local maximum of $\tilde{l}_n$. Letting Γ_{max} be the subset of Γ corresponding to local maxima, define $\check{\theta}_n$ to be the element of Γ_{max} closest to θ_0. See Figure 3.1.

On the set $E_n(\omega_1)$ of conditional probability converging to 1 $a.s.[P_1]$, a root of the weighted likelihood equation (corresponding to local maxima and not depending on ϵ) exists and is within ϵ of the true θ_0. □

Although a conditionally consistent sequence $\{\check{\theta}_n\}$ has been found, it is not necessarily equal to the maximizing sequence $\{\tilde{\theta}_n\}$ for the same reasons as those plaguing theory about roots of the likelihood equation. In fact this sequence may not even be computable because, as defined, each $\check{\theta}_n$ is the local maximum *closest* to the unknown θ_0. The exact same problem comes up in Cramér's proof of consistency of the

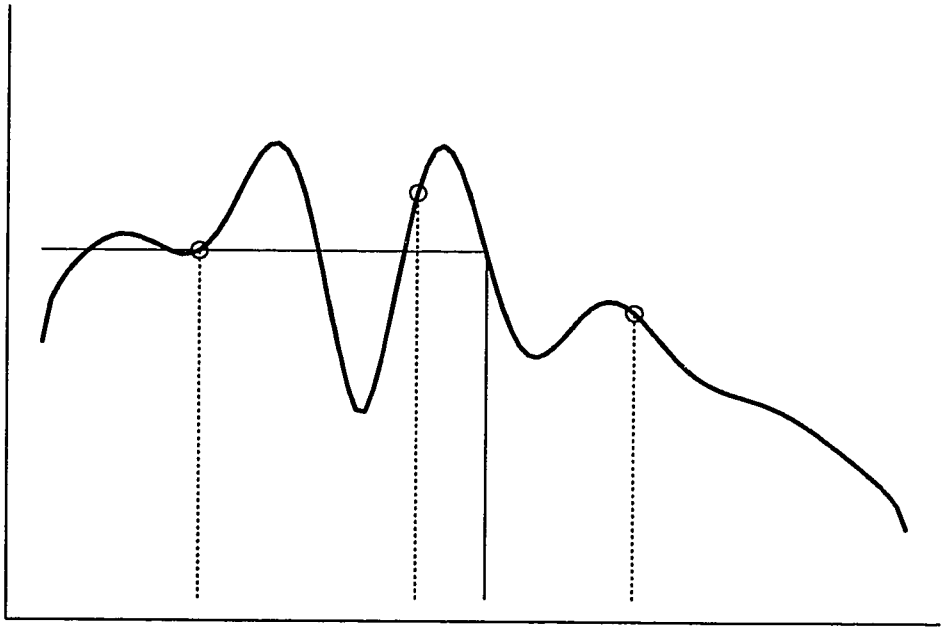


Figure 3.1: A schematic proof of Theorem 5: The curve is the log-weighted likelihood function from a point in $E_n(\omega_1)$. The three dashed lines rise from θ_1, θ_0, and θ_2. Γ_{max} contains two roots.

MLE (Cramér 1946), and it is propagated by Lehmann (1983). The regularity conditions imposed so far do not eliminate the possibility of multiple roots of the weighted likelihood equation. To show that a consistent sequence is essentially unique, more assumptions are needed. Huzurbazar (1948) and Foutz (1977) give such proofs for the maximum likelihood estimator. Foutz's proof, which includes a proof of consistency, is quite elegant and is not restricted to one parameter models. As the simple geometric argument used to prove Theorem 5 does not extend easily to higher dimensions and does not show uniqueness, the method of Foutz is used to get a stronger result. Before stating this conditional consistency result for general K, we study properties of the weighted score and information functions.

3.4 Weighted score and information

To discover more properties of the weighted likelihood, we study the analogs of the score and information functions of classical theory. First we need some notation. Let $\psi_i(\theta)$ be the k vector of first partial derivatives of $\log f_\theta(X_i)$ (all derivatives are with respect to the components of θ). Also let $\psi_i'(\theta)$ be the corresponding matrix of second partials (when they exist). The score and information functions of classical theory are

$$S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\psi_i(\theta) \quad \text{and} \quad J_n(\theta) = \frac{-1}{n}\sum_{i=1}^{n}\psi_i'(\theta)$$

respectively. In parallel with these functions, we define a *weighted score function*

$$\tilde{S}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_{n,i}\psi_i(\theta)$$

and a *weighted information matrix*

$$\tilde{J}_n(\theta) = \frac{-1}{n}\sum_{i=1}^{n} w_{n,i}\psi_i'(\theta)$$

These are, of course, the first and second derivatives of the log weighted likelihood.

Next we define the Fisher information $I(\theta)$, and a slight variant denoted $J(\theta)$

$$[I(\theta)]_{jk} = E_\theta\left[\left(\frac{\partial \log f_\theta(X)}{\partial \theta_j}\right)\left(\frac{\partial \log f_\theta(X)}{\partial \theta_k}\right)\right]$$

and

$$[J(\theta)]_{jk} = -E_{\theta_0}\left(\frac{\partial^2 \log f_\theta(X)}{\partial\theta_j \partial\theta_k}\right).$$

(Throughout, elements of matrices or vectors are described with square brackets and subscripts as above.) The key distinguishing fact about $J(\theta)$ is that the expectation involved is always with respect to the P_{θ_0}.

To study the weighted score and information functions, we require some smoothness conditions on the model.

Regularity Conditions 3

C_5 *All second partial derivatives of* $\log f_\theta(x)$ *with respect to the components of* θ *exist and are continuous for* $\theta \in B$ *at almost all* x.

C_6 *For* $1 \le j,k,l \le K$ *there exists functions* $g_j(x), h_{jk}(x)$ *and* $H_{jk}(x)$ *(possibly depending on* θ_0 *and* B*) such that for all* θ *in* B*, the following relations hold for almost all* x*:*

$$a)\quad \left|\frac{\partial f_\theta(x)}{\partial\theta_j}\right| \le g_j(x) \quad \text{with} \quad \int g_j(x)\,\mu(dx) < \infty,$$

$$b)\quad \left|\frac{\partial^2 f_\theta(x)}{\partial\theta_j\partial\theta_k}\right| \le h_{jk}(x) \quad \text{with} \quad \int h_{jk}(x)\,\mu(dx) < \infty,$$

$$c)\quad \left|\frac{\partial^2 \log f_\theta(x)}{\partial\theta_j\partial\theta_k}\right| \le H_{jk}(x) \quad \text{with} \quad E_{\theta_0}H(X) < \infty,$$

C_7 *The Fisher information* $I(\theta)$ *is positive definite for* $\theta \in B$*, and all its elements are finite.*

The set B in conditions C_5 and C_6 is the same one of conditions C_3 and C_4. There are several important consequences of the above conditions which restrict the behaviour of second derivatives of the density and log density. For example, using the conditions $C_5, C_6(a)$ and $C_6(b)$, it is easy to show that for all $1 \le j,k \le K$

$$[I(\theta)]_{jk} = -E_\theta\left(\frac{\partial^2 \log f_\theta(X)}{\partial\theta_j\partial\theta_k}\right)$$

and

$$E_\theta\left(\frac{\partial \log f_\theta(X)}{\partial \theta_j}\right) = 0\,.$$

Furthermore, continuity of $J(\theta)$ in B follows from conditions C_5 and $C_6(c)$. (Use Theorem 16.8 in Billingsley (1986) for example to prove these results.) The expression above for the Fisher information makes it look quite similar to the definition of $J(\theta)$ in the previous section. Note that the expectation defining $J(\theta)$ is always with respect to θ_0. Therefore under the above conditions, the two functions $I(\theta)$ and $J(\theta)$ coincide at $\theta = \theta_0$ but they are different in general.

With these fairly standard conditions imposed on the model, we can prove some further properties of the weighted likelihood function. First, a property of the weighted score function;

Lemma 4 *If conditions C_1, C_2, C_3, C_4, and $C_6(a)$ hold in the family $\mathcal{P}_\Theta$, then as $n \to \infty$*

$$\| \tilde{S}_n(\theta_0) \| \to_{c.p.} 0 \quad a.s.[P_1]\,.$$

PROOF. By definition,

$$\begin{aligned}
\| \tilde{S}_n(\theta_0) \|^2 &= \| \frac{1}{n}\sum_{i=1}^n w_{n,i}\psi_i(\theta_0) \|^2 \\
&= \frac{1}{\bar{Y}_n^2} \| \frac{1}{n}\sum_{i=1}^n Y_i\psi_i(\theta_0) \|^2, \qquad \bar{Y}_n = n^{-1}\sum_{j=1}^n Y_j \\
&= \frac{1}{\bar{Y}_n^2}\sum_{k=1}^K \left(\frac{1}{n}\sum_{i=1}^n Y_i[\psi_i(\theta_0)]_k\right)^2 .
\end{aligned}$$

The strong law of large numbers gives $\bar{Y}_n \to 1$ with P_2 probability 1, which implies $\bar{Y}_n \to_{c.p.} 1$, $a.s.[P_1]$. Therefore by Lemma 2, it suffices to prove that for any k, $1 \le k \le K$,

$$\frac{1}{n}\sum_{i=1}^n Y_i[\psi_i(\theta_0)]_k \to_{c.p.} 0 \quad a.s.[P_1]\,. \tag{3.4}$$

By condition $C_6(a)$, both $E_{\theta_0}|[\psi_i(\theta_0)]_k| < \infty$ and $E_{\theta_0}[\psi_i(\theta_0)]_k = 0$ (see the discussion after the third set of regularity conditions). Therefore, Lemma 3 can be applied with $g(X_i) = [\psi_i(\theta_0)]_k$ proving (3.4). □

Properties of the weighted information function $\tilde{J}_n(\theta)$ are given in the next three lemmas.

Lemma 5 *If conditions C_1, C_2, C_3, C_4, C_5 and $C_6(c)$ hold on the model $\mathcal{P}_\Theta$, then as $n \to \infty$*

$$\| \tilde{J}_n(\theta) - J(\theta) \| \to_{c.p.} 0 \qquad a.s.[P_1]\,. \tag{3.5}$$

Moreover, if the sample information $J_n(\theta)$ converges $a.s.[P_1]$ to $J(\theta)$ uniformly in $\theta \in B$, then the convergence in (3.5) is also uniform in θ for $\theta \in B$.

PROOF. To prove (3.5), note that by Lemma 15 it suffices to show that for all u, v,

$$P(\, |[\tilde{J}_n(\theta)]_{uv} - [J(\theta)]_{uv}| < \epsilon \mid X_1^n \,) \to 1 \quad a.s.[P_1] \tag{3.6}$$

That is, we must show that each component of $\tilde{J}_n(\theta)$ converges to the corresponding component of $J(\theta)$ in the above sense. To proceed, recall that the $(u,v)^{th}$ element of $\tilde{J}_n(\theta)$ is

$$[\tilde{J}_n(\theta)]_{uv} = \frac{-1}{n\bar{Y}_n} \sum_{i=1}^{n} Y_i [\psi_i'(\theta)]_{uv} \quad . \tag{3.7}$$

Also, the corresponding element of $J(\theta)$ is $-E_{\theta_0}\,[\psi_i'(\theta)]_{uv}$. Condition $C_6(c)$ insures that $E|[\psi_i'(\theta)]_{uv}| < \infty$ and so Lemma 3 can be applied with $g(X_i) := [\psi_i'(\theta)]_{uv}$ giving

$$\frac{1}{n} \sum_{i=1}^{n} Y_i [\psi_i'(\theta)]_{uv} \to_{c.p.} E_{\theta_0}\,[\psi_i'(\theta)]_{uv} \quad a.s.[P_1]\,. \tag{3.8}$$

Since the factor $\bar{Y}_n$ in (3.7) converges to $a.s.[P_2]$ by the strong law, it cannot affect the limiting value of $[\tilde{J}_n(\theta)]_{uv}$ and so the proof of (3.5) is complete.

To establish uniform convergence in (3.5), we need only prove uniform convergence in (3.8). For any $\epsilon > 0$ and letting $s_i(\theta) := -[\psi_i'(\theta)]_{uv}$, we have by Markov's inequality

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i s_i(\theta) - E_{\theta_0}\, s_i(\theta)\right| > \epsilon \mid X_1^n\right) \le \frac{1}{\epsilon^2} E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i s_i(\theta) - [J(\theta)]_{uv} \mid X_1^n\right)^2$$
$$= \frac{1}{\epsilon^2}\left([J(\theta)]_{uv} - [J_n(\theta)]_{uv}\right)^2 + \frac{1}{n^2\epsilon^2}\sum_{i=1}^{n} s_i(\theta)^2\,.$$

The first term on the right converges uniformly in θ to 0 by assumption, $a.s.[P_1]$. Considering the second term, we have by condition $C_6(c)$ that

$$\frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}(s_i(\theta))^2 \leq \frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}(H_{uv}(X_i))^2 \qquad \text{for all } \theta \in B$$
$$\leq \left(\frac{1}{n\epsilon^2}\sum_{i=1}^{n}H_{uv}(X_i)\right) \times \left(\frac{1}{n}\max_{1\leq i\leq n}H_{uv}(X_i)\right).$$

The strong law ensures that the first factor above stays finite $a.s.[P_1]$, and the second factor converges $a.s.[P_1]$ to 0 by Lemma 14 of Appendix A. □

The uniform convergence in the lemma above is crucial in the proof of conditional consistency discussed in Section 3.5. If the sample information functions are members of an *equicontinuous* family, then uniform convergence follows from pointwise convergence. (See for example Stromberg, 1981, Theorem 3.143)

Before proving the next lemma, a final set of regularity conditions is needed to control the size of remainder terms in Taylor expansions. Not all of these conditions are needed for conditional consistency, but combined with the first 7, they are sufficient for conditional asymptotic normality (see the Section 3.6).

Regularity Conditions 4

C_8 *All the third partial derivatives of* $\log f_\theta(x)$ *with respect to the components of* θ *exist and are continuous for* $\theta \in B$.

C_9 *For* $1 \leq j,k,l \leq K$ *there exists a function* $F_{jkl}(x)$ *such that for all* $\theta \in B$

$$\left|\frac{\partial^3 \log f_\theta(x)}{\partial\theta_j\,\partial\theta_k\,\partial\theta_l}\right| \leq F_{jkl}(x) \quad \text{with} \quad E_{\theta_0}(F_{jkl}(X)) < \infty$$

C_{10} *For* $1 \leq j,k,l \leq K$ *there exists a function* $G_{jkl}(x)$ *such that for all* $\theta \in B$

$$\left|\frac{\partial \log f_\theta(x)}{\partial\theta_j}\cdot\frac{\partial^2 \log f_\theta(x)}{\partial\theta_k\,\partial\theta_l}\right| \leq G_{jkl}(x) \quad \text{with} \quad E_{\theta_0}(G_{jkl}(X)) < \infty$$

The next lemma is similar to Lemma 5 above except that the point at which the weighted information is being evaluated may change with n. Not surprisingly, slightly stronger smoothness conditions are required.

Lemma 6 *If $\hat{\theta}_n$ is a strongly consistent estimator of θ_0 and if conditions C_1 to C_9 hold in the model $\mathcal{P}_\Theta$, then for all $\epsilon > 0$ as $n \to \infty$*

$$P(\| \tilde{J}_n(\hat{\theta}_n) - J(\theta_0) \| > \epsilon \,|\, X_1^n) \to 0 \quad a.s.[P_1]\,. \tag{3.9}$$

PROOF. Let $R = \tilde{J}_n(\hat{\theta}_n) - I(\theta_0)$. By Lemma 15, it suffices to show that for each $1 \le j, k \le K$,

$$P(|[R]_{jk}| > \epsilon \,|\, X_1^n) \to 0 \quad a.s.[P_1]\,. \tag{3.10}$$

By definition, the components of R have the form

$$[R]_{jk} = -\frac{1}{n\bar{Y}_n} \sum_{i=1}^{n} Y_i [\psi_i'(\hat{\theta}_n)]_{jk} + E\,[\psi_i'(\theta_0)]_{jk}$$

where Y_i are iid unit exponential random variables. First note that the factor $\bar{Y}_n$ does not affect the asymptotic value of R because it is converging $a.s.[P_2]$ to 1. If as $n \to \infty$, $a.s.[P_1]$,

$$\frac{\sum_{i=1}^n |[\psi_i'(\hat{\theta}_n)]_{jk}|}{n} \to c \quad 0 < c < \infty, \tag{3.11}$$

$$\frac{\sum_{i=1}^n [\psi_i'(\hat{\theta}_n)]_{jk}}{n} \to -[I(\theta_0)]_{jk}\,, \tag{3.12}$$

and

$$\frac{1}{n} \max_{1 \le i \le n} |[\psi_i'(\hat{\theta}_n)]_{jk}| \to 0 \tag{3.13}$$

then (3.10) is a consequence of Theorem 13. It suffices, therefore, to prove (3.11), (3.12), and (3.13). We use the following Taylor expansion of $[\psi_i'(\theta)]_{jk}$ about θ_0 evaluated at $\hat{\theta}_n$ to argue each of (3.11), (3.12), and (3.13). For ease of notation, put $g_i(\theta) = [\psi_i'(\theta)]_{jk}$.

$$g_i(\hat{\theta}_n) = g_i(\theta_0) + g_i'(\hat{\theta}_n, \theta_0)(\hat{\theta}_n - \theta_0) \tag{3.14}$$

where $g_i'(\hat{\theta}_n, \theta_0)$ is the row vector of partial derivatives of $g_i(\theta)$ evaluated at some point θ_i^* on the line between $\hat{\theta}_n$ and θ_0. This expansion is well defined by differentiability conditions C_4, C_5, and C_8.

Considering only points $\omega_1 \in \Omega_1$ where $\hat{\theta}_n$ converges to θ_0, We can find $N(\omega_1)$ such that for all $n > N(\omega_1)$, $\hat{\theta}_n(\omega_1)$ is close enough to θ_0 (say each component is within δ) so that each third order partial derivative of the log likelihood at θ^* can be bounded by a finite mean random variable (condition C_9). Now to prove (3.11), note that by (3.14) for $n > N(\omega_1)$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} |g_i(\hat{\theta}_n)| &\leq \frac{1}{n}\sum_{i=1}^{n} |g_i(\theta_0)| + \frac{1}{n}\sum_{i=1}^{n} |g_i'(\hat{\theta}_n, \theta_0)(\hat{\theta}_n - \theta_0)| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} |g_i(\theta_0)| + \delta\frac{1}{n}\sum_{i=1}^{n} F_{jk}(x_i) \end{aligned} \tag{3.15}$$

where $F_{jk}(x) = \sum_{l=1}^{K} F_{jkl}(x)$ and $F_{jkl}(x)$ is as in condition C_9. As $E\, F_{jk}(X) < \infty$ (C_9), $E\,|g_i(\theta_0)| < \infty$ $(C_6(c))$, and δ is arbitrarily small, the strong law of large numbers implies (3.11). Result (3.12) follows from the strong law again upon noting that for large enough n,

$$\left|\frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\theta}_n) - \frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)\right| \leq \left|\hat{\theta}_n - \theta_0\right| \frac{\delta}{n}\sum_{i=1}^{n} F_{jk}(X_i)$$

and that $E\, g_i(\theta_0) < \infty$ by condition $C_6(c)$. Finally, to prove (3.13), (3.14) gives

$$\frac{1}{n}\max_{1\leq i\leq n} |g_i(\hat{\theta}_n)| \leq \frac{1}{n}\max_{1\leq i\leq n} |g_i(\theta_0)| + \frac{\delta}{n}\max_{1\leq i\leq n} F_{jk}(x_i)$$

for large enough n, $a.s.[P_1]$. As the maxima on the right are being sought on two sets of iid finite mean random variables, Lemma 14 of Appendix A implies (3.13). □

The next lemma gives conditions under which the weighted information matrix is positive definite with high conditional probability.

Lemma 7 *Let $u \in \mathbb{R}^K$ be given with $u \neq 0$. If condtions C_1 to C_7 hold on the model $\mathcal{P}_\Theta$, then uniformly in u, as $n \to \infty$*

$$P(\, u^T \tilde{J}_n(\theta_0) u > 0 \mid X_1^n \,) \to 1 \quad a.s.[P_1]\,. \tag{3.16}$$

If conditions C_8 and C_9 also hold, then for any strongly consistent estimator $\hat{\theta}_n$

$$P(\, u^T \tilde{J}_n(\hat{\theta}_n) u > 0 \mid X_1^n \,) \to 1 \quad a.s.[P_1] \tag{3.17}$$

also uniformly in u.

PROOF. First we prove (3.16). Let $\epsilon > 0$ and $u \in \mathbf{R}^K$ with $u \neq 0$ be given. Suppose that for almost every $\omega_1 \in \Omega_1$, we find $N(\epsilon, \omega_1)$ not depending on u such that for all $n > N(\epsilon, \omega_1)$

$$P\left(\frac{u^T \tilde{J}_n(\theta_0) u}{u^T u} > 0 \mid X_1^n(\omega_1) \right) > 1 - \epsilon \ . \tag{3.18}$$

If so, then (3.16) holds so it suffices to prove (3.16) for vectors u having $u^T u = 1$. For such a vector u, we can write

$$u^T \tilde{J}_n(\theta_0) u = u^T I(\theta_0) u + u^T R u \tag{3.19}$$

where $R = \tilde{J}_n(\theta_0) - I(\theta_0)$. (Recall that $I(\theta_0) = J(\theta_0)$.) The proof amounts to showing that the first term in (3.19) is bounded below by the smallest eigenvalue of $I(\theta_0)$ uniformly in u, and that the second term in (3.19) can be made arbitrarily small with high conditional probability also uniformly in u. To show this, first consider the spectral decomposition of $I(\theta_0) = \Gamma D \Gamma^T$ where Γ is orthogonal and D is a diagonal matrix with the eigenvalues of $I(\theta_0)$ on its diagonal. Since $I(\theta_0)$ is positive definite by condition C_7, the smallest eigenvalue λ_1 is positive. Note that

$$\begin{aligned} u^T I(\theta_0) u &= u^T \Gamma D \Gamma^T u \qquad (3.20) \\ &= s^T D s \qquad\qquad s = \Gamma^T u \quad s^T s = 1 \\ &= \lambda_1 \sum_{k=1}^{K} s_k^2 \lambda_j / \lambda_1 \\ &\geq \lambda_1 \sum_{k=1}^{K} s_k^2 \\ &= \lambda_1. \qquad (3.21) \end{aligned}$$

Now it remains to show that with conditional probability larger than $1 - \epsilon$, $|u^T R u| < \lambda_1$ for large n uniformly in u. The problem can be further simpified by noting that

$$\begin{aligned} |u^T R u| &\leq \sum_{r=1}^{K} \sum_{s=1}^{K} |u_r| |u_s| |R_{rs}| \\ &\leq K^2 M_R \end{aligned}$$

where $M_R < \infty$ bounds the largest magnitude of components of R and using the fact that $|u_r| \leq 1$. Therefore, $|u^T R u| < \lambda_1$ if $M_R < \lambda_1 / K^2$; i. e. if all the elements of

R are smaller than λ_1/K^2 in absolute value for large enough n. Since the conditions for Lemma 5 hold, each component of R converges to zero in conditional probability $a.s.[P_{\theta_0}]$, and so the proof of (3.16) is complete. To prove (3.17) follow the exact same argument as above but with

$$R = \tilde{J}_n(\hat{\theta}_n) - I(\theta_0)\,.$$

By Lemma 6, the components of this R also converge to zero in the above sense. □

3.5 More on conditional consistency

Here is the main theorem on conditional consistency.

Theorem 6 *If the sample information $J_n(\theta)$ converges a.s.$[P_1]$ to $J(\theta)$ uniformly in $\theta \in B$, and if conditions C_1 through C_7 hold in the model $\mathcal{P}_\Theta$, then there exists a conditionally consistent sequence $\{\check{\theta}_n\}$ such that*

$$P\left(\tilde{S}_n(\check{\theta}_n) = 0 \,|\, X_1^n\right) \to 1 \quad a.s.[P_1]\,. \tag{3.22}$$

Moreover, this sequence is essentially unique in the following sense: If $\{\bar{\theta}_n\}$ satisfies the limit above, and is also conditionally consistent, then

$$P\left(\bar{\theta}_n = \check{\theta}_n \,|\, X_1^n\right) \to 1 \quad a.s.[P_1]\,. \tag{3.23}$$

PROOF. To prove this result, we apply the Inverse Function Theorem (see Section A.4 in Appendix A) to $\tilde{S}_n$ following an argument first used by Foutz (1977). In doing so, we find a neighbourhood U_δ of θ_0 such that $\tilde{S}_n(\theta)$ is one-to-one from U_δ onto $\tilde{S}_n(U_\delta)$. Moreover, the image $\tilde{S}_n(U_\delta)$ contains 0 with high conditional probability. On this image, $\tilde{S}_n^{-1}$ is well defined, so $\check{\theta}_n = \tilde{S}_n^{-1}(0)$ is a root of the weighted likelihood equation and is, roughly speaking, close to θ_0. Making this more precise, we have the following proof.

Define

$$\alpha = \frac{1}{4 \parallel J(\theta_0)^{-1} \parallel}$$

using condition C_7 and the fact that $I = J$ at θ_0. By continuity of the second partials of $\log f_\theta$ given in condition C_5, and by the bound in condition $C_6(c)$, it follows that

$J(\theta)$ is continuous on B (use Theorem 16.8(i) of Billingsley (1986)). Therefore, there exists $\delta > 0$ sufficiently small so that

$$\| J(\theta) - J(\theta_0) \| < \alpha/3 \tag{3.24}$$

whenever $\theta \in U_\delta \subset B$ where $U_\delta \equiv \{\theta : \| \theta - \theta_0 \| < \delta\}$. Let $C_n(\omega_1) \subset \Omega_2$ be the set where

$$\| \tilde{J}_n(\theta) - J(\theta) \| < \alpha/3 \qquad \text{for all } \theta \in U_\delta. \tag{3.25}$$

By uniform convergence of $\tilde{J}_n(\theta)$ ensured by Lemma 5, $P(\, C_n \mid X_1^n \,) \to 1$ $a.s.[P_{\theta_0}]$. On the set $D_n(\omega_1) \subset \Omega_2$ where $\tilde{J}_n(\theta_0)$ is invertible, define

$$\alpha_n = \frac{1}{4 \, \| \, \tilde{J}_n(\theta_0)^{-1} \, \|} \, .$$

By Lemma 7 $P(\, D_n \mid X_1^n \,) \to 1$ $a.s.[P_{\theta_0}]$.

