10

Semiparametric Likelihood Inference

10.1 Likelihood

The likelihood function is central to inference in parametric statistical models. Suppose that data y are believed to have come from a distribution F_ψ, where ψ is an unknown $p \times 1$ vector parameter. Then the likelihood for ψ is the corresponding density evaluated at y, namely

$$L(\psi) = f_\psi(y),$$

regarded as a function of ψ. This measures the plausibility of the different values of ψ which might have given rise to y, and can be used in various ways.

If further information about ψ is available in the form of a prior probability density, $\pi(\psi)$, Bayes' theorem can be used to form a posterior density for ψ given the data y,

$$\pi(\psi \mid y) = \frac{\pi(\psi) f_\psi(y)}{\int \pi(\psi) f_\psi(y)\, d\psi}.$$

Inferences regarding ψ or other quantities of interest may then be based on this density, which in principle contains all the information concerning ψ.

If prior information about ψ is not available in a probabilistic form, the likelihood itself provides a basis for comparison of different values of ψ. The most plausible value is that which maximizes the likelihood, namely the *maximum likelihood estimate*, $\hat{\psi}$. The relative plausibility of other values is measured in terms of the log likelihood $\ell(\psi) = \log L(\psi)$ by the *likelihood ratio statistic*

$$W(\psi) = 2\left\{\ell(\hat{\psi}) - \ell(\psi)\right\}.$$

A key result is that under repeated sampling of data from a regular model, $W(\psi)$ has approximately a chi-squared distribution with p degrees of freedom. This forms the basis for the primary method of calculating confidence regions in parametric models. One special feature is that the likelihood determines the shape of confidence regions when ψ is a vector.

Unlike many of the confidence interval methods described in Chapter 5, likelihood provides a natural basis for the combination of information from different experiments. If we have two independent sets of data, y and z, that bear on the same parameter, the overall likelihood is simply $L(\psi) = f(y \mid \psi)f(z \mid \psi)$, and tests and confidence intervals concerning ψ may be based on this. This type of combination is particularly useful in applications where several independent experiments are linked by common parameters; see Practical 10.1.

In applications we can often write $\psi = (\theta, \lambda)$, where the components of θ are of primary interest, while the so-called nuisance parameters λ are of secondary concern. In such situations inference for θ is based on the *profile likelihood*,

$$L_p(\theta) = \max_{\lambda} L(\theta, \lambda), \tag{10.1}$$

which is treated as if it were a likelihood. In some cases, particularly those where λ is high dimensional, the usual properties of likelihood statistics (consistency of maximum likelihood estimate, approximate chi-squared distribution of log likelihood ratio) do not apply without making an adjustment to the profile likelihood. The adjusted likelihood is

$$L_a(\theta) = L_p(\theta)|j_\lambda(\theta, \hat{\lambda}_\theta)|^{-1/2}, \tag{10.2}$$

where $\hat{\lambda}_\theta$ is the MLE of λ for fixed θ and $j_\lambda(\psi)$ is the observed information matrix for λ, i.e. $j_\lambda(\psi) = -\partial^2\ell(\psi)/\partial\lambda\partial\lambda^T$.

Without a parametric model the definition of a parameter is more vexed. As in Chapter 2, we suppose that a parameter θ is determined by a statistical function $t(\cdot)$, so that $\theta = t(F)$ is a mean, median, or other quantity determined by, but not by itself determining, the unknown distribution F. Now the nuisance parameter λ is all aspects of F other than $t(F)$, so that in general λ is infinite dimensional. Not surprisingly, there is no unique way to construct a likelihood in this situation, and in this chapter we describe some of the different possibilities.

10.2 Multinomial-Based Likelihoods

10.2.1 Empirical likelihood

Scalar parameter

Suppose that observations $y_1, \ldots, y_n$ form a random sample from an unknown distribution F, and that we wish to construct a likelihood for a scalar parameter $\theta = t(F)$, where $t(\cdot)$ is a statistical function. One view of the EDF $\hat{F}$ is that it is the nonparametric maximum likelihood estimate of F, with corresponding nonparametric maximum likelihood estimate $t = t(\hat{F})$ for θ (Problem 10.1). The EDF is a multinomial distribution with denominator one and probability vector $(n^{-1}, \ldots, n^{-1})$ attached to the y_j. We can think of this distribution as embedded in a more general multinomial distribution with arbitrary probability vector $p = (p_1, \ldots, p_n)$ attached to the data values. If F is restricted to be such a multinomial distribution, then we can write $t(p)$ rather than $t(F)$ for the function which defines θ. The special multinomial probability vector $(n^{-1}, \ldots, n^{-1})$ corresponding to the EDF is $\hat{p}$, and $t = t(\hat{p})$ is the nonparametric maximum likelihood estimate of θ. This multinomial representation was used earlier in Sections 4.4 and 5.4.2.

Restricting the model to be multinomial on the data values with probability vector p, the parameter value is $\theta = t(p)$ and the likelihood for p is $L(p) = \prod_{j=1}^{n} p_j^{f_j}$, with f_j equal to the frequency of value y_j in the sample. But, assuming there are no tied observations, all f_j are equal to 1, so that $L(p) = p_1 \times \cdots \times p_n$: this is the analogue of $L(\psi)$ in the parametric case. We are interested only in $\theta = t(p)$, for which we can use the profile likelihood

$$L_{EL}(\theta) = \sup_{p:t(p)=\theta} \prod_{j=1}^{n} p_j, \tag{10.3}$$

which is called the *empirical likelihood* for θ. Notice that the value of θ which maximizes $L_{EL}(\theta)$ corresponds to the value of p maximizing $L(p)$