We must show that α_n converges to α in some sense. The lemmas stated above imply that

$$\tilde{J}_n(\theta_0) \to_{c.p} J(\theta_0) \quad a.s.[P_1]$$

and the matrices in this sequence are invertible with high conditional probability $a.s.[P_1]$. Therefore, by continuity of matrix inversion, it must also be true that

$$\tilde{J}_n(\theta_0)^{-1} \to_{c.p} J(\theta_0)^{-1} \quad a.s.[P_1] \, .$$

This convergence implies that the norm of $\tilde{J}_n(\theta_0)^{-1}$ must be converging in conditional probability $a.s.[P_1]$ to the norm of $J(\theta_0)^{-1}$ and hence $\alpha_n \to_{c.p.} \alpha$ $a.s.[P_1]$. Let $E_n(\omega_1)$ be the subset of $C_n(\omega_1) \bigcap D_n(\omega_1)$ where $|\alpha_n - \alpha| < \alpha/2$. On the set $E_n(\omega_1)$, having conditional probability converging to 1 $a.s.[P_1]$, We apply the triangle inequality to get the following bound on the deviation in the weighted information matrix for $\theta \in U_\delta$:

$$\begin{aligned} \| \tilde{J}_n(\theta) - \tilde{J}_n(\theta_0) \| &\leq \| \tilde{J}_n(\theta) - J(\theta) \| + \| J(\theta_0) - \tilde{J}_n(\theta_0) \| + \| J(\theta) - J(\theta_0) \| \\ &< \frac{\alpha}{3} + \frac{\alpha}{3} + \frac{\alpha}{3} = \alpha \\ &\leq 2\alpha_n \, . \end{aligned} \tag{3.26}$$

It is now apparent that the conditions needed to apply the Inverse Function Theorem to $\tilde{S}_n(\theta)$ are satisfied on $E_n(\omega_1)$, $a.s.[P_1]$. This weighted score function is continuously differentiable on B and has a matrix of derivatives which is invertible at $\theta_0 \in B$. The Inverse Function Theorem implies that on $E_n(\omega_1)$, $a.s.[P_1]$,

(a) For every θ_1, θ_2 in U_δ

$$\| \tilde{S}_n(\theta_1) - \tilde{S}_n(\theta_2) \| \geq 2\alpha_n \| \theta_1 - \theta_2 \|,$$

(b) The image set

$$\tilde{S}_n(U_\delta) = \left\{ s \in \mathbf{R}^K \ : \ \tilde{S}_n(\theta) = s, \quad \theta \in U_\delta \right\}$$

contains the open ball of radius $\alpha_n \delta$ centered at $\tilde{S}_n(\theta_0)$.

Implication (a) ensures that on E_n, $a.s.[P_1]$, the weighted score $\tilde{S}_n(\theta)$ is a one-to-one function from U_δ onto the image set $\tilde{S}_n(U_\delta)$ and so the inverse funtion $\tilde{S}_n^{-1}(y)$ mapping $\tilde{S}_n(U_\delta)$ onto U_δ is well-defined. Furthermore, since $|\alpha_n - \alpha| < \alpha/2$ on $E_n(\omega_1)$, $a.s.[P_1]$,

(b^*) The image set $\tilde{S}_n(U_\delta)$ contains the open ball of radius $\alpha\delta/2$ centered at $\tilde{S}_n(\theta_0)$.

Consider the subset $E_n'(\omega_1)$ of $E_n(\omega_1)$ on which $\| \tilde{S}_n(\theta_0) \| < \alpha\delta/2$. By Lemma 5, $\tilde{S}_n(\theta_0)$ is converging in conditional probability to 0, and so

$$P(E_n' \,|\, X_1^n) \to 1 \quad a.s.[P_{\theta_0}] .$$

Also, on $E_n'(\omega_1)$, $0 \in \tilde{S}_n(U_\delta)$, $a.s.[P_1]$.

Let $N \subset \Omega_1$ be the exceptional set of data sequences where all the previous convergences failed. Define the random variable

$$\check{\theta}_n = \begin{cases} \tilde{S}_n^{-1}(0) & \text{if } \omega_1 \in N^c \text{ and } \omega_2 \in E_n'(\omega_1) \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

Clearly, $\{\check{\theta}_n\}$ forms a sequence of roots of the weighted likelihood equation with increasing conditional probability, $a.s.[P_1]$, establishing (3.22). Moreover, since δ can

be made arbitrarily small, this sequence of roots is conditionally consistent:

$$P(\,|\check{\theta}_n - \hat{\theta}_n| < \delta \,|\, X_1^n\,) \geq P(\,E_n'(\omega_1)\,|\,X_1^n\,) \to 1 \quad a.s.[P_1]\,.$$

By the one-to-oneness of $\tilde{S}_n$ on U_δ, any other sequence $\{\bar{\theta}_n\}$ satisfying $\tilde{S}_n(\bar{\theta}_n) = 0$ necessarily lies outside of U_δ with conditional probability going to 1 almost surely. This gives (3.23) completing the proof.

□

3.6 Conditional Asymptotic Normality

First, a lemma about the conditional distribution of linear combinations of the weighted score vector;

Lemma 8 *Let z be a unit vector in $\mathbf{R}^K$, and put $a_{in} = \sum_{k=1}^K z_k[\psi_i(\hat{\theta}_n)]_k$ for a strongly consistent estimator $\hat{\theta}_n$. Under regularity conditions C_1 to C_8 and C_{10}*

$$\frac{1}{n}\sum_{i=1}^n a_{in}^2 \to z^T I(\theta_0) z \quad a.s.[P_1] \tag{3.27}$$

and

$$\frac{1}{n}\max_{1\leq i\leq n} a_{in}^2 \to 0 \quad a.s.[P_1] \tag{3.28}$$

PROOF. To begin, define

$$h_i(\theta) = \left(\sum_{k=1}^K z_k[\psi_i(\theta)]_k\right)^2 .$$

Note that $a_{in}^2 = h_i(\hat{\theta}_n)$. By differentiability conditions C_4, C_5, and C_8, a second order Taylor series expansion of each h_i about θ_0 is justified. Evaluating this expansion at $\hat{\theta}_n$, we have

$$a_{in}^2 = h_i(\theta_0) + [\,h_i'(\theta_0, \hat{\theta}_n)\,]^T(\hat{\theta}_n - \theta_0) \tag{3.29}$$

where $h_i'(\hat{\theta}_n, \theta_0)$ is a vector of derivatives of h_i evaluated at θ_i^* on the line between $\hat{\theta}_n$ and θ_0. By averaging,

$$\frac{1}{n}\sum_{i=1}^n a_{in}^2 = \frac{1}{n}\sum_{i=1}^n h_i(\theta_0) + R_n$$

where R_n is the average of the remainder terms from (3.29). By the strong law of large numbers, the average of the $h_i(\theta_0)$'s converges to its expectation $z^T I(\theta_0) z$ $a.s.[P_1]$. Therefore, (3.27) follows if we prove that $R_n \to 0 \quad a.s.[P_1]$.

The u^{th} element of the vector $h_i'(\theta_0, \hat{\theta}_n)$ can be written

$$(h_i'(\theta_0, \hat{\theta}_n))_u = 2 \sum_{j=1}^{K} \sum_{k=1}^{K} z_j z_k \frac{\partial \log f_\theta(x_i)}{\partial \theta_j} \frac{\partial^2 \log f_\theta(x_i)}{\partial \theta_u \theta_k}$$

where all the derivatives are evaluated at θ_i^*. Suppose that n is large enough so that each component of $\hat{\theta}_n$ is within δ of the corresponding component of θ_0. Suppose further that δ is small enough so that the bounds on the products of partials given in condition C_{10} hold for all $1 \le u, j, k \le K$. Then

$$\begin{aligned} |R_n| &\le \frac{\delta}{n} \sum_{i=1}^{n} \sum_{u=1}^{K} |(h_i'(\theta_0, \hat{\theta}_n))_u| \\ &\le \frac{2\delta}{n} \sum_{i=1}^{n} \bar{G}(x_i) \end{aligned}$$

where

$$\bar{G}(x) = \sum_{u=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} G_{ujk}(x)$$

and the functions G_{ujk} are as in condition C_{10}. Since $E\,\bar{G}(X_i) < \infty$ the strong law of large numbers applies to the average of the $\bar{G}$'s. R_n must converge $a.s.[P_1]$ to 0 since δ is arbitrarily small.

To prove (3.28), we use the expansion of h_i given above to argue that

$$\frac{1}{n} \max_{1 \le i \le n} a_{in}^2 \le \frac{1}{n} \max_{1 \le i \le n} h_i(\theta_0) + \frac{2\delta}{n} \max_{1 \le i \le n} \bar{G}(x_i)\,.$$

Since the maximizations above are over two sets of iid random variables with finite first moments, (3.28) follows immediately from Lemma 14 of Appendix A. □

The next theorem establishes the first order correctness of the weighted likelihood method. That is, $\tilde{\theta}_n$ has an asymptotic distribution which is the same as the asymptotic posterior distribution of the parameter θ under any one of a large class of priors.

Theorem 7 (Asymptotic Distribution) *Suppose that conditions C_1 to C_{10} hold on the family $\mathcal{P}_\Theta$, and that $\{\hat{\theta}_n\}$ is a strongly consistent estimator of θ_0 satisfying*

$$|\sqrt{n}S_n(\hat{\theta}_n)| \to 0 \quad a.s.[P_1]\,. \tag{3.30}$$

If $\{\check{\theta}_n\}$ is a conditionally consistent sequence of roots of the weighted likelihood equation, then for any Borel set $A \subset \mathbf{R}^K$

$$P(\,\sqrt{n}(\check{\theta}_n - \hat{\theta}_n) \in A \,|\, X_1^n\,) \to P(\,Z \in A\,) \quad a.s.[P_1] \tag{3.31}$$

where $Z \sim N_K\,(0, I(\theta_0)^{-1})$.

Note that condition (3.30) is satisfied trivially if $\hat{\theta}_n$ is the maximum likelihood estimator computed by solving the likelihood equation. Also, $\check{\theta}_n$ is a conditionally consistent sequence of roots of the weighted likelihood equation, but it may not equal the maximizing sequence $\tilde{\theta}_n$.

PROOF. Under differentiability conditions C_4, C_5, and C_8, the weighted score function can be represented as a second order Taylor series centered at $\hat{\theta}_n$. This expansion can be evaluated at $\check{\theta}_n$ giving for all $\omega_2 \in \Omega_2$

$$\tilde{S}_n(\hat{\theta}_n) = (\tilde{J}_n(\hat{\theta}_n) - R_n)(\tilde{\theta}_n - \hat{\theta}_n) \tag{3.32}$$

where R_n is a K by K matrix with j^{th} row

$$[R_n]_{j\cdot} = \frac{1}{2\sum_{i=1}^n Y_i}(\check{\theta}_n - \hat{\theta}_n)^T \sum_{i=1}^n Y_i \Psi_i^j(\theta^*)\,. \tag{3.33}$$

Here, θ^* is a point on the line segment joining $\hat{\theta}_n$ to $\check{\theta}_n$, and the matrix $\Psi_i^j(\theta)$ contains third order partial derivatives of $\log f_\theta(X_i)$:

$$[\Psi_i^j(\theta)]_{lk} = \frac{\partial^3 \log f_\theta(X_i)}{\partial\,\theta_j\,\partial\,\theta_k\,\partial\,\theta_l}\,.$$

The first part of the proof amounts to showing that the matrix $\tilde{J}_n(\hat{\theta}_n) - R_n$ both converges in some sense to $I(\theta_0)$ and is invertible with high conditional probability. Letting $B_n(\omega_1) \subset \Omega_2$ be the set where $\tilde{J}_n(\hat{\theta}_n) - R_n$ is invertible, We invoke an argument similar to that in Lemma 7 to show that

$$P(\,B_n \,|\, X_1^n\,) \to 1 \quad a.s.[P_1]\,. \tag{3.34}$$

To prove (3.34), it is sufficient to show that

$$P(\, u^T(\tilde{J}_n(\hat{\theta}_n) - R_n)u > 0 \mid X_1^n \,) \to 1 \quad a.s.[P_1]$$

uniformly for $u \in \mathbf{R}^K$ with $\| u \| = 1$. Since $\tilde{J}_n(\hat{\theta}_n)$ is positive definite with conditional probability converging to 1 $a.s.[P_1]$ by Lemma 7, it suffices to show that $|u^T R_n u|$ can be made arbitrarily small uniformly in u in the same asymptotic sense. Note that

$$\begin{aligned} |u^T R_n u| &= \sum_{j=1}^{K}\sum_{k=1}^{K}[u]_j[u]_k[R_n]_{jk} \\ &\leq K^2 \max_{j,k} |[R_n]_{jk}| \,. \end{aligned}$$

Proving (3.34) has been reduced therefore to showing that each element of R_n can be made arbitrarily small with high conditional probability $a.s.[P_1]$. By construction

$$\begin{aligned} [R_n]_{jk} &= \frac{1}{2\sum_{i=1}^n Y_i}(\check{\theta}_n - \hat{\theta}_n)^T \sum_{i=1}^n Y_i [\Psi_i^j(\theta^*)]_{\cdot k} \\ &= \frac{1}{2\sum_{i=1}^n Y_i} \sum_{i=1}^n Y_i \left(\sum_{l=1}^{K} [\check{\theta}_n - \hat{\theta}_n]_l [\Psi_i^j(\theta^*)]_{lk} \right) . \end{aligned}$$

Let $\eta > 0$ be given. Consider $\epsilon > 0$ sufficiently small so that the bounds on the third order partial derivatives of the log density hold for all $\| \theta - \theta_0 \| < \epsilon$ (see condition C_9). Suppose, moreover, that $\hat{\theta}_n(\omega_1)$ is within an $\epsilon/2$ ball of θ_0 for all $n > N(\omega_1, \epsilon)$. This is possible $a.s.[P_1]$ by the strong consistency assumption on $\hat{\theta}_n$. Also, let $C_n(\omega_1) \subset \Omega_2$ be the set where $\check{\theta}_n$ is within $\epsilon/2$ of θ_0. For $n > N(\omega_1, \epsilon, \eta)$, this set has conditional probability larger than $1 - \eta/2$ $a.s.[P_1]$ by conditional consistency of $\check{\theta}_n$. By the triangle inequality, $\check{\theta}_n$ is within ϵ of $\hat{\theta}_n$ on C_n. Moreover, all points in between $\check{\theta}_n$ and $\hat{\theta}_n$ are within $\epsilon/2$ of θ_0 so the bound on the third order partials is active (they are evaluated at θ^* in between $\hat{\theta}_n$ and $\check{\theta}_n$). On $C_n(\omega_1)$, $a.s.[P_1]$, therefore

$$\begin{aligned} |[R_n]_{jk}| &\leq \frac{\epsilon}{2\sum_{i=1}^n Y_i} \sum_{i=1}^n Y_i \sum_{l=1}^{K} F_{jkl}(X_i) \\ &= \frac{\epsilon}{2\bar{Y}_n} \frac{1}{n} \sum_{i=1}^n Y_i \bar{F}_{jk}(X_i) \,. \end{aligned}$$

The integrable functions $F_{jkl}(x)$ bound the third derivatives of the log density, and $\bar{F}_{jk}(x) = \sum_l F_{jkl}(x)$. By Lemma 3, $\sum Y_i \bar{F}_{jk}/n$ converges in conditional probability

to $\xi = E\,\bar{F}_{jk}(X)$ $a.s.[P_1]$. As noted before, $\bar{Y}_n$ converges to 1 in the same sense. Therefore, on a set $D_n(\omega_1)$ having conditional probability larger than $1 - \eta/2$

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y_i \bar{F}_{jk}(x_i)\frac{1}{2\bar{Y}_n} - \frac{\xi}{2}\right| < \epsilon$$

for large enough n, $a.s.[P_1]$. On the set $C_n(\omega_1) \bigcap D_n(\omega_1)$ having conditional probability larger than $1 - \eta$

$$|[R_n]_{jk}| \leq \epsilon(\epsilon + \frac{\xi}{2})$$

for $n > N(\omega_1, \epsilon, \eta)$. Since $\epsilon > 0$ is arbitrary, it follows that $[R_n]_{jk} \rightarrow_{c.p.} 0$ $a.s.[P_1]$. The proof of (3.34) is complete. The proof has yielded a useful consequence. By proving that the elements of R_n converge to zero in the above sense, it follows from Lemma 6 that

$$(\tilde{J}_n(\hat{\theta}_n) - R_n) \rightarrow_{c.p.} I(\theta_0) \quad a.s.[P_1]\,.$$

Next, define $\tilde{\Sigma}_n$ to be the inverse of $(\tilde{J}_n(\hat{\theta}_n) - R_n)$ when it exists (on the set $B_n(\omega_1)$ and arbitrarily on B_n^c). By continuity of matrix inversion and using the Continuous Mapping Theorem, the above results imply

$$\tilde{\Sigma}_n \rightarrow_{c.p.} I(\theta_0)^{-1} \quad a.s.[P_1]\,.$$

To see what remains to be proved, note that (3.32) can be written

$$\sqrt{n}(\check{\theta}_n - \hat{\theta}_n) = \tilde{\Sigma}_n \cdot \sqrt{n}\tilde{S}_n(\hat{\theta}_n)$$

on the set $B_n(\omega_1)$ having conditional probability converging to 1 $a.s.[P_1]$. The main result (3.31) follows by Slutsky's Theorem (see Bickel and Doksum, 198?, for example) (applied conditionally along P_1-almost every sample path) if

$$\sqrt{n}\tilde{S}_n(\hat{\theta}_n) \rightarrow_{c.d.} N_K(0, I(\theta_0))\,. \tag{3.35}$$

The notation $\rightarrow_{c.d.}$ means convergence in conditional distribution.

Using the Cramer-Wold device (reference), we prove (3.35) by showing that for any $z \in \mathbf{R}^K$ with $|z| = 1$

$$t_n(z) \equiv \sqrt{n} z^T \tilde{S}_n(\hat{\theta}_n) \rightarrow_{c.d.} N_K(0, z^T I(\theta_0) z) \quad a.s.[P_1]\,. \tag{3.36}$$

To show (3.36), write

$$t_n(z) = \sqrt{n} \sum_{k=1}^{K} z_k \left(\frac{\sum_{i=1}^{n} Y_i g_i^k(\hat{\theta}_n)}{\sum_{i=1}^{n} Y_i} \right) \tag{3.37}$$

where $g_i^k(\theta) = \partial \log f_\theta(x_i)/\partial \theta_k$ and Y_i are iid unit exponential random variables determining the weights in the weighted likelihood. Changing the order of summation in (3.37) I have

$$t_n(z) = \frac{1}{\bar{Y}_n} \frac{\sum_{i=1}^{n} a_{in} Y_i}{\sqrt{n}}$$

where $a_{in} = \sum_{k=1}^{K} z_k g_i^k(\hat{\theta}_n)$. Noting that $\bar{Y}_n \to 1 \quad a.s.[P_2]$, it suffices to prove that $\sum a_{in} Y_i/\sqrt{n}$ converges in the same sense and to the same limit as in (3.36). This follows from a corollary of the Lindeberg-Feller-Lévy Central Limit Theorem (see Appendix A) because, by Lemma 8, the following two convergence results hold:

$$\frac{1}{n} \sum_{i=1}^{n} a_{in}^2 \to z^T I(\theta_0) z \quad a.s.[P_1]$$

$$\frac{1}{n} \max_{1 \le i \le n} a_{in}^2 \to 0 \quad a.s.[P_1]$$

and by (3.30)

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} a_{in} \to 0 \quad a.s.[P_1] .$$

The proof is complete. □

3.7 Asymptotic Skewness

For a one dimensional model, we show that the skewness of the conditional distribution of $\tilde{\theta}$ is asymptotically equivalent to the skewness of a postererior distribution. To simply state the theorem, we write

$$\begin{aligned} \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) &= \frac{\sqrt{n}}{J_n(\hat{\theta}_n)} \tilde{S}_n(\hat{\theta}_n) + R_n \\ &= \frac{1}{\sqrt{n} J_n(\hat{\theta}_n)} \sum_i w_{n,i} \psi_i(\hat{\theta}_n) + R_n . \end{aligned} \tag{3.38}$$

In (3.38), $J_n(\hat{\theta}_n)$ is the observed information (negative of the second derivative of the log likelihood), $\tilde{S}_n$ is the weighted score function and R_n is a remainder term. Results in the previous section require that the remainder term R_n converge in conditional probability to 0 along almost every sample path. To guarantee convergence of the skewness, R_n must converge to 0 in a stronger sense. Note, for example, that $R_n = 0$ if θ is a Poisson mean.

Theorem 8 (Conditional Skewness) *Suppose the MLE $\hat{\theta}_n$ is strongly consistent. If conditions C_1 through C_9 and C_{11} hold on the model, and*

$$n^2 E\left(R_n^4 \mid X_1^n\right) \to 0 \quad a.s.[P_1]$$

as $n \to \infty$, then

$$\gamma_n := n^2 E\left((\tilde{\theta}_n - \hat{\theta}_n)^3 \mid X_1^n\right) \to \frac{2E\left((\psi_1(\theta_0))^3\right)}{\left(I(\theta_0)\right)^3} \quad a.s.[P_1]. \tag{3.39}$$

PROOF. Define

$$\begin{aligned} Z_n &:= \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) - R_n \\ &= \frac{1}{\sqrt{n} J_n(\hat{\theta}_n)} \sum_i w_{n,i} \psi_i(\hat{\theta}_n). \end{aligned} \tag{3.40}$$

From (3.39),

$$\gamma_n = \sqrt{n} E\left\{ Z_n^3 + 3Z_n R_n^2 + 3Z_n^2 R_n + R_n^3 \mid X_1^n \right\}. \tag{3.41}$$

Suppose that the following convergence statement holds almost surely $[P_1]$ for some constants c_2 and c_4:

$$E(\, Z_n^p \mid X_1^n) \to c_p \qquad p = 2, 4. \tag{3.42}$$

Assuming (3.42), we can show that the dominant term in (3.41) is $\sqrt{n}E(Z_n^3|X_1^n)$. To see this, recall Liapounov's inequality which states that for any random variable U, $(E|U|^p)^{1/p}$ is increasing in p. Considering the last three terms in (3.41), we have by application of both the Liapounov and Cauchy-Schwartz inequalities that

$$\left|\sqrt{n} E(\, Z_n R_n^2 \mid X_1^n)\right| \leq \left(E(\, Z_n^2 \mid X_1^n)\right)^{1/2} \left(nE(\, R_n^4 \mid X_1^n)\right)^{1/2}$$

$$\leq \left(E(\,Z_n^2 \,|\, X_1^n\,)\right)^{1/2} \left(n^2 E(\,R_n^4 \,|\, X_1^n\,)\right)^{1/2},$$

$$\left|\sqrt{n} E(\,Z_n^2 R_n \,|\, X_1^n\,)\right| \leq \sqrt{n} \left(E(\,Z_n^4 \,|\, X_1^n\,)\right)^{1/2} \left(E(\,R_n^2 \,|\, X_1^n\,)\right)^{1/2}$$

$$\leq \left(E(\,Z_n^4 \,|\, X_1^n\,)\right)^{1/2} \left(n^2 E(\,R_n^4 \,|\, X_1^n\,)\right)^{1/4},$$

and

$$\left|\sqrt{n} E(\,R_n^3 \,|\, X_1^n\,)\right| \leq \sqrt{n} \left(E(\,R_n^4 \,|\, X_1^n\,)\right)^{3/4} \leq \left(n^2 E(\,R_n^4 \,|\, X_1^n\,)\right)^{3/4}.$$

Considering the initial assumption on R_n and under the proviso (3.42), we have

$$\gamma_n = \sqrt{n}\, E\left(\, Z_n^3 \,|\, X_1^n \,\right) + o(1) \qquad a.s.[P_1]$$

and so it suffices to show that $\sqrt{n}\, E(Z_n^3 | X_1^n)$ converges to the limit in (3.39).

First of all, there are sufficient smoothness conditions on the model to ensure that

$$J_n(\hat{\theta}_n) \to I(\theta_0) \quad a.s.[P_1].$$

Thus again under the proviso (3.42), it suffices to show that with $U_n := J_n(\hat{\theta}_n) Z_n$,

$$\sqrt{n}\, E\left(\, U_n^3 \,|\, X_1^n \,\right) \to 2E\left((\psi_X(\theta_0))^3\right) \quad a.s.[P_1]. \tag{3.43}$$

Computing with $a_i = \psi_i(\hat{\theta}_n)$, $\lambda_i = w_{n,i}/n$, and using $\sum_i a_i = 0$, we use the moments of Dirichlet vectors (see (A.7) of Appendix A) to get

$$\begin{aligned}
\sqrt{n}\, E\left(\, U_n^3 \,|\, X_1^n \,\right) &= n^2\, E\left\{ \sum_{i,j,k} \lambda_i \lambda_j \lambda_k a_i a_j a_k \,\middle|\, X_1^n \right\} \\
&= n^2\, E\left\{ \sum_{i\neq j\neq k} \lambda_i \lambda_j \lambda_k a_i a_j a_k + 3 \sum_{i\neq j} \lambda_i^2 \lambda_j a_i^2 a_j + \sum_i \lambda_i^3 a_i^3 \,\middle|\, X_1^n \right\} \\
&= \frac{n^2}{n(n+1)(n+2)} \left(\sum_{i\neq j\neq k} a_i a_j a_k + 6 \sum_{i\neq j} a_i^2 a_j + 6 \sum_i a_i^3 \right) \\
&= \frac{n}{(n+1)(n+2)} \left\{ \left(\sum_i a_i \right)^3 - 3 \sum_{i\neq j} a_i^2 a_j - \sum_i a_i^3 + 6 \sum_{i\neq j} a_i^2 a_j \right. \\
&\quad \left. + 6 \sum_i a_i^3 \right\} \\
&= \frac{n}{(n+1)(n+2)} \left(3 \sum_{i\neq j} a_i^2 a_j + 5 \sum_i a_i^3 \right) \\
&= \frac{2n}{(n+1)(n+2)} \sum_i a_i^3 .
\end{aligned}$$

Proving (3.43) amounts to showing that

$$\frac{1}{n}\sum_i a_i^3 \to E\left((\psi_1(\theta_0))^3\right) \quad a.s.[P_1]\,. \tag{3.44}$$

By definition, $a_i = \psi_i(\hat{\theta}_n)$. A Taylor expansion of each ψ_i about θ_0 leads to

$$\frac{1}{n}\sum_i a_i^3 = \frac{1}{n}\sum_i (\psi_i(\theta_0))^3 + (\hat{\theta}_n - \theta_0)Q_n \tag{3.45}$$

where $Q_n = 3/n\sum_i \psi_i'(\theta_i^*)(\psi_i(\theta_i^*))^2$. Each θ_i^* is in between $\hat{\theta}_n$ and θ_0. We can apply the strong law of large numbers to the first term in (3.45). Therefore (3.43) follows if the remainder term $(\hat{\theta}_n - \theta_0)Q_n$ converges P_1—almost surely to zero. Let $\epsilon > 0$ be given. Suppose ϵ is small enough so that the the length of B is larger than 2ϵ. By strong consistency of $\hat{\theta}_n$, we can find $N(\omega_1, \epsilon)$ large enough so that $|\hat{\theta}_n(\omega_1) - \theta_0| < \epsilon$ for all $n > N(\omega_1, \epsilon)$. Applying the Cauchy-Schwarz inequality and using the bound given by condition C_{11}, we have

$$|(\hat{\theta}_n - \theta_0)Q_n| \le 3\epsilon \left(\frac{1}{n}\sum_i (\psi_i'(\theta_i^*))^2\right)^{1/2} \left(\frac{1}{n}(\psi_i(\theta_i^*))^4\right)^{1/2} \tag{3.46}$$

$$\le 3\epsilon \left(\frac{1}{n}\sum_i (H_1(X_i))^2\right)^{1/2} \left(\frac{1}{n}(H_2(X_i))^4\right)^{1/2}\,. \tag{3.47}$$

Since the averages in (3.46) converge by the strong law to finite expectations, the remainder in (3.45) does converge almost surely $[P_1]$ to zero. Thus (3.43) is proved.

The above conclusion holds under the proviso (3.42). In light of the almost sure convergence of $J_n(\hat{\theta}_n)$, (3.42) is true if constants c_2 and c_4 exist such that

$$E(\,U_n^p \,|\, X_1^n\,) \to c_p \quad a.s.[P_1] \qquad p = 2, 4\,. \tag{3.48}$$

The necessary calculations are similar to one above for the conditional mean of U_n^3. We have

$$E(\,U_n^2 \,|\, X_1^n\,) = n\,E\left\{ \sum_{i,j} \lambda_i\lambda_j a_i a_j \,\middle|\, X_1^n \right\}$$

$$= n\,E\left\{ \sum_{i\neq j} \lambda_i\lambda_j a_i a_j + \sum_i \lambda_i^2 a_i^2 \,\middle|\, X_1^n \right\}$$

$$= \frac{n}{n(n+1)} \left\{ \sum_{i \neq j} a_i a_j + 2 \sum_i a_i^2 \right\}$$
$$= \frac{1}{n+1} \sum_i a_i^2 .$$

Similarly

$$E(U_n^4 \mid X_1^n) = n^2 E \left\{ \sum_{i,j,k,l} \lambda_i \lambda_j \lambda_k \lambda_l a_i a_j a_k a_l \,\middle|\, X_1^n \right\}$$
$$= \frac{n^2}{n(n+1)(n+2)(n+3)} \left\{ \sum_{i \neq j \neq k \neq l} a_i a_j a_k a_l + 12 \sum_{i \neq j \neq k} a_i^2 a_j a_k \right.$$
$$\left. +12 \sum_{i \neq j} a_i^2 a_j^2 + 24 \sum_{i \neq j} a_i^3 a_j + 24 \sum_i a_i^4 \right\}$$
$$= \frac{n}{(n+1)(n+2)(n+3)} \left\{ 6 \sum_{i \neq j \neq k} a_i^2 a_j a_k + 9 \sum_{i \neq j} a_i^2 a_j^2 + 20 \sum_{i \neq j} a_i^3 a_j \right.$$
$$\left. +23 \sum_i a_i^4 \right\}$$
$$= \frac{n}{(n+1)(n+2)(n+3)} \left\{ 3 \sum_{i \neq j} a_i^2 a_j^2 + 8 \sum_{i \neq j} a_i^3 a_j + 17 \sum_i a_i^4 \right\}$$
$$= \frac{3n}{(n+1)(n+2)(n+3)} \left\{ 2 \sum_i a_i^4 + \left(\sum_i a_i^2 \right)^2 \right\} .$$

By an argument analogous to the one used to prove (3.44), we get

$$c_2 = E\left((\psi_1(\theta_0))^2 \right) \quad \text{and} \quad c_4 = 3 \left[E\left((\psi_X(\theta_0))^2 \right) \right]^2 ,$$

thus completing the proof. □

Chapter 4

BOOTSTRAPPING AND PARTIAL LIKELIHOOD

4.1 Introduction and summary

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 while the other part, the partial likelihood, does not depend on the nuisance parameter ψ. The partial likelihood is used, therefore, in inference about θ. In this chapter, the natural analog of weighted likelihood bootstrapping is derived for partial likelihood. After developing the general method, we give a detailed analysis of this bootstrap for the two-sample proportional hazards model of Cox (1972). In this special case, our new bootstrap is at least first order correct, asymptotically, in the following sense: The conditional distribution of $\tilde{\theta}$, the maximizer of the weighted partial likelihood, mimics the sampling distribution of $\hat{\theta}_n$, the maximizer of the partial likelihood, to at least the mean and variance.

Suppose the data X can be transformed into a sequence

$$(U_1, V_1, U_2, V_2, \cdots, U_n, V_n)\,.$$

The size n itself may be random. The complete likelihood function is

$$\begin{aligned} L_n(\theta,\psi) &= f_{\theta,\psi}(u_1, v_1, \cdots, u_n, v_n) \\ &= \prod_{i=1}^{n} f_{\theta,\psi}(u_i, s_i | u_1^{i-1}, v_1^{i-1}) \\ &= \prod_{i=1}^{n} f_{\theta,\psi}(u_i | u_1^{i-1}, v_1^{i-1}) \prod_{i=1}^{n} f_{\theta}(v_i | u_1^{i}, v_1^{i-1}) \end{aligned}$$

The second product in the last line above is called the partial likelihood for θ based on (V_i) in the sequence (U_i, V_i). The value of θ maximizing the partial likelihood is denoted $\hat{\theta}_n$.

The complete weighted likelihood, as defined in Chapter 1, can also be factored

to produce a weighted partial likelihood.

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

The second product in the last line above is called the weighted partial likelihood based on (V_i) in the sequence (U_i, V_i). We let $\tilde{\theta}_n$ be the value of θ maximizing this function. As before, $w_{n,i} = Y_i/\bar{Y}_n$ for iid exponentials Y_i with sample mean $\bar{Y}_n$.

The conditional distribution of $\tilde{\theta}_n$ (given data) can be used as a surrogate for the unknown sampling distribution of $\hat{\theta}_n$. This is a bootstrap procedure called the weighted, partial-likelihood bootstrap. For the two sample proportional hazards model, we show that this bootstrap procedure is asymptotically first-order correct in approximating the sampling distribution of $\hat{\theta}_n$. Our result is restricted to the model without censoring.

4.2 Cox's proportional hazards model

Inference for the proportional hazards model in survival analysis, Cox (1972), provides the quintessential example of partial likelihood. When considering the survival time of an individual with covariate information Z_i, it is assumed that the hazard function takes the form

$$\psi(t)\exp(\theta^T Z_i)$$

where $\psi(t)$ is the baseline hazard function. Thus the ratio of hazards for two individuals depends only on θ and on the differences in their covariate values, and not on the baseline hazard function $\psi(t)$.

Following the standard development, suppose that the distinct, ordered, uncensored survival times are $X_{(1)} < X_{(2)} < \cdots X_{(n)}$. Let $Z_{(j)}$ be the covariate information for the individual with survival time $X_{(j)}$, and let $\mathcal{R}_j$ be the risk set, the set of labels for individuals with censored or uncensored survival times no less than $X_{(j)}$. The complete data set X is the collection of observed survival or censoring times, indicators of censoring, and covariate information. We treat the covariate information as fixed throughout. To develop a partial likelihood, we define U_i and V_i as follows: U_i

indicates what censoring has occured in $[X_{(i-1)}, X_{(i)})$ and that some individual has *failed* at time $X_{(i)}$. V_i specifies the individual in the risk set $\mathcal{R}_i$ who fails at $X_{(i)}$. Accordingly, the partial likelihood for θ based on (V_i) is

$$L_{\mathcal{P}}(\theta) = \prod_{i=1}^{n} \frac{\exp(\theta^T Z_{(i)})}{\sum_{j \in \mathcal{R}_i} \exp(\theta^T Z_{(j)})}.$$

By design, this function does not depend on the baseline hazard. Note that n is the number of observed, uncensored times.

Similarly, the weighted partial likelihood for θ becomes

$$\tilde{L}_{\mathcal{P}}(\theta) = \prod_{i=1}^{n} \left(\frac{\exp(\theta^T Z_{(i)})}{\sum_{j \in \mathcal{R}_i} \exp(\theta^T Z_{(j)})} \right)^{w_{n,i}},$$

where $(w_{n,1}, w_{n,2}, \cdots, w_{n,n})$ is n times a uniform Dirichlet random vector.

4.3 The two sample Cox model; no censoring

In our asymptotic analysis, it is convenient to adhere to the following notation. We suppose that $X_1, X_2, \cdots, X_m \sim_{iid} F$ having density f on the positive line, and similarly $Y_1, Y_2, \cdots, Y_n \sim_{iid} G$ with density g. We suppose that $m = m(n)$ such that the sampling fraction $\lambda_n = m/n \to \lambda \in (0,1)$ as $n \to \infty$. The empirical distribution functions of the respective samples are denoted F_m and G_n, and $T_1, T_2, \cdots T_{m+n}$ denotes the pooled sample of survival times. Covariates Z_i indicate whether or not T_i comes from the sample of Y's. For any $H \in [0,1]$, $\bar{H} = 1 - H$.

In Cox's proportional hazards model, $\bar{F}^\theta = \bar{G}$ for some $\theta = \theta_0 > 0$. (Note that θ here is e^θ from the previous section, and θ_0 is the true parameter determining all sampling distributions.) The partial likelihood for θ is

$$L_{\mathcal{P},n}(\theta) = \prod_{i=1}^{m+n} \frac{\theta^{Z_i}}{\sum_{j=1}^{m+n} \theta^{Z_j} 1[T_j \geq T_i]}$$

and the weighted partial likelihood is

$$\tilde{L}_{\mathcal{P},n}(\theta) = \prod_{i=1}^{m+n} \left(\frac{\theta^{Z_i}}{\sum_{j=1}^{m+n} \theta^{Z_j} 1[T_j \geq T_i]} \right)^{w_{n,i}}$$

where $w_n = (w_{n,1}, \cdots, w_{n,m+n})$ is a vector of uniform Dirichlet weights, times $m+n$. We denote by $\hat{\theta}_n$ the parameter maximizing $L_{\mathcal{P},n}(\theta)$ and $\tilde{\theta}_n$ the parameter maximizing the weighted version $\tilde{L}_{\mathcal{P},n}(\theta)$.

The next two theorems are the main results of this chapter. Let D denote two infinite sequences of data; $D = (X_1, X_2, \cdots, Y_1, Y_2, \cdots)$ (with sample sizes m and n determined so that the limiting sampling fraction $\lambda = \lim_n(m/n)$ is between 0 and 1).

Theorem 9 *For almost every sequence D of data, $\tilde{\theta}_n$ is conditionally consistent in the strong sense. That is, for all $\epsilon > 0$,*

$$P\left(|\tilde{\theta}_n - \theta_0| > \epsilon \, i.o. \, |D\right) = 0 \quad a.s.[F, \theta_0]$$

Theorem 10 *For almost every sequence D of data, $\tilde{\theta}_n$ is conditionally efficient. That is for all $t \in \mathbf{R}$,*

$$P\left(\sqrt{m+n}(\tilde{\theta}_n - \hat{\theta}_n) \leq t \, |D\right) \to \Phi\left\{t\sqrt{I(\theta_0)}\right\} \quad a.s.[F, \theta_0]$$

where

$$I(\theta) = \frac{\lambda\bar{\lambda}}{\theta^2} \int_0^1 \frac{1}{\lambda + \theta\bar{\lambda}u^{1-1/\theta}} \, du \, .$$

The conditional probability $P(\cdot|D)$ is a version of the conditional probability $P(\cdot|\sigma(D))$ in which the data are fixed at their observed values. Under the model, all the data are governed by F and θ_0, hence the notation $a.s.[F, \theta_0]$ referring to almost every data sequence.

It is well known that the partial likelihood estimator $\hat{\theta}_n$ is strongly consistent and efficient. See for example, Tsiatis (1981), Anderson and Gill (198?), or Bickel *et al.* (1991). Thus the above theorems are tantamount to first order correctness of the weighted, partial likelihood bootstrap. Roughly speaking, this bootstrap approximates the mean and the variance of the estimator $\hat{\theta}_n$. To prove these results, we must first establish certain properties of the partial likelihood function itself.

4.4 *Properties of the partial likelihood*

4.4.1 The score function

We define the score function $S_n(\theta)$ to be the normalized derivative of the log of $L_{\mathcal{P},n}(\theta)$. We have

$$\begin{aligned} S_n(\theta) &= \frac{1}{m+n}\frac{\partial \log L_{\mathcal{P},n}(\theta)}{\partial \theta} \\ &= \frac{1}{\theta}\left(\bar{\lambda}_n - \lambda_n \int \frac{\theta\bar{\lambda}_n\bar{G}_n}{\theta\bar{\lambda}_n\bar{G}_n + \lambda_n\bar{F}_m}\, dF_m - \bar{\lambda}_n \int \frac{\theta\bar{\lambda}_n\bar{G}_n}{\theta\bar{\lambda}_n\bar{G}_n + \lambda_n\bar{F}_m}\, dG_n\right), \end{aligned}$$

these integrals being summations over observed data values. In fact, integrals of this type abound in what follows and so the next lemma is particularly useful.

Lemma 9 *As $n \to \infty$, the following convergence statements hold a.s.$[F, \theta_0]$ for $p = 1, 2$:*

$$\int \left(\frac{\theta\bar{\lambda}_n\bar{G}_n}{\theta\bar{\lambda}_n\bar{G}_n + \lambda_n\bar{F}_m}\right)^p dG_n \to \int \left(\frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}}\right)^p dG$$
$$\int \left(\frac{\theta\bar{\lambda}_n\bar{G}_n}{\theta\bar{\lambda}_n\bar{G}_n + \lambda_n\bar{F}_m}\right)^p dF_m \to \int \left(\frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}}\right)^p dF\,.$$

Furthermore, these results also hold if θ on the left hand side is replaced by a fixed sequence (θ_n) converging to θ.

PROOF. Consider the first statement above with $p = 1$. Let K_n be the left side and K the limit on the right side. By the triangle inequality

$$|K_n - K| \le \int \left|\frac{a_n}{a_n + b_n} - \frac{a}{a+b}\right| dF_m + \left|\int \frac{a}{a+b}\, dF - \int \frac{a}{a+b}\, dF_m\right| \qquad (4.1)$$

where, for simplicity, $a_n = \theta\bar{\lambda}_n\bar{G}_n$, $a = \theta\bar{\lambda}\bar{G}$, $b_n = \lambda_n\bar{F}_m$, and $b = \lambda\bar{F}$. Although it is suppressed in the notation, all four of these quantities are functions of a real variable being integrated out. Because $K < \infty$, the second term in equation (4.1) is converging to 0 by the strong law of large numbers, *a.s.*$[F, \theta_0]$. It suffices, therefore, to prove that the first term in equation (4.1) also converges to 0 in this sense. We use an *ϵ-squeezing* argument and the Glivenko–Cantelli Theorem to do so.

For any Borel set A of non-negative, real numbers, define

$$Q_n(A) = \int_A \left| \frac{a_n}{a_n + b_n} - \frac{a}{a+b} \right| dF_m \,.$$

Given any $\epsilon > 0$, we want to find a set A such that

$$Q_n(A) + Q_n(A^c) < \epsilon$$

for large enough n and $a.s.[F, \theta_0]$. Note that we may insist on $\epsilon < 1$ since $Q_n(\mathbb{R}^+) \leq 1$. The idea is that we can make the integrand small on A, and we can bound it on A^c having small probability.

A good choice for A is the set

$$A = \{x : \; F(x) < 1 - \epsilon/4 \}.$$

Note that on A, $a > \delta_1(\epsilon) := \theta\bar{\lambda}\bar{G}[F^{-1}(1-\epsilon/4)]$ and $b > \delta_2(\epsilon) := \lambda\epsilon/4$. These lower bounds δ_1 and δ_2 go to zero as ϵ goes to 0. Consider $Q_n(A)$.

$$Q_n(A) \leq \int_A \left| 1 - \frac{a}{a_n} \frac{a+b}{a_n + b_n} \right| dF_m \,.$$

We avoid the problem of $a_n = 0$ somewhere on A by assuming that n is large enough so that

$$||a_n - a|| < \frac{\epsilon}{6}\delta_1(\epsilon) \,. \tag{4.2}$$

This is possible, almost surely, by the Glivenko–Cantelli Theorem and by the fact that $\lambda_n \to \lambda$. By (4.2), $a_n > \delta_1(\epsilon)(1 - \epsilon/6) > 0$ on A. In fact, a/a_n is close to 1 on A:

$$\frac{1}{1+\epsilon/6} < \frac{1}{1+(\epsilon\delta_1)/(6a)} = \frac{a}{a+\epsilon\delta_1/6} < \frac{a}{a_n} < \frac{a}{a-\epsilon\delta_1/6} < \frac{1}{1-(\epsilon\delta)/(6a)} < \frac{1}{1-\epsilon/6}$$

which gives

$$1 - \frac{7\epsilon}{36} < \frac{a}{a_n} < 1 + \frac{7\epsilon}{36} \tag{4.3}$$

on A for large enough n, almost surely. It is slightly easier to bound $(a_n + b_n)/(a+b)$ close to 1 on A. Again by Glivenko-Cantelli, for n large enough

$$||b_n - b|| < \frac{\epsilon}{6}\delta_2(\epsilon)$$

almost surely, and so

$$1 - \epsilon/6 < \frac{a_n + b_n}{a + b} < 1 + \epsilon/6 . \tag{4.4}$$

It follows from equations (4.3) and (4.4) that for sufficiently large n and almost surely, $Q_n(A) < \epsilon/2$.

Now we turn to $Q_n(A^c)$. By definition of A,

$$\begin{aligned} Q_n(A^c) &= \int_{F^{-1}(1-\epsilon/4)}^{\infty} \left| \frac{a_n}{a_n + b_n} - \frac{a}{a+b} \right| dF_m \\ &\leq 1 - F_m\{F^{-1}(1-\epsilon/4)\} \\ &< 1 - F\{F^{-1}(1-\epsilon/4)\} + \epsilon/4 \quad \text{for large enough } n \\ &= \epsilon/2 \end{aligned}$$

completing the proof of the first statement in the lemma for $p = 1$. Proof of the second statement for $p = 1$ is exactly the same with integrations with respect to G_n replacing F_m. The result for $p = 2$ is also immediate since the errors must be smaller than double those when $p = 1$. The generalization to having a sequence θ_n converging to θ is also immediate by inspection of the ϵ-squeezing argument. □

Using Lemma 9, we get a number of results about the score function introduced at the beginning of this section.

Lemma 10 *For fixed $\theta > 0$, the following convergence statement holds a.s.$[F, \theta_0]$ as $n \to \infty$:*

$$S_n(\theta) \to S(\theta) := \frac{\theta_0 - \theta}{\theta} \int \frac{(\theta\bar{\lambda}\bar{G})(\lambda\bar{F})}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}} \, d\Lambda_F$$

where $\Lambda_F(x)$ is the cumulative hazard function $\int_0^x dF/\bar{F}$.

PROOF. By Lemma 9 and the definition of $S_n(\theta)$, $S_n(\theta)$ converges almost surely to

$$S(\theta) = \frac{1}{\theta} \left(\bar{\lambda} - \lambda \int \frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}} \, dF - \bar{\lambda} \int \frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}} \, dG \right) .$$

Also,

$$\int_0^1 \theta u^{\theta-1} \, du = 1 \Rightarrow \int_0^{\infty} \theta \bar{F}^{\theta-1} \, dF = 1$$

which further implies that $\int \bar{G}\, d\Lambda_F = 1/\theta_0$ by the proportional hazards model. Therefore,

$$S(\theta) = \frac{1}{\theta}\left(\int \theta_0 \bar{\lambda}\bar{G}\, d\Lambda_F - \lambda \int \frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}}\, dF - \bar{\lambda}\int \frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}}\, dG\right).$$

Noting that $\bar{F}\, dG = \theta_0 \bar{G}\, dF$ leads us to the result. □

We turn now to information-like quantities.

4.4.2 Information functions

The score function defined in the previous section can be written

$$S_n(\theta) = \frac{1}{n+m}\sum_{i=1}^{n+m} c_{n,i}(\theta)$$

where

$$c_{n,i}(\theta) = \frac{1}{\theta}\left(Z_i - \frac{\theta\bar{\lambda}_n \bar{G}_n(T_i)}{\theta\bar{\lambda}_n\bar{G}_n(T_i) + \lambda_n F_m(T_i)}\right).$$

We define three *information functions*, $I_{1,n}$, $I_{2,n}$, and I by

$$I_{1,n}(\theta) = \frac{1}{m+n}\sum_{i=1}^{m+n}[c_{n,i}(\theta)]^2$$

$$I_{2,n}(\theta) = -\frac{1}{m+n}\sum_{i=1}^{m+n}\frac{\partial c_{n,i}(\theta)}{\partial\theta}$$

$$I(\theta) = \frac{\lambda\bar{\lambda}}{\theta^2}\int_0^1 \frac{1}{\lambda + \theta\bar{\lambda}u^{1-1/\theta}}\, du.$$

Lemma 11 *If (θ_n) is a sequence which converges a.s.$[F, \theta_0]$ to θ_0 as $n \to \infty$, then $I_{1,n}(\theta_n)$ and $I_{2,n}(\theta_n)$ both converge to $I(\theta_0)$ in the same sense.*

PROOF. By Lemma 9, $I_{1,n}(\theta_n)$ converges almost surely to $I_1(\theta_0)$ satisfying

$$\theta^2 I_1(\theta) = \bar{\lambda} - 2\bar{\lambda}\int \frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}}\, dG + \bar{\lambda}\int\left(\frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}}\right)^2 dG + \lambda\int\left(\frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G} + \lambda\bar{F}}\right)^2 dF.$$

Similarly, $I_{2,n}(\theta_n)$ converges to $I_2(\theta_0)$ satisfying

$$\theta_2 I_2(\theta) = \bar{\lambda} - \bar{\lambda}\int\left(\frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G}+\lambda\bar{F}}\right)^2 dG - \lambda\int\left(\frac{\theta\bar{\lambda}\bar{G}}{\theta\bar{\lambda}\bar{G}+\lambda\bar{F}}\right)^2 dF.$$

It suffices to show that I_1, I_2, and I all agree at θ_0, and to do so is simply a matter of basic algebra and calculus. To simplify notation, we put $\alpha = \theta_0\bar{\lambda}\bar{G}$ and $\beta = \lambda\bar{F}$. Using $\bar{F}\,dG = \theta_0\bar{G}\,dF$, and the definition of Λ_F, we get

$$\begin{aligned}\theta_0^2 I_1(\theta_0) &= \int\left(\alpha - \frac{2\alpha^2}{\alpha+\beta} + \frac{\alpha^2}{\alpha+\beta}\right) d\Lambda_F \\ &= \int\frac{\alpha\beta}{\alpha+\beta}\,d\Lambda_F\end{aligned}$$

and

$$\begin{aligned}\theta_0^2 I_2(\theta_0) &= \int\left(\alpha - \alpha\left(\frac{\alpha}{\alpha+\beta}\right)^2 - \beta\left(\frac{\alpha}{\alpha+\beta}\right)^2\right) d\Lambda_F \\ &= \int\frac{\alpha\beta}{\alpha+\beta}\,d\Lambda_F\,.\end{aligned}$$

The result follows immediately. □

4.5 The weighted score function

Just as the score function $S_n(\theta)$ was defined, we define the weighted score function $\tilde{S}_n(\theta)$ to be the normalized derivative of the log of $\tilde{L}_{\mathcal{P},n}(\theta)$. We have

$$\begin{aligned}\tilde{S}_n(\theta) &= \frac{1}{m+n}\frac{\partial\log\tilde{L}_n(\theta)}{\partial\theta} \\ &= \frac{1}{m+n}\sum_{i=1}^{m+n} w_{n,i}c_{n,i}(\theta)\end{aligned}$$

where, again, the weights $(w_{n,1},\cdots,w_{n,m+n})$ are distributed as $m+n$ times a uniform Dirichlet random vector. Specifically, if $U_1,\cdots,U_{m+n}$ are iid exponential random variables with average $\bar{U}_n$, then we may take $w_{n,i} = U_i/\bar{U}_n$.

Of interest is the conditional distribution of $\tilde{S}_n$ given a fixed set of data. We frequently refer to a set D which represents two infinite sequences of data for which the

sampling fraction of X's to Y's stays between 0 and 1: $D = (X_1, X_2, \cdots, Y_1, Y_2, \cdots)$. This infinite sequence of data is governed by F and θ_0, whereas the weights $w_{n,i}$ are governed by $\tilde{P}$, say. We are careful to make explicit the various forms of convergence and to which probability measures they refer.

The next two results are precursors to a proof on strong conditional consistency of $\tilde{\theta}_n$.

Lemma 12 *For data D in a set of probability 1 (with respect to F, θ_0), and for large enough n, there exists $\tilde{\theta}_n$ for which $\tilde{S}_n(\tilde{\theta}_n) = 0$ and at which the weighted partial likelihood function attains its global maximum. Weights ensuring this live in a set of probability 1 (with respect to $\tilde{P}$).*

PROOF. To show existence of a zero of $\tilde{S}_n(\theta)$, note that $h_n(\theta) := \theta \tilde{S}_n(\theta)$ can be written $h_n(\theta) = a_n - b_n(\theta)$. The point a_n is in between 0 and 1, and b_n increases from 0 to 1 as θ increases from 0. (More specifically, $b_n(\theta)$ increases as long as not all weights are 0, and at least one of $\bar{G}(T_i)$ is positive. Thus the caveat, *almost surely.*) Consequently, h_n has a zero at some positive value and therefore $\tilde{S}_n(\theta)$ also has a zero, called $\tilde{\theta}_n$.

To show uniqueness of the zero, note that $\tilde{\theta}_n$ must satisfy

$$\sum_{i=1}^{m+n} w_{n,i} Z_i = \sum_{i=1}^{m+n} w_{n,i} \frac{\tilde{\theta}_n \bar{\lambda}_n \bar{G}_n(T_i)}{\tilde{\theta}_n \bar{\lambda}_n \bar{G}_n(T_i) + \lambda_n \bar{F}_n(T_i)}.$$

Using this, we write

$$\tilde{S}_n(\theta) = (\tilde{\theta}_n - \theta) K_n(\theta)$$

where $K_n(\theta) > 0$ for all $\theta > 0$ (notwithstanding the almost sure restriction). By a standard argument, we see that $\tilde{\theta}_n$ globally maximizes $\tilde{L}_{\mathcal{P},n}(\theta)$. □

Like the score function (Lemma 10), the weighted score function coverges to $S(\theta)$.

Lemma 13 *For data D in a set of probability 1 (with respect to F, θ_0), and for fixed $\theta > 0$, as $n \to \infty$*

$$\tilde{S}_n(\theta) \to S(\theta) \quad a.s.[\tilde{P}]$$

where $S(\theta)$ is the same as in Lemma 10.

PROOF. We have

$$\begin{aligned}\tilde{S}_n(\theta) - S(\theta) &= \frac{1}{m+n}\sum_{i=1}^{m+n}(w_{n,i}-1)c_{n,i}(\theta) \\ &= \frac{1}{(m+n)\bar{U}_n}\sum_{i=1}^{m+n}(U_i-\bar{U}_n)c_{n,i}(\theta) \\ &= \frac{1}{(m+n)\bar{U}_n}\sum_{i=1}^{m+n}(U_i-1)c_{n,i}(\theta) + (\bar{U}_n-1)\frac{1}{m+n}\sum_{i=1}^{m+n}c_{n,i}(\theta)\end{aligned}$$

where, as defined earlier, U_i are iid exponentials (with mean 1) and $\bar{U}_n$ is their average. Note that $\bar{U}_n$ is converging to 1 with $\tilde{P}$ probability 1 by the strong law. Since the average of the $c_{n,i}(\theta)$ is converging to $S(\theta) < \infty$ by Lemma 10, the second term above is converging in that sense to 0. It suffices therefore to show strong convergence to 0 of the first term above. To do so, we invoke Chow's (1966) strong law (see Appendix A) to $\sum(U_i - 1)a_{n,i}$ where

$$a_{n,i} = \frac{c_{n,i}(\theta)}{\sqrt{\sum[c_{n,i}(\theta)]^2}}.$$

We use Lemma 11 on the average of $[c_{n,i}(\theta)]^2$. □

4.6 *Proof of strong conditional consistency*

Using the tools forged in the previous sections, we are now ready to prove Theorem 9 on the consistency of $\tilde{\theta}_n$.

PROOF. Let $0 < \epsilon < \theta_0$ be given. First of all, by Lemma 13 and the definition of $S(\theta)$,

$$\tilde{S}_n(\theta_0 - \epsilon) \to S(\theta_0 - \epsilon) > 0$$

and

$$\tilde{S}_n(\theta_0 + \epsilon) \to S(\theta_0 + \epsilon) < 0$$

as $n \to \infty$. These convergence statements hold for almost every data set D and almost every sequence of weights (w_n). Therefore, the set A_n defined

$$A_n = \{w_n : \tilde{S}_n(\theta_0 - \epsilon) > 0\} \cap \{w_n : \tilde{S}_n(\theta_0 + \epsilon) < 0\}$$

is such that

$$P(A_n^c \; i.o. \; |D) = 0 \quad a.s.[F, \theta_0] \, .$$

Now by Lemma 12, any weight vector $w_n \in A_n$ is such that $|\tilde{\theta}_n - \theta_0| < \epsilon$, and so the result necessarily follows. □

4.7 *Proof of conditional efficiency*

Below is a proof of Theorem 10.

PROOF. A one term Taylor series expansion of the weighted score function gives

$$\sqrt{m+n}\,(\tilde{\theta}_n - \hat{\theta}_n) = \sqrt{m+n}\,\tilde{S}_n(\hat{\theta}_n) \Big/ \left(-\frac{\partial \tilde{S}_n}{\partial \theta}(\theta_n^*)\right) \tag{4.5}$$

for some point θ_n^* in between $\hat{\theta}_n$ and $\tilde{\theta}_n$. We consider the numerator and denominator separately. First, the denominator:

$$\begin{aligned} \frac{\partial \tilde{S}_n}{\partial \theta}(\theta_n^*) &= \frac{1}{m+n} \sum_{i=1}^{m+n} w_{n,i} \frac{\partial c_{n,i}}{\partial \theta}(\theta_n^*) \\ &= \frac{1}{\bar{U}_n(m+n)} \sum_{i=1}^{n} (U_i - 1) \frac{\partial c_{n,i}}{\partial \theta}(\theta_n^*) - \frac{1}{\bar{U}_n} I_{2,n}(\theta_n^*) \, . \end{aligned} \tag{4.6}$$

Here $I_{2,n}$ is as given in Lemma 11 and the U_i are iid unit exponentials with mean $\bar{U}_n$. Since θ_n^* is in between $\tilde{\theta}_n$ and $\hat{\theta}_n$, it is converging with $\tilde{P}$ probability 1 to θ_0 for almost every data sequence D. Applying Lemma 11, we see that the second term in equation (4.6) is converging to $I(\theta_0)$. It suffices, therefore, to show that the first term in equation (4.6) is converging to 0. Using Lemma 9 again, we can show that in the correct almost sure sense,

$$\frac{1}{m+n} \sum_{i=1}^{m+n} \left(\frac{\partial c_{n,i}}{\partial \theta}(\theta_n^*) \right)^2 \to \text{constant} \; < \infty \, .$$

Therefore, Chow's result (see Appendix A) shows that the first term in equation (4.6) does indeed coverge to 0.

Turning now to the numerator of equation (4.5), we have

$$\begin{aligned}\sqrt{m+n}\,\tilde{S}_n(\hat{\theta}_n) &= \frac{1}{\sqrt{m+n}} \sum_{i=1}^{m+n} w_{n,i} c_{n,i}(\hat{\theta}_n) \\ &= \frac{1}{\bar{U}_n \sqrt{m+n}} \sum_{i=1}^{m+n} (U_i - 1) c_{n,i}(\hat{\theta}_n) \\ &= \frac{1}{\bar{U}_n} \sqrt{I_{1,n}(\hat{\theta}_n)} \sum_{i=1}^{m+n} (U_i - 1) a_{n,i}\end{aligned}$$

where $I_{1,n}$ is as in Lemma 11 and the constants $a_{n,i}$ satisfy $\sum_i a_{n,i}^2 = 1$. By the Central Limit Theorem in Appendix A,

$$\sum_{i=1}^{m+n} (U_i - 1) a_{n,i} \xrightarrow{\mathcal{D}} N(0,1)$$

for almost every sequence of data D; this distribution being conditional on the data and thus on $a_{n,i}$. We need only verify the asymptotic negligibility condition

$$r_n := \left(\sum [c_{n,i}(\hat{\theta}_n)]^2\right)^{-1} \max_{1 \le i \le (m+n)} \left[c_{n,i}(\hat{\theta}_n)\right]^2 \to 0 \tag{4.7}$$

as $n \to \infty$ and for almost every sequence of data D. Restriction (4.7) is true because

$$r_n \le \frac{1}{m+n} \left(I_{1,n}(\hat{\theta}_n)\right)^{-1}.$$

Now using Slutsky's Theorem and Lemma 11, we see that conditionally, the numerator converges to a mean 0 normal variable with variance $I(\theta_0)$. Combining results for the numerator and denominator in (4.5), the theorem follows from another application of Slutsky's Theorem. □

4.8 Discussion

The close analogy between Efron's nonparametric bootstrap and the weighted likelihood bootstrap does not hold in this extension to partial likelihood. For example, if we tried the standard nonparametric bootstrap here, then the risk set would change with every bootstrap sample, unlike for the weighted partial likelihood. Moreover, we would add the complication of ties in the data. A bootstrapping procedure has been developed for the proportional hazards model; see Hjort (1989). Hjort's procedure

is truly in the spririt of Efron's bootstrap, in that it involves simulation from the fitted statistical model. It is in fact a semiparametric bootstrap, and so the partial likelihood of a bootstrap sample is not simply the original partial likelihood with factors raised to multinomial powers. Hjort does claim first order consistency of his bootstrap, although this is not rigourously proved. His bootstrap and the WLB, while both consistent, are structurally quite different and it is not clear what higher order properties either procedure possesses.

Chapter 5

A BOOTSTRAP RECYCLING ALGORITHM FOR PREPIVOTING

5.1 Introduction

In frequentist inference problems, one often compares the observed value of a random variable to a quantile of that variable's estimated distribution. Here, such a random variable is called a *root* of the inference problem, extending the definition in Beran (1987). A root which has the same distribution regardless of the distribution generating the data is a pivotal quantity and inference based upon it is exact in the frequency sense. When the distribution of the root is not invariant however, such inference is only approximate. That is, the actual level of a confidence set or size of a test differs from the nominal level or size. The method of prepivoting introduced by Beran (1987, 1988) can reduce errors in these inference problems. To prepivot, one bases inference upon the new root obtained by transforming the original root via its estimated cumulative distribution function. The prepivoting operation can be iterated to further reduce errors.

While prepivoting provides an elegant solution in principle, it can be difficult to implement in practice. The iterated-bootstrap algorithm suggested by Beran (see also Loh, 1987) uses an exponentially growing number of bootstrap samples as the number of iterations of prepivoting increases. In this chapter, we introduce a *bootstrap recycle* algorithm to implement iterated prepivoting. This Monte Carlo algorithm is motivated by the fact that a root obtained by prepivoting is actually an expectation with respect to a measure fitted from a bootstrap sample. Different roots are expectations with respect to different estimated measures. Using a simple relation, we express all roots as expectations with respect to a common measure, which need not be in the model. Consequently, we obtain approximations to the distributions of prepivoted roots such that the number of bootstrap samples required for simulation-consistency grows linearly with the number of iterations of prepivoting.

Like the iterated-bootstrap algorithm, the bootstrap recycle algorithm involves

a number of levels or stages of bootstrap sampling. The first stage is the familiar simulation of data sets from the fitted model and calculation of initial roots. The algorithms differ at subsequent stages of sampling. In the bootstrap recycle algorithm, each bootstrap sample beyond the first stage is recycled to take the place of a sample that would be required in an iterated-bootstrap algorithm. To compensate for this recycling, certain weighting functions are invoked to yield estimates of the desired distributions. The price paid for recycling bootstrap samples is dependence between the terms averaged in our estimates. This price is not too great in one sense because these estimates are simulation-consistent; they converge to the right answers as the amount of computing resources gets large.

5.2 Bootstrap algorithms and prepivoting

Suppose that we model a random variable x by a family $\mathcal{P}$ of probability distributions on the measurable space $(\mathbf{X}, \mathcal{A})$. The variable x may represent any sort of data: a random sample of size n from some population, a time series, or data with more complex dependency structure. It can have an unobservable component, which may not be data in the traditional sense but rather parameters as in an empirical Bayes, or random-effects model. The unobservable component might also be the future value of a time series in a prediction problem. (Straying somewhat from conventional notation, we do not distinguish between random variables and their realizations.) Also, the model $\mathcal{P}$ can be quite general. It may or may not be parameterized by a finite-dimensional space, for example. We suppose that for any $x \in \mathbf{X}$, there is an *estimate* $P_x \in \mathcal{P}$ of the point $P \in \mathcal{P}$ which is presumed to have generated x. For definiteness, we suppose that this estimate depends only on the observable part of x when this is relevant.

In performing hypothesis tests and in constructing confidence sets for some parameter $\theta = T(P)$, one often compares the observed value of a real-valued, random variable to a quantile of that variable's estimated distribution. Similar comparisons are made when doing inference on an unobservable component x_u of x. Such a random variable at the base of the inference problem is called a *root*, and is typically denoted by $R(x; P)$. The notation is meant to convey several bits of information, and so to avoid excessive notation or confusion, we adhere to the following conventions throughout:

1. The root $R(x; P)$ can be functionally dependent on the observable part of x and on the unknown object of interest (parameter or unobserved random variable). It cannot be functionally dependent on unknown quantities which are not the focus of inference.

2. The random variable in the first position has a distribution given by the fixed element in the second position *unless explicitly stated otherwise.* For example, $y \sim P_x$ given x in $R(y; P_x)$.

This general framework seems to encompass the domain of earlier works on bootstrapping. Beran (1987, 1988) is restricted to inference about fixed parameters, while Efron (1987) has hinted at the more general set up in discussion of empirical Bayes models. In Section 5.4, we illustrate these concepts with three examples.

Generally, the root $R(x; P)$ has left-continuous distribution function

$$H(u, P) = P\left(R(x; P) < u\right) .$$

Unless $R(x; P)$ is a pivotal quantity, $H(\cdot, P)$ really depends on P, the unknown distribution assumed to generate x. Inference proceeds, nonetheless, along at least one of two possible avenues. Firstly, the root $R(x; P)$ may have a tractable asymptotic distribution which may be known, or may depend on a parameter which can be estimated. Alternatively, $H(\cdot, P)$ can be estimated by its bootstrap distribution

$$H(u, P_x) = P_x\left(\, R(y; P_x) < u \,|\, x \,\right) .$$

Here we are taking the broad definition of bootstrapping as inference under the fitted statistical model. Typically, this bootstrap distribution can be approximated arbitrarily well by simulation using the following algorithm:

Bootstrap algorithm:

1. Generate bootstrap samples

$$y^1, y^2, \cdots y^m \sim_{iid} P_x \text{ given } x .$$

2. Compute $a_i = R(y^i; P_x)$ for each i.

3. Approximate $H(\cdot, P_x)$ by the empirical distribution

$$\hat{H}(u, P_x) = \frac{1}{m} \sum_{i=1}^{m} 1\left[a_i < u\right] .$$

Note that the samples above are generated under the fitted model P_x. As we shall see in the next section, these samples can be generated by any measure which dominates P_x. As an aside, sometimes we cannot generate iid samples, but we can construct a Markov chain on $\mathbf{X}$ having the right invariant distribution: P_x. Metropolis-Hastings algorithms are examples.

Since $H(\cdot, P_x)$ is generally different from the desired distribution $H(\cdot, P)$, the inference using this bootstrap algorithm is in error no matter how many bootstrap samples are generated. That is, the actual level of the confidence set or the actual size of the test is different from the nominal level or size. To reduce this error, Beran suggests prepivoting. Define a new root $R_1(x; P)$ by transforming the original root via its estimated cumulative distribution function as follows:

$$R_1(x; P) = H\left\{R(x; P), P_x\right\} . \tag{5.1}$$

This root, which is a monotone, data-dependent transformation of the original root, has left-continuous distribution function $H_1(\cdot, P)$.

In a relatively small class of problems, the prepivoted root $R_1(x; P)$ is exactly pivotal; indeed the name prepivoting is meant to suggest that R_1 is closer to being pivotal than R. If, for instance, the original root is pivotal and has a continuous distribution function, then $R_1(x; P)$ is uniformly distributed on $(0, 1)$ by the probability integral transformation. The prepivoted root can be pivotal even if the original root is not. For example, suppose that $x = (x_1, x_2, \cdots, x_n)$ is a random sample of uniform random variables on the interval $(0, \theta)$ and that inference about θ is of interest. A natural initial root is

$$R(x; P) = n(\theta - x_{(n)}), \qquad \text{where } x_{(n)} = \max_i x_i .$$

By a straightforward calculation,

$$H(u, P) = 1 - \left(1 - \frac{u}{n\theta}\right)^n$$

and with P_x determined by $\hat{\theta}(x) = x_{(n)}$, the maximum likelihood estimate, the prepivoted root is

$$\begin{aligned} R_1(x;P) &:= H\left(R(x;P), P_x\right) \\ &= 1 - \left(1 - \frac{\theta - x_{(n)}}{x_{(n)}}\right)^n . \end{aligned}$$

Wheras the initial root is not pivotal in this example, the prepivoted root is; its distribution depends only on $\mathcal{P}$, not a particular $P \in \mathcal{P}$.

Now $R_1(x;P)$ is generally not pivotal but its distribution can be estimated by the bootstrap distribution

$$\begin{aligned} H_1(u, P_x) &= P_x \{\, R_1(y;P_x) < u \,|\, x \,\} \\ &= P_x \{\, P_y [\, R(z;P_y) < R(y;P_x) \,|\, x, y \,] < u \,|\, x \,\} . \end{aligned} \tag{5.2}$$

The iterated bootstrap algorithm given below (from Beran 1987) can be used to approximate $H_1(\cdot, P_x)$ by simulation:

Iterated bootstrap algorithm:

1. Generate first stage bootstrap samples

$$y^1, y^2, \cdots, y^m \sim_{iid} P_x \text{ given } x .$$

2. Compute the initial roots $b_i = R(y^i; P_x)$ for all i.

3. For each i generate second stage bootstrap samples

$$z^{i,1}, z^{i,2}, \cdots, z^{i,n} \sim_{iid} P_{y^i} \text{ given } y^i .$$

4. Compute the initial roots $a_{i,j} = R(z^{i,j}; P_{y^i})$ for all i and j.

5. For each i compute

$$\hat{R}_1(y^i; P_x) = \frac{1}{n} \sum_{j=1}^{n} 1\,[a_{i,j} < b_i] .$$

6. Approximate $H_1(\cdot, P_x)$ by the empirical distribution

$$\hat{H}_1(u, P_x) = \frac{1}{m}\sum_{i=1}^{m} 1\left[\hat{R}_1(y^i; P_x) < u\right] .$$

The prepivoting operation adjusts the calculations from the initial inference problem. We study its effect more closely in Section 5.4. Importantly, the prepivoting operation can be iterated. For the same reason the original bootstrap inference is in error, there may be error in the once-prepivoted inference. Applying Beran's argument to the once-prepivoted root gives the twice-prepivoted root

$$R_2(x; P) = H_1\left\{ R_1(x; P), P_x\right\}$$

with left-continuous distribution function $H_2(\cdot, P)$. Generally, we can construct a p^{th} prepivoted root

$$R_p(x; P) = H_{p-1}\left\{ R_{p-1}(x; P), P_x\right\}$$

having left-continuous distribution function $H_p(\cdot, P)$. The iterated bootstrap algorithm can be used, in principle, to approximate the distribution functions $H_p(\cdot, P_x)$ for each p. Of course, the number of bootstrap samples required to approximate these distributions then grows exponentially with p.

5.3 Bootstrap recycling

5.3.1 The algorithm

The major drawback of the iterated bootstrap algorithm is that for every first stage bootstrap sample y^i, we must generate an entire set of second stage bootstrap samples $z^{i,j}$ $j = 1, 2, \cdots, n$. In the bootstrap recycle algorithm described here, the same set of second level bootstrap samples is used with every bootstrap sample from the first level. The algorithm is based on the following simple relation: Let P_1 and P_2 be two measures on $(\mathbf{X}, \mathcal{A})$ such that $P_1 << P_2$. For an integrable function U,

$$\begin{aligned} E_1\left\{U(x)\right\} &= \int U(x)\, dP_1(x) \\ &= \int U(x)W(x)\, dP_2(x) \\ &= E_2\left\{U(x)W(x)\right\} \end{aligned}$$

where W is a version of the Radon-Nikodym derivative

$$W = \frac{dP_1}{dP_2}.$$

By averaging objects sampled under P_2, an expectation with respect to P_1 can be estimated using the relationship above. This fact is the basis of importance sampling (Hammersley and Hanscomb 1964), in which P_2 is chosen so that the variance of the subsequent estimator is as small as possible. Indeed, other researchers have used this idea in connection with bootstrapping, (Johns, 1988, Davison, 1988, Hinkley and Shi, 1989). In bootstrap recycling, it is more *convenience* sampling than *importance* sampling which drives the estimators.

Given x and an initial root $R(x;P)$, iterated prepivoting can be implemented if we have Monte Carlo approximations to the distribution functions

$$H(\cdot, P_x),\ H_1(\cdot, P_x),\ \cdots,\ H_p(\cdot, P_x).$$

Suppose that for two measures P_1 and P_2 with $P_1 << P_2$, $P_1 \in \mathcal{P}$, we can sample $z^1, z^2, \cdots, z^n \sim_{iid} P_2$ and we want to approximate $H(u, P_1)$. By the argument above, the natural estimate is

$$\hat{H}(u, P_1) = \frac{1}{n}\sum_{j=1}^{n} 1\left[R(z^j; P_1) < u\right] \frac{dP_1}{dP_2}(z^j). \tag{5.3}$$

While this gives us an estimate of the null distribution $H(u, P_x)$ (by putting $P_1 = P_x$), we need more to get the higher order distributions. Furthermore, it is typically the case that the original root $R(x;P)$ can be computed easily, but the higher order roots need to be approximated by Monte Carlo. From (5.1) and (5.3), we have, by substitution, the estimate

$$\begin{aligned}\hat{R}_1(y; P_1) &= \hat{H}\left(R(y; P_1), P_y\right) \\ &= \frac{1}{n}\sum_{j=1}^{n} 1\left[R(z^j; P_y) < R(y; P_1)\right] \frac{dP_y}{dP_2}(z^j)\end{aligned}$$

for some $y \in \mathbf{X}$. From the definition of H_1 (e.g. (5.2)), the subsequent estimate of $H_1(u, P_1)$ is

$$\hat{H}_1(u, P_1) = \frac{1}{m}\sum_{i=1}^{n} 1\left[\hat{R}_1(y^i; P_1) < u\right] \frac{dP_1}{dP_3}(y^i)$$

where $y^1, \cdots, y^m$ are sampled from some other point P_3 which dominates P_1. We are lead to the following algorithm to approximate the distribution of a once prepivoted root.

Bootstrap recycle algorithm for one prepivot (method I):

1. Generate first stage bootstrap samples

$$y^1, y^2, \cdots, y^m \sim_{iid} P_1 >> P_x \text{ given } x.$$

2. Compute the initial roots $b_i = R(y^i; P_x)$ for all i.
3. Generate second stage bootstrap samples independent of all y^i

$$z^1, z^2, \cdots, z^n \sim_{iid} P_2 >> P_x \text{ given } x.$$

4. Compute the initial roots $a_{i,j} = R(z^j; P_{y^i})$ for all i and j.
5. Construct an $m \times n$ matrix W of weights: for all i, j set

$$W(i,j) = \frac{dP_{y^i}}{dP_2}(z^j).$$

6. For each i compute

$$\hat{R}_1(y^i; P_x) = \frac{1}{n} \sum_{j=1}^{n} 1\left[a_{i,j} < b_i\right] W(i,j).$$

7. Approximate $H_1(\cdot, P_x)$ by the weighted empirical distribution

$$\hat{H}_1(u, P_x) = \frac{1}{m} \sum_{i=1}^{m} 1\left[\hat{R}_1(y^i; P_x) < u\right] \frac{dP_x}{dP_1}(y^i).$$

There are two essential differences between this algorithm and the iterated bootstrap algorithm of the previous section. Firstly, first stage samples are generated from some measure P_1 which dominates the fitted model P_x; samples need not be generated from P_x as in classical bootstrapping. Secondly and more substantially, the second stage samples z^j in recycling are all generated from the same measure P_2, instead of from a host of different measures. Importantly, the measures P_1 and P_2

need not be elements of the model $\mathcal{P}$. They must be chosen so that if $y \sim P_1$, then P_y is absolutely continuous with respect to P_2, denoted $P_y << P_2$. Indeed, an efficient algorithm may be had by forming a mixture over elements in $\mathcal{P}$.

By recycling the secondary samples with each initial sample, we achieve an enormous reduction in the number of samples that are required (from $m \times n$ to $m + n$). The price we pay is in dependence between the terms which we average to get $\hat{H}_1$. These terms are independent in the iterated bootstrap algorithm, but in the recycle algorithm, $\hat{R}_1(y^i; P_x)$ is formed from all the z^j, for each y^i. This dependence is not too strong, however, because the procedure works. In Section 5.5, we prove that this bootstrap recycling algorithm is simulation consistent. That is, given enough computing resources, $\hat{H}_1(\cdot, P_x)$ can be made arbitrarily close to $H_1(\cdot, P_x)$ with *simulation* probability one. Of course we may lose something in efficiency, but the algorithm nonetheless brings certain inaccessible problems into the realm of possibility.

The *change-of-measures* argument used above leads to variations of this basic algorithm. We do not need to generate the secondary samples (step 3) from the same P_2. It may be more convenient, for example, to set $m = n$ and to generate each z^i from P_{y^i}. This will tend to spread out the secondary samples throughout the sample space $\mathbf{X}$. The only consequence for the algorithm above is that the weight matrix W must be computed differently (step 5). Under this new sampling scheme (called **method II**), we have

$$W(i,j) = \frac{dP_{y^i}}{dP_{y^j}}(z^j)\,.$$

Different restrictions on the model are required for this **method II** recycling algorithm to work. We must have $P_{y^i} << P_{y^j}$ for all i, j. If P_2 is taken to be a mixture of P_{y^i}'s, then there may be little difference between the **method I** and **method II** algorithms. Another variation of the recycling algorithm (**method III**) can be used in exponential family models where the weights can be simplified in a special way. We elaborate on this in the next section. In fact, this third method works whenever the model densities are known up to an intractable normalizing constant.

The bootstrap recycle algorithm can be generalized to compute approximations to the distributions of the first p prepivoted roots for any $p \geq 1$. The principle is substitution: with $\hat{H}_{k-1}$ in hand, compute

$$\hat{R}_k(y^i; P_1) = \hat{H}_{k-1}\left(R_{k-1}(y^i; P_1), P_{y^i}\right)$$

and then

$$\hat{H}_k(u, P_1) = \frac{1}{m}\sum_{i=1}^{n} 1\left[\hat{R}_k(y^i; P_1) < u\right] \frac{dP_1}{dP_2}(y^i)$$

where y^i are sampled from P_2. A set of independent bootstrap samples is required at each stage. We include the general algorithm in Section 5.6 for completeness. As written, it requires on the order of $(p+1) \times m$ samples be generated (supposing that m samples are generated at each stage). It also requires on the order of $p \times m^2$ units of storage to keep track of weight matrices and intermediate roots. In fact, it is precisely for the higher order iteration that the bootstrap recycle algorithm is important. Since it requires on the order of m^2 units of storage, it may be similar in speed to the regular iterated bootstrap for one prepivot. For two or more iterations, it drastically outperforms the iterated bootstrap. Moreover, if estimation is expensive, it will be faster than the iterated bootstrap even for one prepivot.

5.3.2 Forming the weights

The Radon-Nikodym derivative dP_1/dP_2 used in the recycle algorithm may not be difficult to compute. If, as is often true in parametric models, measures P_j have densities f_j with respect to a σ–finite measure μ on $(\mathbf{X}, \mathcal{A})$ such that $f_j(x) > 0$ for all x, then

$$\frac{dP_j}{dP_k}(y) =_{a.s.} \frac{f_j(y)}{f_k(y)}.$$

Efron's original bootstrap provides another special case. Here x is an iid sample $x_1, \cdots, x_n$, and $P_1 = P_x$ is formed from the empirical distribution of the sample. Bootstrap samples y and z are simply samples with replacement from $x_1, \cdots, x_n$. It is easy to show that

$$\frac{dP_y}{dP_x}(z) = \prod_{i=1}^{n}\sum_{j=1}^{n} 1\left[y_j = z_i\right].$$

Note that **method II** does not work here because two bootstrap samples y^1, y^2 may provide different support for P_{y^1}, P_{y^2}.

For exponential families, there are two ways to construct weights for the recycle algorithm. The first is simply to use the ratio of densities described above. An

alternative, based on *exponential tilting*, is appealing because the weights have mean 1 and are computable even if the ratio of densities is not. Consider first a *standard* exponential family of densities with respect to the Borel measure μ on $(\mathbf{R}^k, \mathcal{B}^k)$:

$$f_\theta(x) = \frac{1}{c(\theta)} e^{\langle x,\theta \rangle} \quad x \in \mathbf{R}^k, \quad \theta \in \Theta \subset \mathbf{R}^k$$

and Θ nonempty. Here $\langle , \rangle$ is inner product, and $c(\theta)$ is the Laplace transform

$$c(\theta) = \int e^{\langle x,\theta \rangle} \, d\mu(x) \, .$$

Each density f_θ determines a distribution P_θ on $(\mathbf{X}, \mathcal{A})$. For two points $\theta, \psi \in \Theta$,

$$\frac{dP_\theta}{dP_\psi}(x) = \frac{c(\psi)}{c(\theta)} e^{\langle x,\theta-\psi \rangle} \, .$$

Furthermore,

$$\frac{c(\theta)}{c(\psi)} = E_\psi e^{\langle x,\theta-\psi \rangle} \, .$$

Therefore, by sampling $x^1, x^2, \cdots, x^m$ from P_ψ we can estimate dP_θ / dP_ψ by

$$\widehat{\frac{dP_\theta}{dP_\psi}}(x) = \frac{m e^{\langle x,\theta-\psi \rangle}}{\sum_i e^{\langle x^i,\theta-\psi \rangle}} \, . \tag{5.4}$$

This may be very useful, especially if $c(\theta)/c(\psi)$ is difficult to calculate analytically. A similar result holds for exponential families which are not necessarily *standard*. If densities have the form

$$f_\psi(x) = a(\psi) d(x) e^{\langle t(x),b(\psi) \rangle}$$

for some functions a, b, d, and t, then the Radon-Nikodym derivative of P_ϕ with respect to P_ψ is

$$\frac{dP_\phi}{dP_\psi}(x) = \frac{e^{\langle t(x),b(\phi) \rangle}}{E_\psi \left(e^{\langle t(x),b(\phi) \rangle} \right)} \, ,$$

so we get an estimator like that in (5.4). Note that this third method applies to any model where densities are known up to an intractable normalizing constant.

5.4 *Applications*

5.4.1 *Testing independence in a sparse table*

Suppose that a random sample of N objects are classified in a two-way rxs contingency table and the resulting matrix of counts is denoted $x = (x_{i,j})$. The sample space $\mathbf{X}$ is the collection of all $r \times s$ tables having a total of N elements. The *full* model for x is the set of possible multinomial probability matrices:

$$\mathcal{P} = \{P = (p_{i,j}) : \ 0 \leq p_{i,j} \leq 1 \ \ p_{\cdot,\cdot} = 1\} .$$

A dotted subscript means that that index has been summed over. The null hypothesis of independence is the sub-model

$$\mathcal{P}_0 = \{P \in \mathcal{P} : \ p_{i,j} = p_{i,\cdot} p_{\cdot,j}\} .$$

Under this null hypothesis, the natural estimator of P is P_x given by

$$(P_x)_{i,j} = \frac{x_{i,\cdot} x_{\cdot,j}}{N^2} .$$

The likelihood-ratio statistic is often used to test the hypothesis of independence:

$$R(x; P) = \sum_{i=1}^{r} \sum_{j=1}^{s} x_{i,j} \log \frac{x_{i,j}}{E_{i,j}}$$

where $E_{i,j} = N(P_x)_{i,j}$. For test statistics as this one above, $R(x; P)$ is functionally the same for all P, although its distribution depends on P.

When the table is sparse, the statistician has little confidence that the $H(\cdot, P)$ is close to the limiting chi-squared distribution with $(r-1) \times (s-1)$ degrees of freedom. Consider the following sparse table analyzed by Mehta and Patel, 1983, and Guo and Thompson, 1989, for example.

1	2	2	1	1	0	1
2	0	0	2	3	0	0
0	1	1	1	2	7	3
1	1	2	0	0	0	1
0	1	1	1	1	0	0

Figure (5.1) shows the results of prepivoting the likelihood-ratio statistic for this data set. The upper graph compares the distribution of $R(y; P_x)$ (the bootstrap distribution) with the limiting chi-square distribution. The lower graph shows the result of prepivoting via the bootstrap recycle algorithm (method III, normalized weights). Several aspects of these graphs are noteworthy. First, the bootstrap distribution of the likelihood ratio statistic is much different than the asymptotic approximation. This is, in effect, a diagnosis of error in the asymptotic test. Second, the asymptotic distribution of the bootstrap p-value is uniform, the solid line of the lower graph. (This follows from the Polya lemma using only that the limiting cdf of the root is continuous.) That the bootstrap distribution of the prepivoted root (as determined by recycling) is much different than uniform is a diagnosis of error in the first stage bootstrap test. Even with recycling, the lower histogram takes a long time to compute.

The prepivoting operation adjusts the calculations from the initial inference problem. The original (nominal) size α test is typically to reject when

$$R(x; P) \geq H^{-1}\{1 - \alpha, P_x\}. \tag{5.5}$$

Here, H^{-1} is the right-continuous inverse of H thus the critical value of the test is the largest $(1 - \alpha)$ quantile of the estimated distribution of $R(x; P)$. The prepivoted test is to reject when

$$R_1(x; P) \geq H_1^{-1}\{1 - \alpha, P_x\},$$

or equivalently reject when

$$R(x; P) \geq H^{-1}\left\{H_1^{-1}[1 - \alpha, P_x], P_x\right\}.$$

By prepivoting, the original test has been adjusted to account somewhat for the fact that $H(\cdot, P)$ is not known exactly. The adjustment leads to a new critical value for the original test statistic. A heuristic reason why prepivoting reduces the errors in inference is simple. Instead of naively assuming that $R_1(x; P)$ is uniformly distributed (as we do essentially in the original test considering (5.1) and (5.5)), we estimate the distribution of this new root. In the parlance of hypothesis testing, $1 - R_1(x; P)$ is the $p-$value of the original test. By prepivoting, we adjust for the fact that this $p-$value generally is not uniformly distributed.

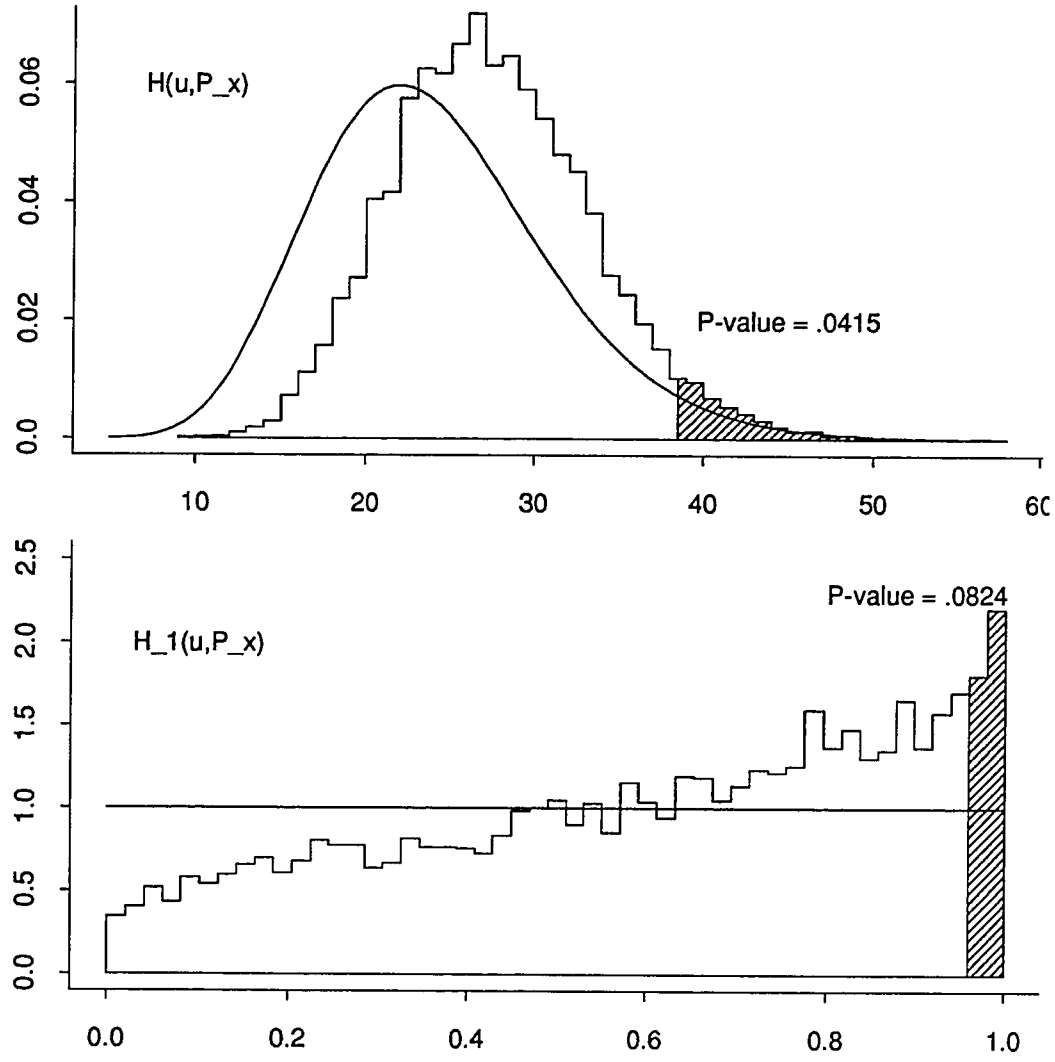


Figure 5.1: Goodness-of-fit test for a contingency table: The upper graph compares $H(\cdot, P_x)$, the bootstrap distribution of the LR statistic, to the limiting chi-square (solid line) for the example of Section 5.4.1. The bootstrap histogram is based on $1e4$ samples and yields a p-value of $.0415 \pm .002$. The lower graph shows the estimated distribution $H_1(\cdot, P_x)$ of this *naive* p-value. This bootstrap distribution is estimated by bootstrap recycling using $5e3$ first stage samples and $25e4$ second stage samples. The *prepivoted* P-value is double the naive one, at $.0824 \pm .004$. Note the p-value from the asymptotic test is .031. (CPU times: 51 seconds, 27 hours.)

We are interested in the behavior of the bootstrap recycle algorithm. In Figures (5.2) and (5.3), we examine how many second stage samples are needed to get adequate approximations $\hat{R}_1(y^i; P_x)$ of $R_1(y^i; P_x)$ for the contingency table example. To do so, we compare bootstrap recycling and the iterated bootstrap with a fixed set of 1000 first stage tables y^i. In each of Figures (5.2) and (5.3), the prepivoted roots $R_1(y^i; P_x)$ are estimated by iterated bootstrapping with $n = 1000$ (this gives $1e6$ second stage samples) and by recycling with 4 different sizes of second stage sampling. In recycling, the the second stage tables are generated from $P_2 = P_x$, in this case. We see that a lot of samples need to be generated to produce comparable estimates. Also, a lot of CPU time is used to do the recycling, and so there is no gain here in using recycling instead of iterated bootstrapping.

The reason that so many second stage samples are required in this example is because we have made a poor choice of P_2. Whatever this distribution is, it should spread its probability mass over the sample space more broadly than P_x. A natural choice for P_2 is an equal mixture over the first level fitted models:

$$P_2 = \frac{1}{m} \sum_{i=1}^{m} P_{y^i} .$$

Using this P_2, we redid the computations for prepivoting the likelihood ratio statistic, with quite dramatic results. Figures 5.4 and 5.5 compare recycling with this new P_2 to the iterated bootstrap for $m = 1000$ and various second stage sample sizes. The improvement over using $P_2 = P_x$ is astounding, as only a few thousand second stage samples give a good approximation. There is, of course, a price to be paid in using this mixture as P_2. The Radon-Nikodym weights are more difficult to compute. Nevertheless, the recycling is more efficient than the iterated bootstrap in this case. (Note that the CPU times for recycling quoted in the captions of Figures 5.4 and 5.5 are longer than optimal because both unnormalized and normalized weights are computed on the same run.)

5.4.2 Likelihood-based confidence sets

The set of parameter points having likelihood larger than a given cutoff is a natural confidence set in parametric models. For the set to have a desired nominal coverage probability, this cutoff must be chosen by appealing to an estimate of the distribution

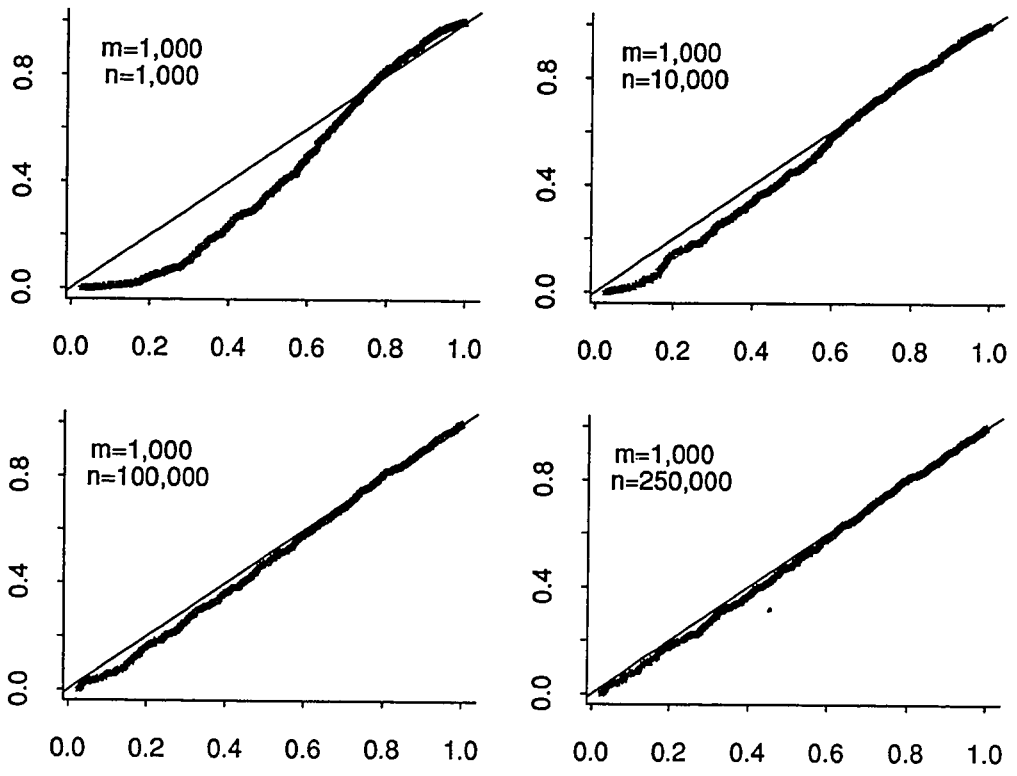


Figure 5.2: Empirical qq plots: On the horizontal axis are the ordered values of $\hat{R}_1(y^i; P_x)$ from the iterated bootstrap algorithm using 1000 by 1000 sampling (that's 1e6 second stage samples). On the vertical axis are values of $\hat{R}_1(y^i; P_x)$ from bootstrap recycling (method III, normalized weights). The same first level samples y^i are used in each case. The number of second stage samples n is shown, and increases clockwise from the upper left plot. (CPU times in minutes: 1.4, 13.3, 132.3, 330 for the recycled bootstraps and 41.6 for the iterated bootstrap.)

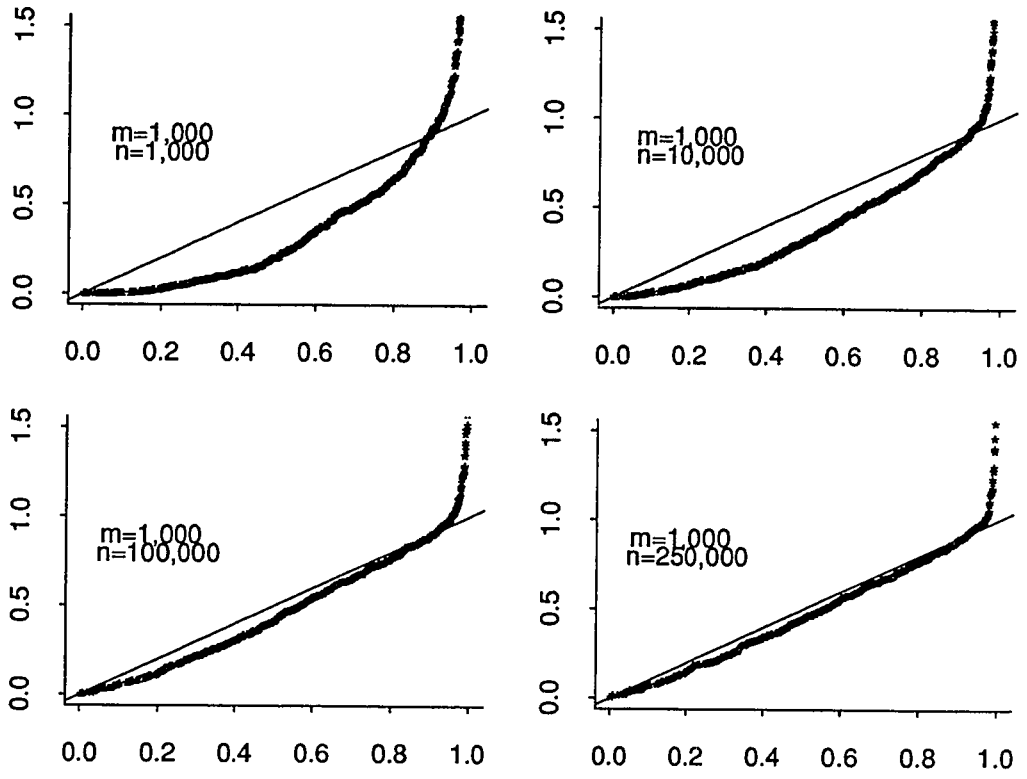


Figure 5.3: Empirical qq plots: On the horizontal axis are the ordered values of $\hat{R}_1(y^i; P_x)$ from the iterated bootstrap algorithm using 1000 by 1000 sampling. On the vertical axis are values of $\hat{R}_1(y^i; P_x)$ from bootstrap recycling (method I). The same first level samples y^i are used in each case. The number of second level samples n is shown, and increases clockwise from the upper left plot. Note that the recycle estimates are not constrained to be less than 1 when the weights are not normalized. In fact, extreme values have not been plotted. The numbers of points not shown are respectively, 62, 32, 11, and 12 out of 1000. (CPU times in minutes: 1.1, 10.6, 104.8, 261.8 for the recycled bootstraps and 41.7 for the iterated bootstrap.)

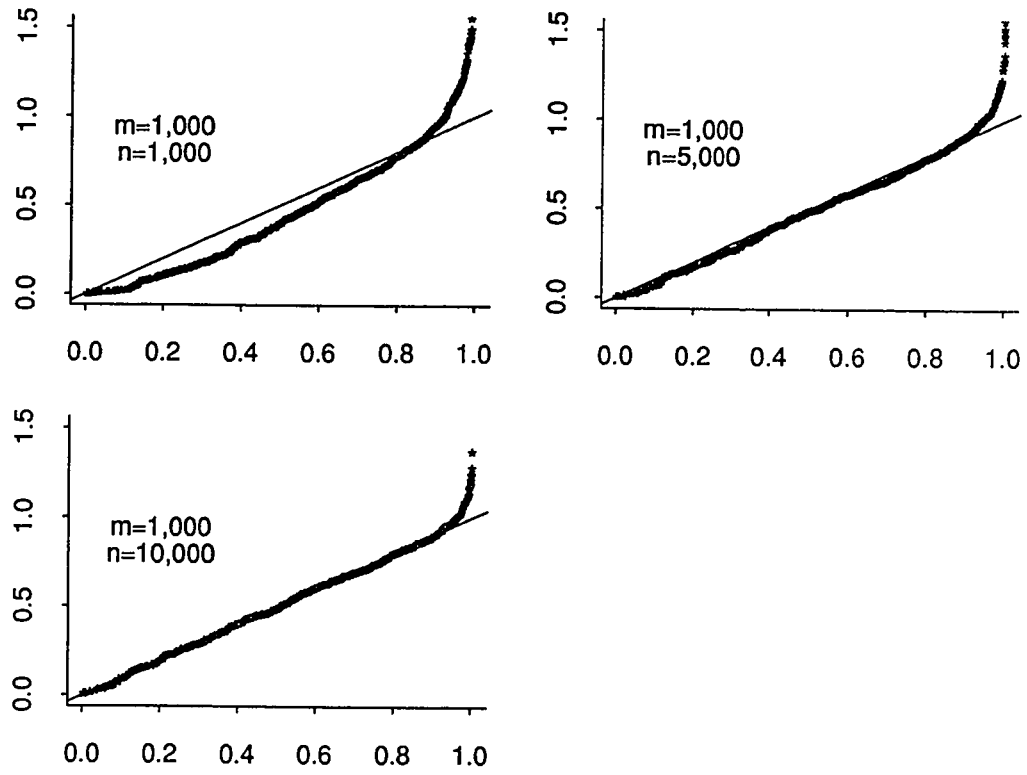


Figure 5.4: Same as Figure 5.2 except second stage samples come from a mixture over fitted models from the first stage, and weights are normalized. The number of second stage samples n is shown, and increases clockwise from the upper left plot. (CPU times in minutes: 8.5, 41.2, 87, for the recycled bootstraps and 41.7 for the iterated bootstrap.)

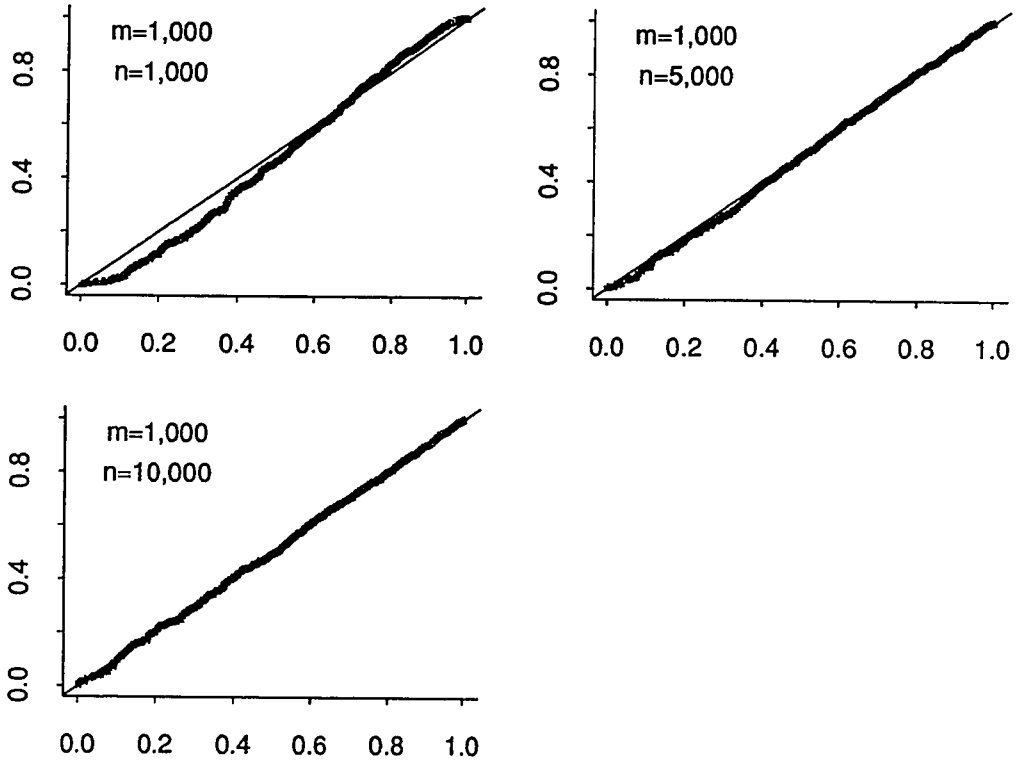


Figure 5.5: Same as Figure 5.3 except second stage samples come from a mixture over fitted models from the first stage. The number of second stage samples n is shown, and increases clockwise from the upper left plot. Note that the recycle estimates are not constrained to be less than 1 because the weights are not normalized. In fact, extreme values have not been plotted. The numbers of points not shown are respectively, 22, 3, and 0 out of 1000. (CPU times in minutes: 8.5, 41.2, 87, for the recycled bootstraps and 41.7 for the iterated bootstrap.)

of a root. In this section, the use of bootstrapping and prepivoting to calculate this cutoff is studied for a stochastic model derived and discussed in Guttorp *et al.* (1990). In this particular example, the random x has both an unobserved and an observed component. The unobserved component forms a finite-state, continuous time Markov process on the state-space

$$\{0, 1, \cdots, N\}$$

for some unknown N which represents the number of active stem cell clones in the bone marrow of a cat in the experiment. (For further explanation, see Guttorp *et al.* 1990). The unobservable process $U(t)$ has a distribution determined by N and two real parameters p and λ which determine the rate of fluctuations in $U(t)$. The observed component of x is a time series of proportions $O(t_1), O(t_2), \cdots, O(t_m)$; the $t_1, \cdots, t_m$ being experimental sampling times. Given $U(t_1), \cdots, U(t_m)$, the observed proportions are modeled as binomially distributed with fixed sample sizes $n_1, \cdots, n_m$ and success probabilities $U(t_1)/N, \cdots, U(t_m)/N$. Based on the observed proportions, there is interest in inference about the unknown parameters N, λ, p.

Figure 5.6 shows a single time series of proportions from the study of Abkowitz *et al.* (1990), which was analyzed using the methods of Guttorp *et al.* (1990) described above. As external information provided an estimate for p, inference in this case is reduced to statements about the two parameters $\theta = (N, \lambda)$. Using an efficient recursive updating algorithm, the likelihood of any particular point θ can be computed given the observed proportions. This allows the calculation of the likelihood function on a large grid, and so confidence sets can be drawn using a contour plotting routine as soon as the nominal distances are chosen — these nominal distances indicate how far down to go from the maximum likelihood to draw the 95% confidence set line, for example. Figure 5.7 shows such a likelihood surface with contour lines determined by prepivoting using the recycle algorithm. Details of the calculations are described below.

The natural root of inference in this confidence set construction is the likelihood ratio

$$R(x; P) = -2 \log \frac{L_{x_o}(\theta)}{L_{x_o}(\hat{\theta}(x_o))}$$

where x_o is the observed part of x; that is the time series of proportions, L_{x_o} is

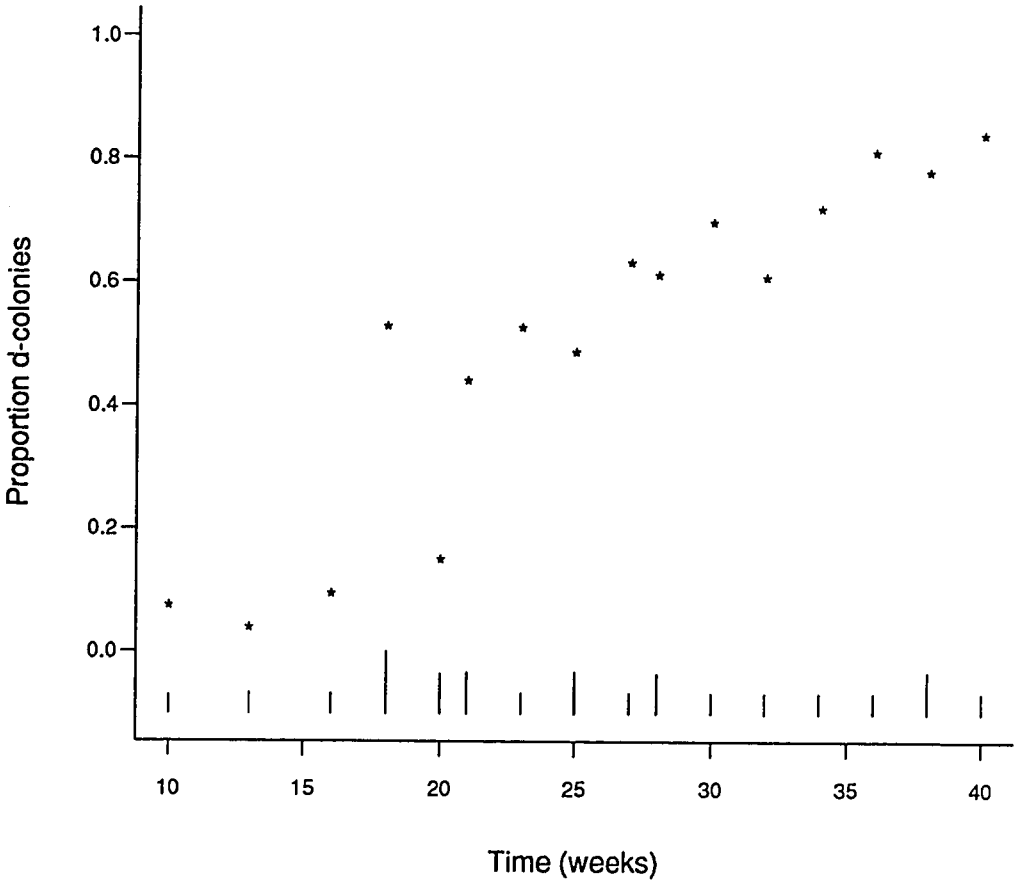


Figure 5.6: Data from Guttorp *et al.* (1990): Each point represents the proportion of a particular kind of cell in a sample of cells from a special kind of cat. The vertical bars below 0 represent the sample sizes determining each proportion. The average sample size is about 100, while the largest is 227.

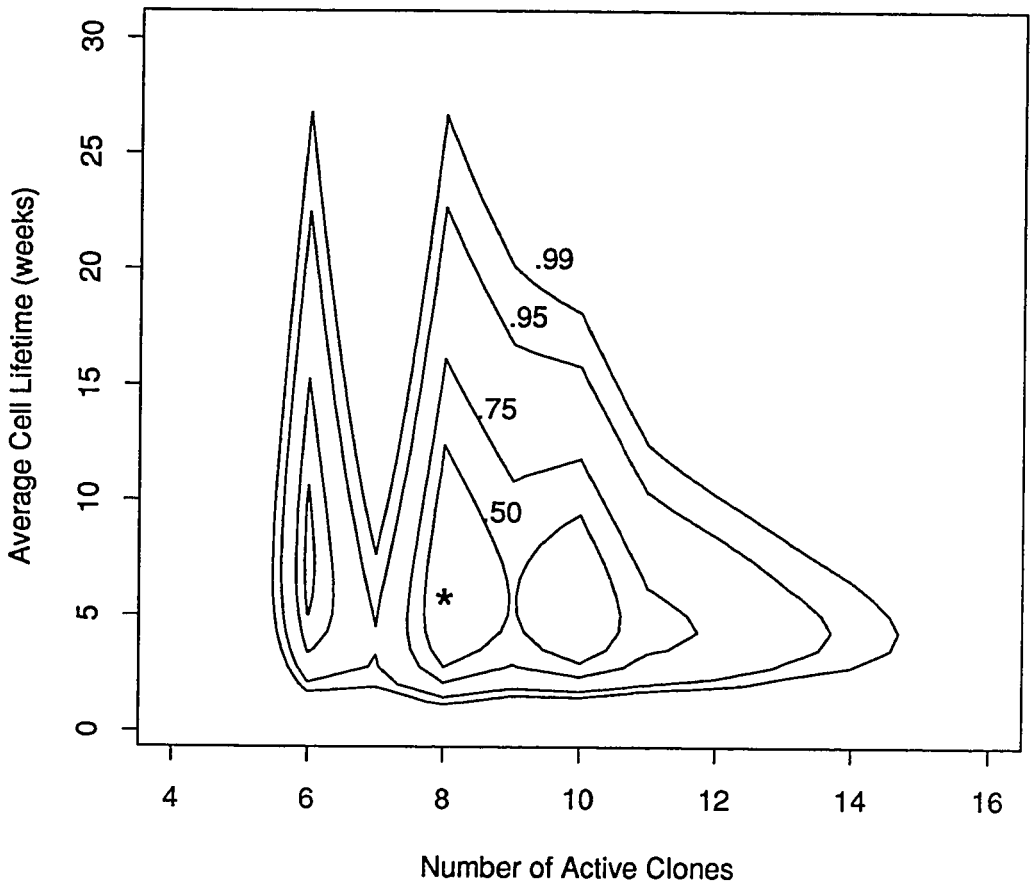


Figure 5.7: This map shows the likelihood surface for the parameters (N, λ) of the model described in Section 5.4.2 and from the data shown in Figure 5.6. The contour lines are determined by recycling the likelihood ratio. The grid on which the likelihood is evaluated is determined by 40 values of λ and each value of N, and the MLE is $(\hat{N}, \hat{\lambda}) = (8, 5.8)$.

the likelihood function given data x_o and $\hat{\theta}(x_o)$ is the MLE of θ based on x_o. A likelihood–based confidence set for θ takes the form

$$\{\theta = T(P) : \ R(x;P) \leq c\,\}$$

where c is chosen to give this set some desired nominal coverage. One might expect, from the famous result of Wilks (1938), that $R(x;P)$ is asymptotically χ–squared on 2 degrees of freedom. That N is an integer parameter means that the sufficient conditions for Wilks' result are not satisfied, and the actual limiting distribution of $R(x;P)$ in this model is unknown. Bootstrapping $R(x;P)$ is relatively straightforward. First, the unobservable continuous time Markov process is generated under the fitted model P_x. A time series of binomial proportions is then generated on top of these unobserved proportions. Figures 5.8 and 5.9 show the two step simulation of 3 such bootstrap samples. The lower right panels of these figures show the actual data and the estimated unobserved proportions, for comparison. Figure 5.10 compares the cumulative distribution function of a χ–squared 2 random variable with the empirical distribution of 240 bootstrapped roots:

$$R(y^i;P_x) = -2\log\frac{L_{y_o^i}(\hat{\theta}(x_o))}{L_{y_o^i}(\hat{\theta}(y_o^i))}.$$

Bootstrap inference can be based on quantiles of this empirical distribution. Note that the confidence sets based on bootstrapping are larger than those based on the χ–squared approximation.

Prepivoting by the iterated bootstrap is virtually impossible in this example because maximum likelihood estimates must be calculated from every bootstrap sample. This calculation is very intensive as it requires evaluation of the likelihood function on a large grid of parameter values. Recycling, on the other hand, is quite feasible on the available computing equipment. The results of recycling, based on 240 second stage samples generated from $P_2 = P_x$, are summarized in Figure 5.11. This plot compares the empirical distribution of 240 roots $\hat{R}_1(y^i;P_x)$ with the uniform cdf. There is close agreement, however there is some indication that the first stage bootstrap inference creates confidence sets which are a bit too small.

As a check on the adequacy of the recycling calculation, we look at the normalized Radon-Nikodym weights which are used. For the i^{th} first stage sample y^i, we have a

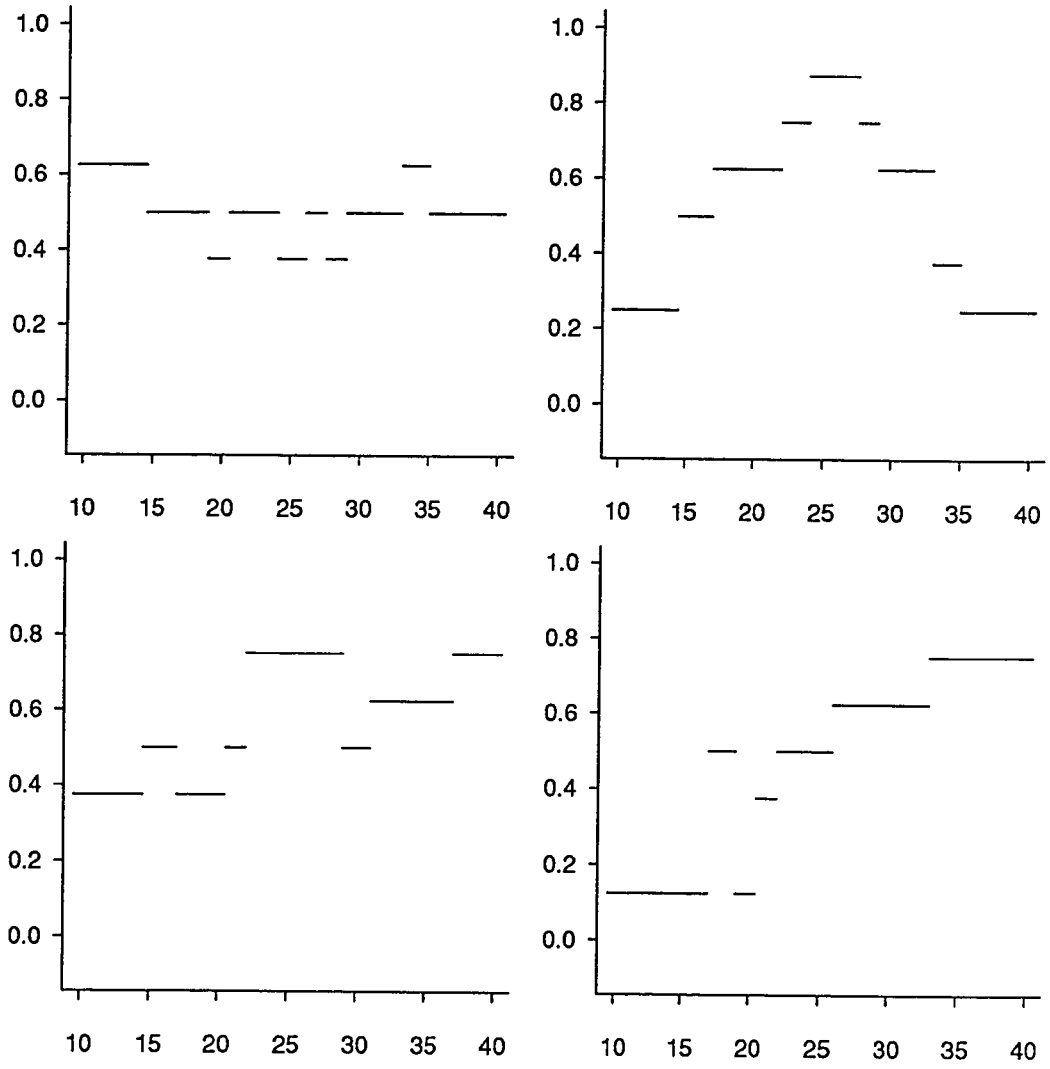


Figure 5.8: Except for the bottom right panel, these plots show examples of the simulated unobserved process $U(t_1)/N, \cdots, U(t_m)/N$ determined by the estimated model. The plot on the bottom right shows the estimated states for the actual data.

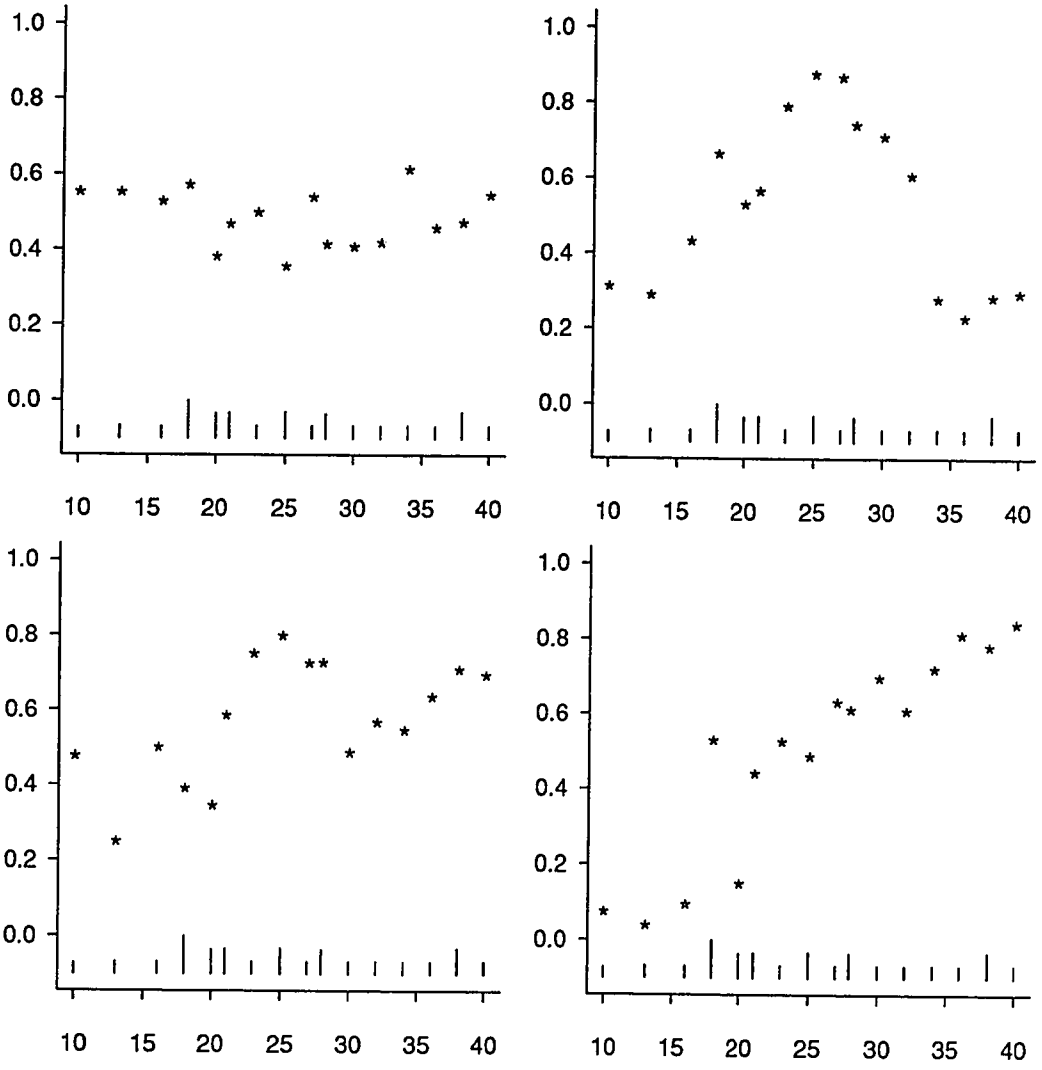


Figure 5.9: Except for the bottom right panel, these plots show examples of simulated observed proportions $O(t_1), \cdots, O(t_m)$ with mean values determined by the unobserved processes shown in Figure 5.8. The plot on the bottom right shows the actual data.

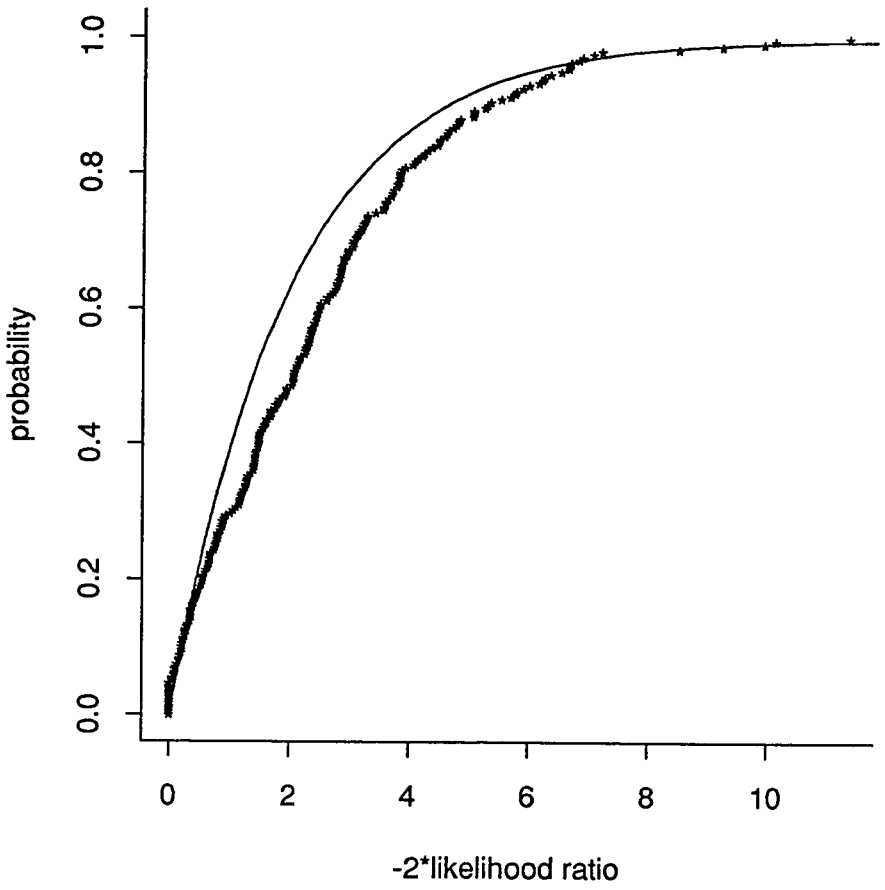


Figure 5.10: The dots show the empirical distribution function of 240 bootstrapped likelihood ratios of Section 5.4.2. The solid curve is the cdf of a χ−squared random variable on 2 degrees of freedom. Quantiles of these distributions can be used to get cutoff distances for confidence set construction.

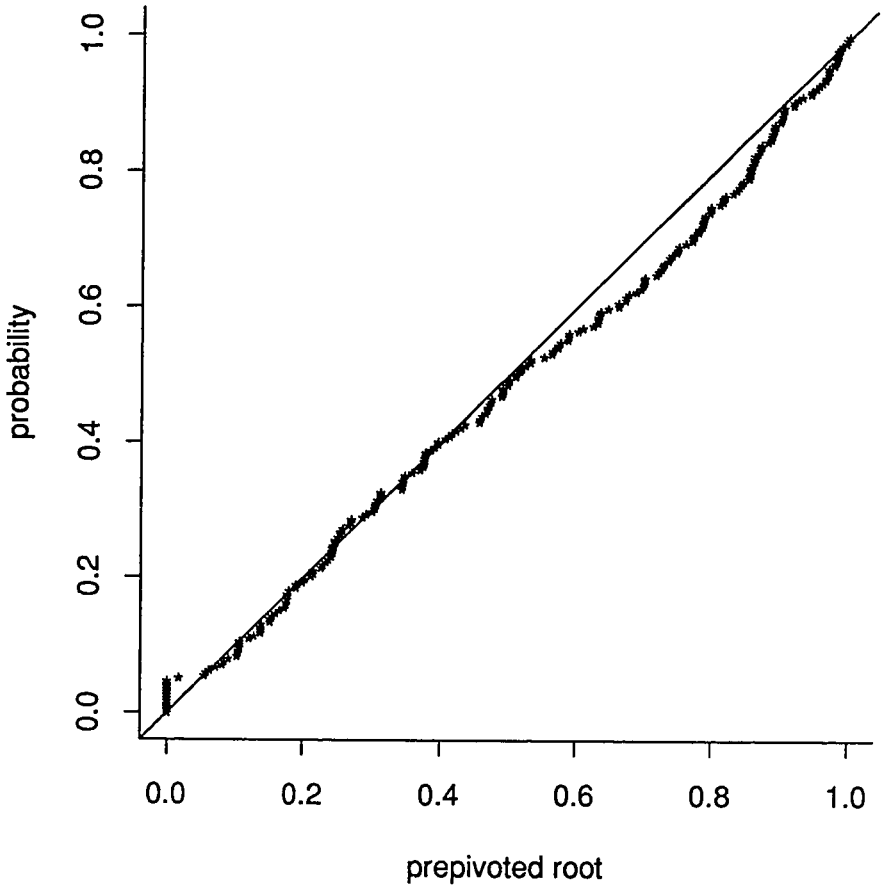


Figure 5.11: The dots show the empirical distribution function of 240 bootstrapped roots $\hat{R}_1(y^i; P_x)$. These are prepivoted likelihood ratios of Section 5.4.2 determined by recycling on 240 second stage bootstrap samples from $P_2 = P_x$. The solid curve is the cdf of a uniform random variable. Quantiles of these distributions are used to adjust the probabilities determining quantiles of the distribution of the likelihood ratio. These adjusted quantiles give cutoff distances for confidence set construction.

vector of weights

$$w^i = (w_1^i, \cdots, w_n^i)$$

of length equal to the number of second stage samples $z^1, \cdots, z^n$. Because these second stage samples come from $P_2 = P_x$ in this example, the weights satisfy

$$w_j^i \propto \frac{P_{y^i}(z^j)}{P_x(z^j)}$$

and sum to n across j. Recall that the prepivoted root, estimated by recycling, is

$$\hat{R}_1(y^i; P_x) = \frac{1}{n} \sum_{j=1}^{n} w_j^i 1 \left[R(z^j; P_{y^i}) < R(y^i; P_x) \right] .$$

The efficiency of recycling depends on the the distribution of the weights w_j^i induced by P_2. A natural goal is to find P_2 so the weights are as uniform as possible. Figure 5.12 shows boxplots indicating the distribution of the smallest percent of second stage samples which carry a given amount $(25\%, 50\%, 75\%)$ of the total weight. This plot is generated by first sorting in descending order the elements of the weight vector w^i and then computing for each i the first index in this sorted vector such that the cumulative weight exceeds p. The suggestion is that the weights are reasonably well distributed, even under $P_2 = P_x$.

5.4.3 Conditional likelihood inference

Suppose that $\mathcal{P}$ is indexed by a parameter space $\{(\theta, \psi)\}$ and that each $P \in \mathcal{P}$ has a density $f_{\theta,\psi}$ with respect to a $\sigma-$finite measure on $(\mathbf{X}, \mathcal{A})$. Here, θ is a parameter of interest and ψ is a nuisance parameter. A possible confidence set for θ is

$$\{\theta \ : \ f_{\theta,\hat{\psi}(x)}(x) \geq c_\alpha \} .$$

This conditional likelihood region can be constructed given the data x, an estimator $\hat{\psi}(x)$ of the nuisance parameter, and a cutoff value c_α chosen to give the region some desired nominal coverage $1 - \alpha$. Here, the root of the confidence set is

$$R(x; P) = f_{\theta,\hat{\psi}(x)}(x) .$$

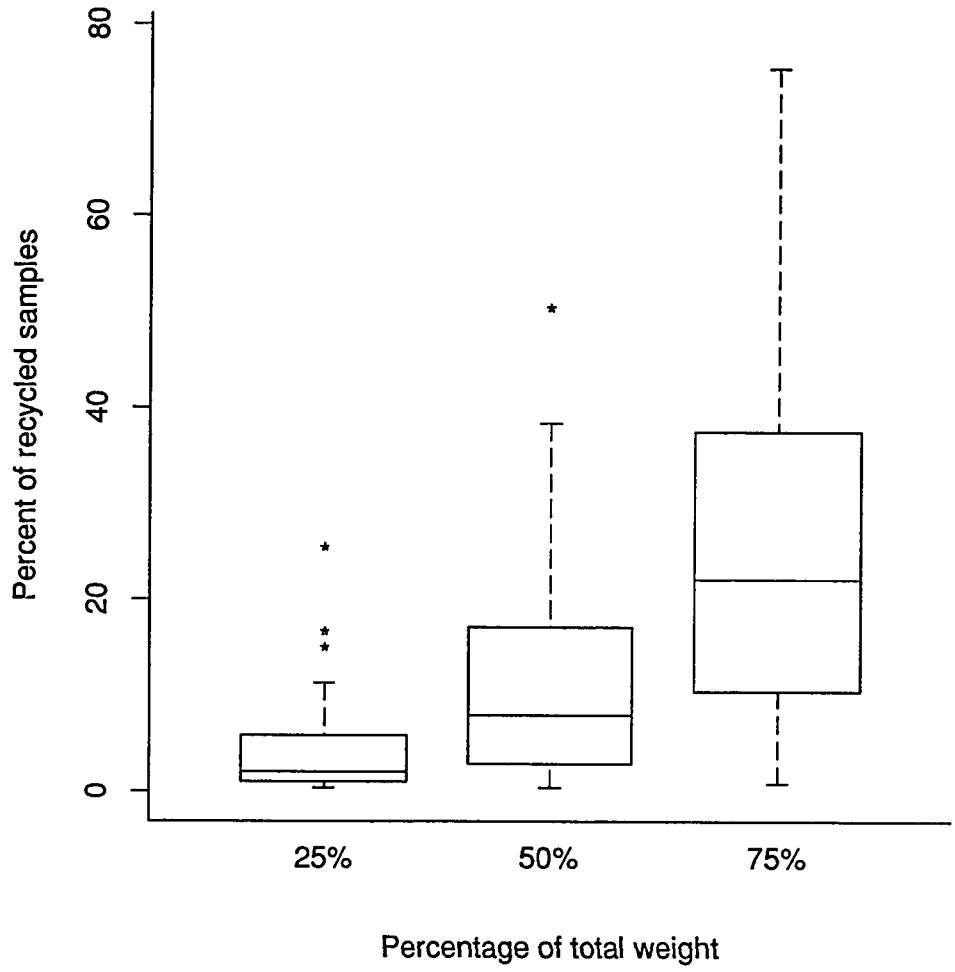


Figure 5.12: This plot indicates the extent to which the normalized Radon-Nikodym weights are uniform for the example of Section 5.4.2. For example, 75% of the weight is carried by about 20% of the second stage samples (on average over first stage samples).

As an illustration, suppose that $x = (x_1, x_2, \cdots, x_n)$ is a random sample of n real-valued random variables. A very simple *robust* model for x is to assume that each x_i has distribution function

$$P(\cdot) = G_\psi(\cdot - \theta)$$

where G_ψ is the cumulative distribution function of either a standard normal ($\psi = 1$) or a standard Cauchy ($\psi = 2$). The corresponding densities are g_ψ. This is a single model with two sub-models; the normal and the Cauchy and is indexed by

$$\{(\theta, \psi) : \theta \in \mathbf{R}, \text{ and } \psi = 1 \text{ or } 2\}.$$

Estimating P involves choosing the Cauchy or normal sub-model and then estimating θ. Thus estimation itself may involve a hypothesis test. In this example, we take the median as a consistent estimator of θ, and we use the likelihood ratio to get a consistent estimator of ψ

$$\hat{\psi}(x) = \begin{cases} 1 & \text{if } \prod_i g_1(x_i - \hat{\theta}(x)) > \prod g_2(x_i - \hat{\theta}(x)) \\ 2 & \text{otherwise.} \end{cases}$$

A confidence set for θ can be constructed in several ways. For example, we might estimate the sampling distribution of an estimator $\hat{\theta}$ of θ, and then form the confidence interval from percentiles of this estimated distribution. In this case, we are basing inference on the root

$$R(x; P) = \hat{\theta}(x) - \theta.$$

Alternatively, we can consider a likelihood-based confidence interval

$$\{\theta : R(x; P) := L_x(\theta, \hat{\psi}(x)) \geq c_\alpha\} \tag{5.6}$$

where $L_x(\theta, \psi)$ is the likelihood function

$$\prod_{i=1}^{n} g_\psi(x_i - \theta).$$

This set, containing parameter values θ having high conditional likelihood, can be constructed given the data and the constant c_α which is chosen to give the interval the desired coverage probability.

As you might guess, if we get c_α from the bootstrap distribution of the root $R(y; P_x)$, we are still prone to substantial error because we are conditioning on one of the submodels being true. By prepivoting, the error incurred by conditioning on a submodel is reduced. It is reduced not by *Bayesianly* integrating out uncertainty, but rather by finding a root having a distribution less sensitive to the measure P generating the data.

5.4.4 Empirical Bayes confidence sets

Consider the parametric empirical Bayes models described in Morris (1983). In the notation of Carlin and Gelfand (1990), unobserved real *parameters* $\theta_1, \theta_2, \cdots, \theta_m$ are viewed as i.i.d. from some continuous *prior* distribution $\pi(\cdot|\eta)$ for a *hyperparameter* η. Given θ_i, we observe data

$$\begin{aligned} y_i &= (y_{i,1}, y_{i,2}, \cdots, y_{i,n_i}) \\ y_{i,j} &\sim_{iid} f(\cdot|\theta_i)\,. \end{aligned}$$

This specifies a model $\mathcal{P}$ indexed by η, for the random vector $x = (y, \theta_1, \theta_2, \cdots, \theta_m)$, where $y = (y_1, y_2, \cdots, y_m)$. Any point P in the model induces a *posterior* c.d.f. $F(\cdot|y_i, \eta)$ for each θ_i. This is not a Bayesian posterior because η is considered unknown. Note that given y_i and η, the other data sets y_j, $j \neq i$ provide no further information about θ_i.

If $\hat{\eta}(y)$ is some estimate of η, then the *naive* (equi-tailed) empirical Bayes confidence set for θ_i (of nominal level $1 - \alpha$) is

$$\left\{\theta_i \,:\, \frac{\alpha}{2} \leq F(\theta_i|y_i, \hat{\eta}(y)) \leq 1 - \frac{\alpha}{2}\right\}. \tag{5.7}$$

In our terminology, the root of inference about the unobserved θ_i is

$$R(x; P) = F(\theta_i|y_i,\, \hat{\eta}(y))\,. \tag{5.8}$$

This is a case where the focus of inference is an unobserved random variable.

The naive construction (5.7) is based on assuming that $R(x; P)$ is uniformly distributed, regardless of P. Since P is determined by the hyperparameter η, this assumption would be true if η was known, but because η is estimated from data,

(5.7) is typically too short (as measured by conditional or unconditional coverage). Essentially, (5.7) is the asymptotic confidence interval based on the root in (5.8) where both $m \to \infty$ and $n_i \to \infty$ for all $1 \leq i \leq m$.

Another way to view (5.7) is as the *conditional bootstrap* interval based on the root

$$R_0(x; P) = \theta_i \,.$$

We say *conditional bootstrap* because there are two ways to bootstrap in this problem, as in any problem where inference is about an unobserved random variable. The unconditional bootstrap of $R_0(x; P)$ is simply based on fitting the marginal distribution of θ_i; $\pi(\cdot|\hat{\eta}(y))$. The conditional bootstrap, on the other hand, involves the fitted conditional distribution of θ_i given $b(y) = y_i$ having c.d.f.

$$F(\cdot|y_i, \hat{\eta}(y))$$

as in (5.8). We see that $R(x; P)$ in (5.8) is a *conditionally prepivoted* version of the root $R_0(x; P)$.

The naive interval (5.7) can be improved by prepivoting. Instead of getting interval endpoints from the uniform distribution, we can estimate the distribution of $R(x; P)$ from the sample. Let $H(\cdot, P)$ be the c.d.f of $R(x; P)$ *conditional on* $b(y) = y_i$. A once prepivoted empirical Bayes confidence interval for θ_i is

$$\left\{\theta_i \,:\, H^{-1}\left(\alpha/2, P_x\right) \leq F(\theta_i|y_i, \hat{\eta}(y)) \leq H^{-1}\left(1-\alpha/2, P_x\right)\right\} . \tag{5.9}$$

Since

$$H(u, P) = P\left\{\, F(\theta_i|y_i, \hat{\eta}(y)) < u \,|\, y_i, \eta \,\right\} \qquad u \in (0,1)\,,$$

we have

$$H(u, P_x) = P\left\{\, F(\theta_i|y_i^*, \hat{\eta}(y^*)) < u \,|\, y_i^*, \hat{\eta}(y) \,\right\} \qquad u \in (0,1)\,, \tag{5.10}$$

where y^* has marginal distribution determined by $\hat{\eta}(y)$ instead of η, as does θ_i. To construct a prepivoted interval, the lower and upper bounds of θ_i are still found from quantiles of $f(\theta_i|y_i, \hat{\eta}(y))$. These are no longer the $\alpha/2$ and $1-\alpha/2$ quantiles, but

are adjusted by prepivoting. These modified cutoff probabilities can be computed by a double bootstrap or by bootstrap recycling.

If y_i replaces y_i^* in (5.10), the prepivoted interval (5.9) is exactly the same as the bias-corrected interval of Carlin and Gelfand (1990).

5.5 Simulation consistency

We usually think of consistency as a property that estimators have when the observed sample size gets large. We might, therefore, give the name *simulation consistency* to bootstrap or Monte Carlo approximations which converge to the right answer as the amount of computing resources gets large, for a fixed data set. For straightforward sampling procedures, the strong law of large numbers ensures simulation consistency. For more complicated procedures, like bootstrap recycling, it is not a trivial matter to prove simulation consistency of the algorithm.

In this section, we demonstrate simulation consistency for the algorithm described in Section 5.3.1, (bootstrap recycling, method I). For simplicity, suppose that each stage of bootstrap sampling uses the same number m of bootstrap samples. Also, suppose that the stage samples are ordinary bootstrap samples; that is $y^i \sim P_1 = P_x$. The algorithm produces an estimator $\hat{H}_{1,m}(\cdot, P_x)$ of the distribution function $H_1(\cdot, P_x)$. We have the following result:

Theorem 11 *Let P_1 and P_2 of the (method I) bootstrap recycling algorithm be chosen so that $P_1 = P_x$, $P_x << P_2$, and if $y \sim P_1$ then $P_y << P_2$. The method I bootstrap recycling algorithm is simulation consistent. That is, as $m \to \infty$,*

$$\hat{H}_{1,m}(u, P_x) \to H_1(u, P_x)$$

with simulation probability 1. Moreover, if the limiting distribution is continuous, then the convergence is uniform in u.

PROOF. There are two independent sets of simulated data

$$y^1, y^2, \cdots, y^m \sim_{iid} P_1 = P_x$$
$$z^1, z^2, \cdots, z^m \sim_{iid} P_2 >> P_x .$$

All the y^i's and z^j's in the roots used below are distributed as above. We introduce some simplifying notation. For fixed u in $[0,1]$, put

$$\begin{aligned} h_m &:= \hat{H}_{1,m}(u, P_x) \\ &= \frac{1}{m}\sum_{i=1}^{m} 1\left[\hat{R}_{1,m}(y^i; P_x) < u\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} 1\left[\sum_{j=1}^{m} g(y^i, z^j) < mu\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} V_{m,i}\,, \end{aligned}$$

where

$$\begin{aligned} g(y,z) &= 1\left[R(z; P_y) < R(y; P_x)\right]\frac{dP_y}{dP_2}(z) \\ V_{m,i} &= 1\left[\sum_{j=1}^{m} g(y^i, z^j) < mu\right]\frac{dP_x}{dP_1}(y^i)\,. \end{aligned}$$

Finally, let h be the desired limit of h_m, namely $H_1(u, P_x)$.

The proof proceeds by dividing h_m into two bits,

$$h_m = \eta_m + e_m \tag{5.11}$$

where e_m is an error term converging to 0 with simulation probability one, and η_m is fairly easy to handle. A convenient choice of η_m is

$$\eta_m = E(V_{m,i}|\mathcal{Z}_m)\,, \tag{5.12}$$

where $\mathcal{Z}_m$ is the σ–field generated by $z^1, z^2, \cdots, z^m$. (Note that the probability space involved is the product space of all the z^i, and y^j. Almost sure statements are with respect to the probability measure on this space: the simulation probability.) With this choice of η_m, the error term e_m becomes

$$e_m = \frac{1}{m}\sum_{i=1}^{m}(V_{m,i} - \eta_m)\,.$$

Importantly, the variables $V_{m,i}$ are (for fixed m) conditionally independent given $\mathcal{Z}_m$. Using this along with the fact that $|V_{m,i}| \leq 1$ and properties of conditional

expectations, we have that for any $\epsilon > 0$,

$$\begin{aligned} P(|e_m| \geq \epsilon) &= E\{P[\,|e_m| \geq \epsilon \mid \mathcal{Z}_m\,]\} \\ &\leq \frac{1}{m^4\epsilon^4} E\left\{E\left[\left(\sum_{i=1}^{m}(V_{m,i} - \eta_m)\right)^4 \mid \mathcal{Z}_m\right]\right\} \\ &\leq \frac{1}{m^4\epsilon^4}\left(3m(m-1)+m\right) \\ &< \frac{3}{m^2\epsilon^4}. \end{aligned}$$

Therefore, $\sum_m P(|e_m| \geq \epsilon) < \infty$, and so by the Borel-Cantelli lemmas, e_m converges to 0 with simulation probability one.

From (5.11), it remains to show that η_m converges to the desired limit h. Now from (5.12), η_m is a conditional expectation of $V_{m,1}$. Using a corollary to the Ergodic Theorem (see Section B of Appendix B), we have that

$$V_{m,1} = 1\left\{\frac{1}{m}\sum_{j=1}^{m} g(y^1, z^j) < u\right\} \rightarrow 1\left\{E\left[\,g(y^1, z^1)\,\middle|\,y^1\,\right] < u\right\} := V_{\infty,1}$$

as $m \rightarrow \infty$ with simulation probability 1. Also, since $V_{m,1}$ depends only on the first m variables $z^1, \cdots, z^m$, we have

$$\eta_m = E(V_{m,1}|\mathcal{Z}_m) = E(V_{m,1}|\mathcal{Z}_\infty) \quad a.s.$$

where $\mathcal{Z}_\infty$ is the σ–field determined by the union of the $\mathcal{Z}_m$ for all m. Using the almost sure convergence and boundedness of $V_{m,1}$, we can apply the dominated convergence theorem for conditional expectations (for example, Theorem 2, page 208 of Chow and Teicher, 1988) to get

$$\eta_m \rightarrow_{a.s.} E(V_{\infty,1}|\mathcal{Z}_\infty).$$

Finally, since $V_{\infty,1}$ is independent of all z^j, it is independent of $\mathcal{Z}_\infty$ and so

$$\begin{aligned} \eta_m \rightarrow_{a.s.} \quad & E\{V_{\infty,1}\} \\ &= E\left\{1\left[E\left(g(y^1, z^1)\,|\,y^1\right) < u\right]\right\} \\ &= P_x\left\{P_{y^1}\left[R(z^1; P_{y^1}) < R(y^1; P_x)\,\middle|\,x, y^1\right] < u \mid x\right\} \\ &= h \end{aligned}$$

completing the proof for fixed u. Uniformity in u when the limit is continuous is an immediate consequence of a result in Chung (1974), (lemma, page 133). □

5.6 A recycle algorithm for p prepivots

To compute an approximation for $H_k(\cdot, P_x)$ for $k = 1, 2, \cdots, p$ we need $(p+1)$ independent sets of bootstrap samples. Let $m_0, m_1, \cdots, m_k$ be the number of bootstrap samples in each set and denote by I_k the integers from 1 to m_k. Put $I_{-1} = \{0\}$. We denote the bootstrap samples by double superscripts as $y^{k,i}$, $i \in I_k$ and $0 \leq k \leq p$. Thus we get a set of bootstrap samples for each k. For each $k = 0, 1, \cdots, p$, we introduce two two-dimensional arrays

$$a_k(u, v) \text{ and } b_k(u, v)$$

for non-negative integers u, v. For **method I**, we need a sequence of weight matrices W_k with elements

$$W_k(u, v) = \frac{dP_{y^{k-1,u}}}{dP_1}(y^{k,v})$$

for $1 \leq k \leq p$, $u \in I_{k-1}$ and $v \in I_k$. Throughout it is assumed that $P_1 = P_x$. As remarked earlier, if the bootstrap sample $y^{k,i}$ is distributed $P_{y^{k-1,i}}$ (**method II**), then the weights are

$$W_k(u, v) = \frac{dP_{y^{k-1,u}}}{dP_{y^{k-1,v}}}(y^{k,v}) .$$

An algorithm for recycling is given below.

Bootstrap recycle algorithm for p prepivots:

Stage 1; the bootstrap

1. Generate bootstrap samples $y^{0,i}$ for $i \in I_0$.
2. For $i \in I_0$, set $a_0(i, 0) = R(y^{0,i}; P_x)$.
3. Construct the approximation

$$\hat{H}(u, P_x) = \frac{1}{m_0} \sum_{i=1}^{m_0} 1\left[a_0(i, 0) < u\right] .$$

4. Prepare for next stage; for $i \in I_0$ set

$$b_0(i,0) \leftarrow a_0(i,0)$$

Stage 2; the first prepivot

1. Generate bootstrap samples $y^{1,j}$ for $j \in I_1$.
2. For $i \in I_0$ and $j \in I_1$, compute the weight $W_1(i,j)$.
3. Calculate new roots: for $i \in I_0$ and $j \in I_1$,

$$a_0(j,i) \leftarrow R(y^{1,j}; P_{y^{0,i}})$$
$$a_1(i,0) \leftarrow \frac{1}{m_1} \sum_{j=1}^{m_1} 1\left[a_0(j,i) < b_0(i,0)\right] W_1(i,j)$$

4. Construct the approximation

$$\hat{H}_1(u, P_x) = \frac{1}{m_0} \sum_{i=1}^{m_0} 1\left[a_1(i,0) < u\right] .$$

5. Prepare for the next stage; for $i \in I_0$ and $j \in I_1$ set

$$b_0(j,i) \leftarrow a_0(j,i)$$
$$b_1(i,0) \leftarrow a_1(i,0)$$

$\vdots$

Stage $p+1$; the p^{th} prepivot

1. Generate bootstrap samples $y^{p,j}$ for $j \in I_p$.
2. For $i \in I_{p-1}$ and $j \in I_p$, compute the weight $W_p(i,j)$.
3. Calculate new roots

   For $i \in I_{p-1}$ and $j \in I_p$

$$a_0(j,i) \leftarrow R(y^{p,j}; P_{y^{p-1,i}})$$

Do $k = 1, 2, \cdots p$: for $i \in I_{p-k-1}$ and $j \in I_{p-k}$

$$a_k(j,i) \leftarrow \frac{1}{m_{p-k+1}} \sum_{l=1}^{m_{p-k+1}} 1\left[a_{k-1}(l,j) < b_{k-1}(j,i)\right] W_{p-k+1}(j,l)$$

4. Construct the approximation

$$\hat{H}_p(u, P_x) = \frac{1}{m_0} \sum_{i=1}^{m_0} 1\left[a_p(i,0) < u\right] .$$

5. Prepare for the next stage;
   Do $k = 0, p$: for $i \in I_{p-k-1}$ and $j \in I_{p-k}$

$$b_k(j,i) \leftarrow a_k(j,i)$$

Of course, other algorithms trading off storage and speed may be constructed to produce the same results.

BIBLIOGRAPHY

[1] J. L. Abkowitz, M. L. Linenberger, M. A. Newton, G. H. Shelton, R. L. Ott, and P. Guttorp. Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. *Proc. Natl. Acad. of Sci.* USA, 87:9062–9066, 1990.

[2] J. Aitchison, and I. R. Dunsmore. *Statistical Prediction Analysis*. University Press, Cambridge, 1975.

[3] R. Beran. Prepivoting to reduce level error of confidence sets. *Biometrika* 74:457–468, 1987.

[4] R. Beran. Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* 83:687–697, 1988.

[5] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. R. Statist. Soc. B* 2:192–236, 1974.

[6] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1991.

[7] P. Billingsly. *Probability and Measure (2nd ed.)*. John Wiley & Sons, New York, 1986.

[8] G. D. Birkhoff. Proof of the ergodic theorem. *Proc. Nat'l. Acad. Sci.* 17:656–660, 1931.

[9] D. D. Boos and J. F. Monahan. Bootstrap Methods Using Prior Information. *Biometrika* 73:77–83, 1986.

[10] L. Breiman. *Probability*. Addison-Wesley, Reading, Mass., 1968.

[11] Norm Breslow. Personal communication. 1987.

[12] S. G. Brush. Foundations of Statistical Mechanics 1845–1915. *Archive for History of Exact Sciences* 4:145–183, 1967.

[13] B. P. Carlin and A. E. Gelfand. Approaches for empirical Bayes confidence intervals. *J. Amer. Statist. Assoc.* 85:105–114, 1990.

[14] Y. S. Chow. Some convergence theorems for independent random variables. *Ann. Math. Statist.* 37:1482–1492, 1966.

[15] Y. S. Chow and H. Teicher. *Probability Theory: Independence, Interchangeability, Martingales (2nd ed.).* Springer Verlag, New York, 1988.

[16] K. L. Chung. *A Course in Probability Theory (2nd ed.).* Academic Press, New York, 1974.

[17] D. R. Cox. Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* 34:187–220, 1972.

[18] D. R. Cox. Partial Likelihood, *Biometrika* 62:269–276, 1975.

[19] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events.* Methuen, London, 1966.

[20] H. Cramér. *Mathematical Methods of Statistics.* Princeton University Press, Princeton, N. J., 1946.

[21] A. C. Davison. Discussion of paper by D. V. Hinkley. *J. R. Statist. Soc. B* 50:356-357, 1988.

[22] A. P. Dawid. Present position and potential developments: Some personal views. Statistical theory—The prequential approach (with Discussion). *J. R. Statist. Soc. A* 147:278–292, 1984.

[23] A. R. Dempster, N. M. Laird and D. B. Rubin. Maximimum Likelihood from Incomplete Data via the *EM* Algorithm (with discussion). *J. R. Statist. Soc. B* 39:1–38, 1977.

[24] J. L. Doob. *Stochastic Processes.* John Wiley & Sons, New York, 1953.

[25] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7:1–26, 1979.

[26] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* Society of Industrial and Applied Mathematics, Philadelphis, 1982.

[27] B. Efron. Comment on: Empirical Bayes confidence intervals byased on bootstrap samples, by N. M. Laird and T. A. Lewis. *J. Amer. Statist. Assoc.* 82:754, 1987.

[28] B. S. Everitt. Mixture Models. *Encyclopedia of Statistical Sciences* S. Kotz and N. Johnson eds., John Wiley & Sons, 1985.

[29] R. V. Foutz. On the Unique Consistent Solution to the Likelihood Equations. *J. Amer. Statist. Assoc.* 72:147–148, 1977.

[30] A. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal posterior densities. *J. Amer. Statist. Assoc.* 85:398–409, 1990.

[31] E. Giné and J. Zinn. Bootstrapping general empirical measures. *Ann. Prob.* 18:851–869, 1990.

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

[33] S. W. Guo and E. A. Thompson. Analyis of sparse contingency tables: Monte Carlo estimation of exact p-values. *Technical Report* 187, Department of Statistics, University of Washington, Seattle, 1989.

[34] P. Guttorp, M. A. Newton, and J. L. Abkowitz. A Stochastic Model for Hematopoiesis in Cats. *Institue for Mathematical Applications Journal of Mathematics Applied in Medicine and Biology* 7: 125–143, 1990.

[35] E. Haeusler, D. M. Mason, and M. A. Newton. Weighted Bootstrapping of Means. *To appear, CWI*, 1991.

[36] P. Hall. Theoretical comparision of bootstrap confidence intervals. *Annals of Statistics* 16:927–953, 1988.

[37] P. Hall and C. C. Heyde. *Martingale Limit Theory and its Application* Academic Press, New York, 1980.

[38] I. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods.* John Wiley, New York, 1964.

[39] J. A. Hartigan. *Bayes Theory* Springer–Verlag, New York, 1983.

[40] D. V. Hinkley and S. Shi. Importance sampling and the nested bootstrap. *Biometrika* 76:435–446, 1989.

[41] N. L. Hjort. Bayesian Nonparametric Bootstrap Confidence Intervals. *Technical Report* 20, Stanford, November, 1985.

[42] V. S. Huzurbazar. The Likelihood Equation, Consistency and Maxima of the Likelihood Function. *Annals of Eugenics* 14(3):185–200, 1948.

[43] E. T. Jaynes. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics* Vol. ssc-4, No. 3, 1968.

[44] M. V. Johns. Importance sampling for bootstrap confidence intervals. *J. Amer. Statist. Assoc.* 83:709–714, 1988.

[45] R. A. Johnson. An asymptotic expansion for posterior distributions. *Ann. Math. Statist.* 38:1899–1907, 1967.

[46] R. A. Johnson. Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* 41:851–864, 1970.

[47] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimate in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27:887–906.

[48] R. E. Kass. The Geometry of Asymptotic Inference, with discussion. *Statistical Science* 4:188–234, 1989.

[49] H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17:1217–1241, 1989.

[50] N. M. Laird and T. A. Lewis. Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* 82:739–757, 1987.

[51] K. W. Lee. Bootstrapping logistic regression models with random regressors. *Comm. Stat.* 19:2527–2593, 1990.

[52] E. L. Lehmann. *Theory of Point Estimation.* John Wiley & Sons, New York, 1983.

[53] D. V. Lindley. Approximate Bayesian methods. *Bayesian Statistics*, (eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith). University Press, Valencia, 1980.

[54] A. Y. Lo. A large sample study of the Bayesian bootstrap. *Annals of Statistics* 15:360–375, 1987.

[55] W. Loh. Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* 82:739–750, 1987.

[56] D. M. Mason, and M. A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Technical Report no. 190*, Department of Statistics, University of Washington. Under revision for *Ann. Statist.*, 1990.

[57] P. McCullagh and J. A. Nelder. *Generalized Linear Models (2nd ed.)*. Chapman and Hall, 1989.

[58] G. S. McLachlan and K. E. Basford. *Mixture Models: inference and applications to clustering*. Marcel Dekker, New York, 1988.

[59] C. R. Mehta and N. R. Patel. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* 78:427–434, 1983.

[60] C. N. Morris. Parametric Empirical Bayes Inference: Theory and Applications. *J. Amer. Statist. Assoc.* 78:47–59, 1983.

[61] C. R. Nelson and C. I. Plosser. Trends and Random Walks in Macroeconomic Time Series: some evidence and implications. *Journal of Monetary Economics* 10:139–162, 1982.

[62] S. Noll, P. Waibel, R. Cook, and J. Witmer. Biopotency of menthionine sources for young turkeys. *Poultry Science* 63:2458–2470, 1984.

[63] I. Olkin, L. J. Gleser, and C. Derman. *Probability Models and Applications*. Macmillan, New York, 1980.

[64] A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75:237–249, 1988.

[65] A. Owen. Empirical likelihood ratio confidence regions. *Annals of Statistics* 18:90–120, 1990.

[66] E. S. Pearson and H. O. Hartley. *Biometrika Tables for Statisticians*. Vol. 1, 3rd ed. Cambridge University Press, Cambridge, 1976.

[67] M. D. Perlman. On the strong consistency of approximate maximum likelihood estimators. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* 1:263–282, Univ. of California Press, 1972.

[68] F. Proschan. Theoretical Explanation of Observed Decreasing Failure Rate. *Technometrics* 5:375–383, 1963.

[69] A. Racine-Poon. A Bayesian approach to nonlinear calibration problems. *J. Amer. Statist. Assoc.* 83(403):650–656, 1988.

[70] C. R. Rao. *Linear Statistical Inference and its Applications*. John Wiley & Sons, New York, 1973.

[71] D. B. Rubin. The Bayesian bootstrap. *Annals of Statistics* 9:130–134, 1981.

[72] D. B. Rubin. Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics 3* (eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith), University Press, Oxford, 395–402, 1988.

[73] W. Rudin. *Principles of Mathematical Analysis*. McGraw–Hill, New York, 1964.

[74] F. W. Scholtz. Maximum Likelihood Estimation. *Encyclopedia of Statistical Sciences* Vol. 5, John Wiley & Sons, New York, 1985.

[75] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980.

[76] K. R. Stromberg. *An Introduction to Classical Real Analysis*. Wadsworth, 1981.

[77] M. Tanner and W. Wong. The Calculation of Posterior Densities by Data Augmentation (with discussion). *J. Amer. Statist. Assoc.* 82:528–550, 1987.

[78] G. R. Terrel. The Maximal Smoothing Principle in Density Estimation. *J. Amer. Statist. Assoc.* 85:470–477, 1990.

[79] A. Thrum. A remark on almost sure convergence of weighted sums. *Probab. Th. Rel. Fields* 75:425–430, 1987.

[80] L. Tierney and J. Kadane. Accurate Approximations for Posterior Moments and Marginal Densities. *J. Amer. Statist. Assoc.* 81:82–86, 1986.

[81] L. Tierney, R. E. Kass, and J. Kadane. Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* 84:710–716, 1989.

[82] A. Tsiatis. A large sample study of Cox's regression model. *Annals of Statistics* 9:93–108, 1981.

[83] S. S. Wilks. The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* 9:60, 1938.

[84] A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* 20:595–600, 1949.

[85] A. M. Walker. Asymptotic behaviour of posterior distributions. *J. R. Statist. Soc. B* 31:80–88, 1969.

[86] S. Weisberg. *Applied Linear Regression.* John Wiley & Sons, New York, 1985.

[87] C. S. Weng. On a second-order asymptotic property of the Bayesian bootstrap. *Annals of Statistics* 17:705–710, 1989.

[88] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics* 11:95–103, 1983.

[89] Z. Zheng and D. Tu. Random weighting method in regression models. *Scientia Sinica A* XXXI:1442–1459, 1988.

Appendix A

AUXILIARY RESULTS

This chapter contains results necessary in many proofs but tangential to the main line of thought in the thesis.

A.1 Properties of Dirichlet vectors

For non-negative constants $\alpha_1, \alpha_2, \cdots, \alpha_n$, not all equal to 0, define random variables $Y_1, Y_2, \cdots, Y_n$ as $Y_i = 0$ with probability one if $\alpha_i = 0$, and Y_i has the Gamma density

$$\frac{y^{\alpha_i - 1} e^{-y}}{\Gamma(\alpha_i)} \qquad y > 0 \tag{A.1}$$

with respect to Lebesgue measure if $\alpha_i > 0$. The random vector

$$(W_1, W_2, \cdots, W_n) := \frac{1}{\sum_j Y_j}(Y_1, Y_2, \cdots, Y_n) \tag{A.2}$$

is said to have a Dirichlet distribution with parameters $\alpha_1, \alpha_2, \cdots, \alpha_n$, denoted

$$(W_1, W_2, \cdots, W_n) \sim \text{Dirichlet}_n(\alpha_1, \alpha_2, \cdots, \alpha_n)\,. \tag{A.3}$$

For any permutation π of $\{1, 2, \cdots, n\}$,

$$(W_{\pi_1}, W_{\pi_2}, \cdots, W_{\pi_n}) \sim \text{Dirichlet}_n\left(\alpha_{\pi_1}, \alpha_{\pi_2}, \cdots, \alpha_{\pi_n}\right)\,. \tag{A.4}$$

The components W_i are exchangeable if and only if all α_i are the same. Let $1 < m_1 \leq m_2 \leq \cdots, \leq m_k < n$ be integers for some $k \geq 1$. Dirichlet vectors have a nice *collapsing* property:

$$\left(\sum_{j=1}^{m_1} W_j, \sum_{j=m_1+1}^{m_2} W_j, \cdots, \sum_{j=m_k+1}^{n} W_j\right) \sim \text{Dirichlet}_{k+1}\left(\gamma_1, \gamma_2, \cdots, \gamma_{k+1}\right) \tag{A.5}$$

where

$$(\gamma_1, \gamma_2, \cdots, \gamma_{k+1}) = \left(\sum_{j=1}^{m_1} \alpha_j, \sum_{j=m_1+1}^{m_2} \alpha_j, \cdots, \sum_{j=m_k+1}^{n} \alpha_j \right) .$$

If all α_i are positive, then the first $n-1$ components $W_1, \cdots, W_{n-1}$ have the following density with respect to Lebesgue measure on $\mathbf{R}^{n-1}$:

$$p(w_1, w_2, \cdots, w_{n-1}) = \frac{\Gamma\left(\sum_{j=1}^n \alpha_j\right)}{\prod_{j=1}^n \Gamma(\alpha_j)} \prod_{j=1}^n w_j^{\alpha_j - 1} \qquad w_j > 0, \quad \sum_{j=1}^n w_j = 1 . \quad \text{(A.6)}$$

We are particularly interested in the case $\alpha_i = 1$, which we call the uniform Dirichlet distribution. The moments of the uniform Dirichlet are

$$E \prod_{j=1}^n W_j^{r_j} = \frac{\Gamma(n)}{\Gamma(n + \sum r_j)} \prod_{j=1}^n \Gamma(1 + r_j) \qquad \text{(A.7)}$$

for some vector $r_1, \cdots, r_n$.

A.2 Certain moments in exponential families

Suppose that X has a one-parameter exponential family density

$$f_\theta(x) = \exp\{a(\theta)x - c(\theta) + d(x)\}$$

and $\theta = EX$. With sufficiently smooth a and c, it is readily shown that

$$EX = \frac{c^{(1)}(\theta)}{a^{(1)}(\theta)}$$

using superscipt notation for derivatives. Thus $c^{(1)}(\theta) = \theta a^{(1)}(\theta)$. With $\phi_X(\theta) = \log f_\theta(X)$, define

$$I(\theta) := -E\phi_X^{(2)}(\theta)$$
$$K(\theta) := E\phi_X^{(3)}(\theta) .$$

We have the following interesting result:

Theorem 12

$$\begin{aligned}\frac{1}{2}K(\theta) &= -a^{(2)}(\theta) \\ &= -I^{(1)}(\theta) \\ &= E\left(\phi_X^{(1)}(\theta)\right)^3 .\end{aligned} \tag{A.8}$$

PROOF. The first two equalities are easily established using the definition of ϕ_X, the fact that $\theta = EX$, and the relation $c^{(1)}(\theta) = \theta a^{(1)}(\theta)$. To prove the third equality, the second and third moments of X must be computed in terms of θ and $a(\theta)$. These moments can be more easily determined in the natural parameterization of the model:

$$\eta = a(\theta), \qquad A(\eta) = c(\theta) .$$

From Lehmann (1983) page 31 for example, the moment generating function of X is

$$M_(t) = \exp\{A(\eta + t) - A(\eta)\}$$

which implies that

$$\begin{aligned}EX &= A^{(1)}(\eta) \\ EX^2 &= A^{(2)}(\eta) + \left(A^{(1)}(\eta)\right)^2 \\ EX^3 &= A^{(3)}(\eta) + 3A^{(2)}(\eta)A^{(1)}(\eta)\left(A^{(1)}(\eta)\right)^3 .\end{aligned}$$

Furthermore, by implicit differentiation of $A(a(\theta)) = c(\theta)$,

$$\begin{aligned}A^{(1)}(\theta) &= \theta \\ A^{(2)}(\theta) &= \frac{1}{a^{(1)}(\theta)} \\ A^{(3)}(\theta) &= -\frac{a^{(2)}(\theta)}{\left(a^{(1)}(\theta)\right)^3} .\end{aligned}$$

Using this, the third moment of $\phi_X(\theta)$ can be computed and shown to equal $K(\theta)/2$. □

A.3 An expansion of posterior moments

By arguments similar to those in Lindley (1980), we derive an asymptotic expansion for the ratio of integrals

$$R_n = \frac{N_n}{D_n} = \frac{\int w(\theta)e^{nl_n(\theta)}\,d\theta}{\int v(\theta)e^{nl_n(\theta)}\,d\theta} . \tag{A.9}$$

Here $l_n(\theta)$ is the scaled log-likelihood $(1/n)\sum \log f_\theta(X_i)$ and w and v are two smooth functions of θ (possible dependent upon n). When $v(\theta)$ is a prior density and $w(\theta) = g(\theta)v(\theta)$, then R_n is the posterior moment $E(g(\theta)|X_1^n)$. For notation, let $\hat{\theta}$ be the maximum likelihood estimator of θ and denote k^{th} derivatives by superscripts (e.g. $l_n^{(1)}(\hat{\theta}) = 0$). We assume that the model and the functions w and v are sufficiently smooth to justify the Taylor expansions to follow.

First we expand the numerator N_n of (A.9). An expansion of the denominator is analagous. To start, we expand both l_n and w about $\hat{\theta}$ giving

$$l_n(\theta) = l_n(\hat{\theta}) + \frac{1}{2!}l_n^{(2)}(\hat{\theta})(\theta-\hat{\theta})^2 + \frac{1}{3!}l_n^{(3)}(\hat{\theta})(\theta-\hat{\theta})^3 + \cdots \tag{A.10}$$

$$w(\theta) = w(\hat{\theta})\left\{1 + \frac{w^{(1)}(\hat{\theta})}{w(\hat{\theta})}(\theta-\hat{\theta}) + \frac{w^{(2)}(\hat{\theta})}{2!w(\hat{\theta})}(\theta-\hat{\theta})^2 + \cdots\right\}. \tag{A.11}$$

Define coefficients a_k and d_k by

$$a_k = \frac{l_n^{(k)}(\hat{\theta})}{k!(-l_n^{(2)}(\hat{\theta}))^{k/2})} \quad d_k = \frac{w^{(k)}(\hat{\theta})}{k!w(\hat{\theta})(-l_n^{(2)}(\hat{\theta}))^{k/2})}.$$

These coefficients clearly depend upon n, however under suitable smoothness conditions they will be almost surely $O(1)$. We suppress their dependence upon n in the notation. Defining a new variable

$$u = \sqrt{-nl_n^{(2)}(\hat{\theta})}(\theta-\hat{\theta})$$

the numerator N_n becomes

$$N_n = c_n \int \frac{e^{-u^2/2}}{\sqrt{2\pi}}\left\{1 + \frac{ud_1}{n^{1/2}} + \frac{ud_2}{n} + \cdots\right\}\exp\left\{\frac{u^3a_3}{n^{1/2}} + \frac{u^4a_4}{n} + \cdots\right\} du \tag{A.12}$$

where

$$c_n = \frac{\sqrt{2\pi}w(\hat{\theta})e^{nl_n(\hat{\theta})}}{\sqrt{-nl_n^{(2)}(\hat{\theta})}}.$$

We can formally expand the exponential factor in the integrand giving

$$N_n = c_n \int \frac{e^{-u^2/2}}{\sqrt{2\pi}}\left\{1 + \frac{ud_1}{n^{1/2}} + \frac{u^2d_2}{n} + \cdots\right\}$$
$$\cdot\left\{1 + \left(\frac{u^3a_3}{n^{1/2}} + \frac{u^4a_4}{n} + \cdots\right) + \frac{1}{2!}\left(\frac{u^3a_3}{n^{1/2}} + \frac{u^4a_4}{n} + \cdots\right)^2 + \cdots\right\} du.$$

Collecting terms of like order in n, we have

$$N_n = c_n \int \frac{e^{-u^2/2}}{\sqrt{2\pi}} \left\{ 1 + \frac{A_1(u)}{n^{1/2}} + \frac{A_2(u)}{n} + \frac{A_3(u)}{n^{3/2}} + \frac{A_4(u)}{n^2} + O\left(\frac{1}{n^{5/2}}\right) \right\} du$$

where

$$\begin{aligned}
A_1(u) &= u^3 a_3 + u d_1 \\
A_2(u) &= u^6 \left(\frac{1}{2} a_3^2\right) + u^4 (a_4 + a_3 d_1) + u^2 d_2 \\
A_3(u) &= u^9 \left(\frac{1}{6} a_3^3\right) + u^7 \left(a_3 a_4 + \frac{1}{2} a_3^2 d_1\right) + u^5 (a_5 + a_4 d_1 + a_3 d_2) + u^3 d_3 \\
A_4(u) &= u^{12} \left(\frac{1}{12} a_3^4\right) + u^{10} \left(\frac{1}{6} a_3^3 d_1 + \frac{1}{2} a_3^2 a_4\right) + u^8 \left(\frac{1}{2} a_4^2 + a_3 a_5 + a_3 a_4 d_1 + \frac{1}{2} a_3^2 d_2\right) \\
&\quad + u^6 (a_6 + a_5 d_1 + a_4 d_2 + a_3 d_3) + u^4 d_4 .
\end{aligned}$$

Note that for odd integers k, $A_k(u)$ is an odd function of u. Together with the fact that

$$\int u^{2k} e^{-u^2/2} \, du = \frac{\sqrt{2\pi}(2k)!}{2^k k!} ,$$

we get, by integrating term by term in the expression for N_n above, that

$$N_n = c_n \left\{ 1 + \frac{\alpha_2}{n} + \frac{\alpha_4}{n^2} + O\left(\frac{1}{n^3}\right) \right\} \quad a.s.$$

where

$$\begin{aligned}
\alpha_2 &= d_2 + 3(a_4 + a_3 d_1) + \frac{15}{2} a_3^2 \\
\alpha_4 &= 3d_4 + 15(a_3 d_3 + a_4 d_2 + a_5 d_1 + a_6) + 105 \left(\frac{1}{2} a_3^2 d_1 + a_3 a_4 d_1 + a_3 a_5 + \frac{1}{2} a_4^2\right) \\
&\quad + 945 \left(\frac{1}{6} a_3^3 d_1 + \frac{1}{2} a_3^2 a_4\right) + \frac{3465}{4} a_3^4 .
\end{aligned}$$

We can produce an analagous expansion for the denominator D_n of (A.9) giving

$$D_n = c_n^* \left\{ 1 + \frac{\beta_2}{n} + \frac{\beta_4}{n^2} + O\left(\frac{1}{n^3}\right) \right\} \quad a.s.$$

where the derivatives $v^{(k)}(\hat{\theta})$ replace those of w in all the coefficients above. This leads finally to the expression

$$R_n = \frac{w(\hat{\theta})}{v(\hat{\theta})} \left\{ 1 + \frac{\alpha_2 - \beta_2}{n} + \frac{\alpha_4 - \beta_4}{n^2} + O\left(\frac{1}{n^3}\right) \right\} \quad a.s. . \qquad (A.13)$$

The second order term $\alpha_2 - \beta_2$ is fairly easy to write in terms of the original functions w, v, and l_n:

$$\alpha_2 - \beta_2 = \frac{1}{2l_n^{(2)}(\hat{\theta})}\left(\frac{v^{(2)}(\hat{\theta})}{v(\hat{\theta})} - \frac{w^{(2)}(\hat{\theta})}{w(\hat{\theta})}\right) + \frac{l_n^{(3)}(\hat{\theta})}{2(l_n^{(2)}(\hat{\theta}))^2}\left(\frac{w^{(1)}(\hat{\theta})}{w(\hat{\theta})} - \frac{v^{(1)}(\hat{\theta})}{v(\hat{\theta})}\right).$$

The fourth order term $\alpha_4 - \beta_4$ has quite an unweildy expression in terms of these original functions.

Taking $v(\theta)$ as a prior density and $w(\theta) = \theta v(\theta)$ in (A.13) we get the second order approximation of the posterior mean:

$$E(\theta|X_1^n) = \hat{\theta} - \frac{1}{nl_n^{(2)}(\hat{\theta})}\left\{\frac{v^{(1)}(\hat{\theta})}{v(\hat{\theta})} - \frac{l_n^{(3)}(\hat{\theta})}{2l_n^{(2)}(\hat{\theta})}\right\} + O\left(\frac{1}{n^2}\right) \quad a.s.. \qquad \text{(A.14)}$$

Because it involves third derivatives of the log likelihood function, (A.14) is difficult to generalize to higher dimensions. To avoid this problem, Tierney and Kadane (1986) have developed an $O(1/n^2)$ approximation to the same posterior moment by expanding each integrand (for the numerator and the denominator) about its mode. The approximation (A.14) will suffice for our purposes since we are mainly interested in the role of the prior in this expression.

A.4 The Inverse Function Theorem

This well-known theorem from analysis (see for example Rudin, 1964, page 193) is key to the general proof of conditional consistency and so is stated here for completeness (as in Foutz, 1977).

First, for a square matrix M define $\| M \|$ to be the least upper bound of $|Mx|$ over all vectors x with $|x| \leq 1$. Here $|\cdot|$ is Euclidean distance.

Inverse Function Theorem *Suppose that f is a mapping from an open set $\Theta \subset \mathbf{R}^r$ into $\mathbf{R}^r$, $r \geq 1$, that the partial derivatives of f exist and are continuous on Θ, and that the matrix of derivatives $f'(\theta^*)$ has inverse $f'(\theta^*)^{-1}$ at some point $\theta^* \in \Theta$. Define*

$$\lambda = \left(4 \| f'(\theta^*)^{-1} \|\right)^{-1},$$

and the ball $U_\delta \subset \Theta$ of θ^* of radius $\delta > 0$ sufficiently small so that

$$\| f'(\theta) - f'(\theta^*) \| < 2\lambda \qquad \text{whenever } \theta \in U_\delta .$$

Then for every θ_1 and θ_2 in U_δ,

$$|f(\theta_1) - f(\theta_2)| \geq 2\lambda|\theta_1 - \theta_2|$$

and the image set $f(U_\delta)$ contains the open neighborhood with radius $\lambda\delta$ about $f(\theta^*)$.

The theorem implies that f is one-to-one on U_δ and that its inverse f^{-1} is well defined on the image set $f(U_\delta)$.

A.5 Miscellaneous asymptotics

Except for Theorem 13, the results in this section are all *well known* and have been proved elsewhere. They are, however, particularly important for establishing the main results of Chapter 3, and so are included for completeness. We assume that *very well known* results, like the strong law of large numbers for example, do not require statement here.

The first result about maxima is well known in the sense that it is a homework problem in several probability texts.

Lemma 14 *Let $X, X_1, X_2, \cdots$ be a sequence of iid random variables. $E|X| < \infty$ if and only if*

$$\frac{M_n}{n} := \frac{1}{n} \max_{1 \leq i \leq n} |X_i| \to_{a.s.} 0 .$$

PROOF. Suppose that $M_n/n \to_{a.s.} 0$. Since each X_i is no larger than the maximum, it is true that $|X_n|/n \to_{a.s.} 0$, or equivalently that for any $\epsilon > 0$

$$P(|X_n| > n\epsilon \ i.o.) = 0 \ .$$

Setting $\epsilon = 1$, and using the Borel Cantelli lemmas

$$\sum_{n=1}^{\infty} P(|X_n| > n) < \infty \ .$$

Since the X_i's are identically distributed

$$\sum_{n=1}^{\infty} P(|X| > n) < \infty$$

The finite moment result follows when we recall another well known inequality

$$\sum_{n=1}^{\infty} P(|X| > n) \leq E|X| \leq \sum_{n=0}^{\infty} P(|X| > n)$$

For the converse, suppose that $E|X| < \infty$, and fix $\epsilon > 0$. By the inequality above,

$$\begin{aligned} \frac{E|X|}{\epsilon} < \infty &\Rightarrow \sum_{n=1}^{\infty} P(\,|X| > n\epsilon\,) < \infty \\ &\Rightarrow \sum_{n=1}^{\infty} P(\,|X_n| > n\epsilon\,) < \infty \end{aligned}$$

Therefore, by the first Borel-Cantelli lemma,

$$P(|X_n| > n\epsilon \ i.o.) = 0$$

or equivalently, $X_n/n \rightarrow_{a.s.} 0$ as $n \rightarrow \infty$. For a point ω where the convergence happens, there exists $N = N(\omega, \epsilon)$ such that $|X_n(\omega)/n| < \epsilon/2$ for all $n > N$. For $n > N$

$$\begin{aligned} \frac{M_n(\omega)}{n} &= \frac{M_N(\omega)}{n} \vee \frac{1}{n} \max_{1 \leq i \leq n} |X_i(\omega)| \\ &\leq \frac{M_N(\omega)}{n} + \max_{1 \leq i \leq n} \frac{|X_i(\omega)|}{i} \\ &< \frac{M_N(\omega)}{n} + \frac{\epsilon}{2} \end{aligned}$$

by convergence of $X_n(\omega)/n$ to 0. Moreover, since $M_N(\omega)$ does not depend upon n, we can make n large enough so that $M_N(\omega)/n < \epsilon/2$ and consequently $M_n(\omega)/n < \epsilon$. As this is true for almost every ω, the result follows. □

The next theorem is very useful in studying the asymptotic behaviour of weighted likelihood.

Theorem 13 *Let $Y_1, Y_2, \cdots$ be iid random variables with mean 1 and finite variance σ^2. If $a_{n,i}$ $i = 1, 2, \cdots n$ form a triangular array of real numbers satisfying*

$$\frac{1}{n}\sum_{i=1}^{n} |a_{n,i}| \to c \quad 0 < c < \infty \qquad \text{and} \qquad \frac{1}{n}\max_{1\le i\le n} |a_{n,i}| \to 0$$

as $n \to \infty$ then

$$\frac{1}{n}\sum_{i=1}^{n} a_{n,i}Y_i \to_P \alpha$$

where $\alpha = \lim_n \sum_i a_{n,i}/n$.

PROOF. Let $\epsilon > 0$ be given and put $\bar{a}_n = \sum a_{n,i}/n$. By Markov's inequality

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} a_{n,i}Y_i - \alpha\right| \ge \epsilon\right) &\le \frac{1}{\epsilon^2} E\left(\frac{1}{n}\sum_{i=1}^{n} a_{n,i}Y_i - \alpha\right)^2 \\ &= \frac{1}{\epsilon^2} E\left((\bar{a}_n - \alpha) + \frac{1}{n}\sum_{i=1}^{n} a_{n,i}(Y_i - 1)\right)^2 \\ &= \frac{1}{\epsilon^2}(\bar{a}_n - \alpha)^2 + \frac{\sigma^2}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n} a_{n,i}^2 . \end{aligned}$$

Since $\bar{a}_n \to \alpha$, it suffices to prove

$$\frac{1}{n^2}\sum_{i=1}^{n} a_{n,i}^2 \to 0$$

which follows readily from the assumptions of the theorem. □

A theorem of Chow is stronger than the *in probability* result of Theorem 13.

Theorem 14 *(Chow 1966) If $X_1, X_2, \cdots, X_n$ are iid with mean 0 and some finite variance, and if constants $(a_{n,i})$ satisfy $\sum_{i=1}^{n} a_{n,i}^2 = 1$ for all $n \ge 1$, then*

$$n^{-1/2}\sum_{i=1}^{n} a_{n,i}X_i \to_{a.s.} 0$$

as $n \to \infty$.

For an extension of this theorem, see Thrum, 1987.

The following central limit theorem which is a consequence of the Lindeberg-Feller-Lévy Central Limit Theorem is used in the proof of Theorem 7.

Theorem 15 *Let $Z_1, Z_2, \cdots$ be iid random variables with mean μ and variance σ^2. Let $(a_{n,i})$ be a non-vanishing sequence of constants satisfying*

$$\frac{\sum_{i=1}^n a_{n,i}^2}{\max_{1\le i\le n} a_{n,i}^2} \to \infty \tag{A.15}$$

as $n \to \infty$. Put $T_n = \sum_{i=1}^n a_{n,i} Z_i$, $\mu_n = \mu \sum_{i=1}^n a_{n,i}$, and $\sigma_n^2 = \sigma^2 \sum_{i=1}^n a_{n,i}^2$. Then

$$\frac{T_n - \mu_n}{\sigma_n} \to_d N(0,1)$$

as $n \to \infty$.

PROOF. Show that $Y_{n,i} = a_{n,i} Z_i$ satisfies the conditions of the Lindeberg-Feller-Lévy Central Limit Theorem. □

In studying weighted likelihood for multiparameter models, we use the following Euclidean norm:

$$\| A \| = (trace\{A^T A\})^{1/2}$$

Lemma 15 *Let $\{A_n\}$ be a sequence of matrices such that for all i, j, the sequence of $(A_n)_{ij}$ converges to a number A_{ij}; the matrix of such numbers denoted simply by A, then as $n \to \infty$*

$$\| A_n - A \| \to 0$$

PROOF. The lemma follows from equivalence of norms in finite dimensional Banach spaces. □

Appendix B

A COROLLARY TO THE ERGODIC THEOREM

B.1 Preliminaries

Suppose that $X = (X_1, X_2, \cdots)$ is a real-valued, stationary process on a probability space $(\Omega, \mathcal{A}, P)$. A set $A \in \mathcal{A}$ is *invariant* if there exists a set $B \in \mathcal{B}^\infty$ (the Borel σ–field on $\mathbb{R}^\infty$) such that $\forall n \geq 1$

$$A = \{\omega \in \Omega : (X_n(\omega), X_{n+1}(\omega), \cdots) \in B\}.$$

The collection $\mathcal{T} \subset \mathcal{A}$ of all invariant events is a σ–field and is referred to as the invariant σ–field. The famous Ergodic Theorem can be stated as follows (from Breiman 1968):

Ergodic Theorem *If* $E|X_1| < \infty$, *then as* $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} E(X_1|\mathcal{T}).$$

The stationary process X is called *ergodic* if the invariant σ–field contains only sets of probability 0 or 1. If so, then the limit above becomes $E(X_1)$. The Ergodic Theorem thus generalizes one direction of the strong law of large numbers.

The ergodic hypothesis originated in the work of Boltzmann and Maxwell as a proposition about the nature of physical systems. The Ergodic Theorem, a formal statement of this hypothesis, was first rigourously proved by Birkhoff in 1931. For an early history see Brush (1967).

B.2 The main result

The Ergodic Theorem has the following interesting corollary.

Corollary *Let* $Z = (Z_1, Z_2, \cdots)$ *be a real-valued, stationary, ergodic process, and let* Y *be a random vector independent of* Z. *If the real-valued, measurable function* f

satisfies

$$E|f(Y, Z_1)| < \infty,$$

then as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} f(Y, Z_i) \to_{a.s.} E(f(Y, Z_1)|\sigma(Y)).$$

PROOF. The random vector Y is defined on a probability space $(\Omega_0, \mathcal{A}_0, P_0)$ and thus induces a measure $\hat{P}_0$ on the measurable space $(\mathbb{R}^K, \mathcal{B}^K)$. Similarly Z is defined on $(\Omega_1, \mathcal{A}_1, P_1)$ and induces a measure $\hat{P}_1$ $(\mathbb{R}^\infty, \mathcal{B}^\infty)$. More generally, Y and Z each live on the product probability space

$$(\Omega, \mathcal{A}, P) := (\Omega_0 \times \Omega_1, \mathcal{A}_0 \times \mathcal{A}_1, P_0 \times P_1).$$

The measure P is the product measure by independence of Y and Z. By definition,

$$\begin{aligned}\sigma(Y) &:= \{A \in \mathcal{A} : \ A = \{\omega \in \Omega : \ Y(\omega) \in B \text{ for some } B \in \mathcal{B}^K\}\} \\ &= \{A \in \mathcal{A} : \ A = A_0 \times \Omega_1 \text{ for some } A_0 \in \mathcal{A}_0\};\end{aligned}$$

the latter being a consequence of the independence of Y and Z.

We can define a new process $X = (X_1, X_2, \cdots)$ on the product space $(\Omega, \mathcal{A}, P)$ by

$$X_i := f(Y, Z_i).$$

That X is stationary is an immediate consequence of stationarity of Z and independence of Z and Y. Therefore as $n \to \infty$, we have by the Ergodic Theorem that

$$\frac{1}{n}\sum_{i=1}^{n} f(Y, Z_i) = \frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} E(X_1|\mathcal{T})$$

where $\mathcal{T}$ is the invariant σ–field for the process X. It suffices to show that

$$E(X_1|\mathcal{T}) = E(X_1|\sigma(Y)) \ \ a.s.\,[P]. \tag{B.1}$$

To do this, we study the contents of the two σ–fields $\sigma(Y) \subset \mathcal{A}$ and $\mathcal{T} \subset \mathcal{A}$.

Actually, $\sigma(Y)$ and $\mathcal{T}$ differ only by null sets. That is, if $A \in \mathcal{T}$ there exists $E \in \sigma(Y)$ such that $P(A \triangle E) = 0$. By lemma 16, proving this is equivalent to proving that if $A \in \mathcal{T}$, there exists $E \in \sigma(Y)$ such that

$$P(A) = P(E) = P(A \cap E)\,.$$

To do so, we first study the structure of sets $A \in \mathcal{T}$. Typically, A is not a product set like those in $\sigma(Y)$. However, A can be formed from the following *sections*:

$$A_{\omega_0} = \{\omega_1 \in \Omega_1 : \ (\omega_0, \omega_1) \in A\} \ \text{ for } \omega_0 \in \Omega_0\,.$$

It is certainly true that for each $\omega_0 \in \Omega_0$, $A_{\omega_0} \in \mathcal{A}_1$, (Billingsley, 1986, Theorem 18.1). Moreover, because $A \in \mathcal{T}$ there exists $B \in \mathcal{B}^\infty$ such that for all $n \geq 1$

$$A = \{\omega \in \Omega : \ (f(y, Z_n(\omega_1)), f(y, Z_{n+1}(\omega_1)), \cdots) \in B, \ \ y = Y(\omega_0)\}\,.$$

Therefore with $y = Y(\omega_0)$,

$$A_{\omega_0} = \{\omega_1 \in \Omega_1 : \ (f(y, Z_n(\omega_1)), f(y, Z_{n+1}(\omega_1)), \cdots) \in B\} \ \ \forall n \geq 1\,.$$

Define the set $C_y \subset \mathbb{R}^\infty$ by

$$C_y = \{z \in \mathbb{R}^\infty : \ (f(y, z_1), f(y, z_2), \cdots) \in B\}\,.$$

Since f is measurable, $C_y \in \mathcal{B}^\infty$. Also,

$$A_{\omega_0} = \{\omega_1 \in \Omega_1 : \ (Z_n(\omega_1), Z_{n+1}(\omega_1), \cdots) \in C_y\} \ \ \forall n \geq 1.$$

Thus $A_{\omega_0} \in \mathcal{T}_1$, the invariant σ–field for the stationary process Z. (This is considering Z marginally on $(\Omega_1, \mathcal{A}_1, P_1)$ rather than on the joint space; $\mathcal{T}_1 \subset \mathcal{A}_1$.) Furthermore since Z is ergodic, $P_1(A_{\omega_0})$ is either 0 or 1.

We are now ready to confirm that for any $A \in \mathcal{T}$, there exists $E \in \sigma(Y)$ such that $P(E) = P(A) = P(E \cap A)$. For $A \in \mathcal{T}$, if $g(\omega) = 1_A(\omega)$ for $\omega \in \Omega$, then $g(\omega) = 1_{A_{\omega_0}}(\omega_1)$. Now

$$P(A \cap F) = \int_F g\, dP$$

$$
\begin{aligned}
&= \int_{F_0 \times \Omega_1} g \, dP \\
&= \int_{F_0} \left(\int_{\Omega_1} 1_{A_{\omega_0}}(\omega_1) \, dP_1(\omega_1) \right) dP_0(\omega_0) \quad \text{by Fubini's theorem} \\
&= \int_{F_0} P_1(A_{\omega_0}) \, dP_0 \\
&= \int_F P_1(A_{\omega_0}) \, dP \, .
\end{aligned}
$$

Putting

$$
E = \{\omega = (\omega_0, \omega_1) \in \Omega : \; P_1(A_{\omega_0}) = 1\} \, ,
$$

it follows that $P(A \cap F) = P(E \cap F)$. Thus when $F = \Omega$, we get $P(A) = P(E)$ and when $F = E$, $P(A \cap E) = P(E)$. Since $P_1(A_{\omega_0})$ is $\sigma(Y)$–measurable, $E \in \sigma(Y)$. We have proved that $\sigma(Y)$ equals $\mathcal{T}$ up to null sets.

It remains to show that $E(X_1|\mathcal{T}) = E(X_1|\sigma(Y))$ almost surely. First we show that $\sigma(Y) \subset \mathcal{T}$. For $A = A_0 \times \Omega_1 \in \sigma(Y)$ construct $B \in \mathcal{B}^\infty$ by

$$
B = \{x \in \mathbb{R}^\infty : \; x = (X_1(\omega), X_2(\omega), \cdots) \text{ for } \omega \in A\} \, .
$$

From our first perspective, Y is a function of $\omega_0 \in \Omega_0$ alone (it is constant for fixed ω_0) and similarly Z is a function of $\omega_1 \in \Omega_1$ alone. Thus we can write

$$
\begin{aligned}
B = \{x \in \mathbb{R}^\infty : \; & x = (f(y, Z_1(\omega_1)), f(y, Z_2(\omega_1)), \cdots), \\
& \text{where } y = Y(\omega_0), \; \omega_0 \in A_0 \text{ and } \omega_1 \in \Omega_1\}.
\end{aligned}
$$

Since ω_1 is *unconstrained* ($\omega \in \Omega_1$), we have for $n \geq 1$ that

$$
\begin{aligned}
B &= \{x \in \mathbb{R}^\infty : \; x = (f(y, Z_n(\omega_1)), f(y, Z_{n+1}(\omega_1)), \cdots), \\
&\qquad \text{where } y = Y(\omega_0), \; \omega_0 \in A_0 \text{ and } \omega_1 \in \Omega_1\} \\
&= \{x \in \mathbb{R}^\infty : \; x = (X_n(\omega), X_{n+1}(\omega), \cdots), \; \text{where } \omega \in A\} \, .
\end{aligned}
$$

We have demonstrated that if $A \in \sigma(Y)$ then $A \in \mathcal{T}$.

Since $\sigma(Y) \subset \mathcal{T}$, $E(X_1|\sigma(Y))$ is $\mathcal{T}$–measurable and is thus a candidate for the conditional probability of X_1 given $\mathcal{T}$. We must show that for all $A \in \mathcal{T}$, the following equation holds:

$$
\int_A E(X_1|\sigma(Y) \, dP = \int_A X_1 \, dP \, . \tag{B.2}
$$

For $A \in \mathcal{T}$ let $E \in \sigma(Y)$ be such that $P(A \triangle E) = 0$. Since $P(A \cap E^c) = 0$, (B.2) is true if

$$\int_{A \cap E} E(X_1|\sigma(Y)\, dP = \int_{A \cap E} X_1\, dP\,. \tag{B.3}$$

Furthermore since $P(E) = P(A \cap E)$, (B.3) is true if

$$\int_{E} E(X_1|\sigma(Y)\, dP = \int_{E} X_1\, dP\,. \tag{B.4}$$

Now (B.4) is true by the definition of conditional expectation since $E \in \sigma(Y)$. Therefore $E(X_1|\sigma(Y))$ is a version of the conditional expectation of X_1 given $\mathcal{T}$ completing the proof. □

Lemma 16 *Let $(\Omega, \mathcal{A}, P)$ be a probability space. For $A, B \in \mathcal{A}$,*

$$P(A \triangle B) = 0 \Longleftrightarrow P(A \cap B) = P(A) = P(B)\,.$$

PROOF. Simply use the facts

$$\begin{aligned} A &= (A \cap B) \cup (A \cap B^c), \\ B &= (A \cap B) \cup (A^c \cap B), \\ A \triangle B &= (A \cap B^c) \cup (A^c \cap B)\,. \end{aligned}$$

□

VITA

Michael Abbott Newton was born on July 19, 1964 in Baddeck, Victoria County, Nova Scotia, Canada. In June, 1982, he graduated from Riverview Rural High School, Coxheath, Nova Scotia. In the fall of that year, he entered Dalhousie University, Halifax, Nova Scotia to pursue a degree in engineering. In 1985, he received a Diploma in Engineering, but continued work on his mathematics degree instead of engineering. In June, 1986, he received a B.Sc. in mathematics and statistics combined, with first class honours from Dalhousie University. He joined the Statistics Department at the University of Washington in the fall on 1986, and earned an M. S. in Statistics in 1988, and a Ph. D. in 1991.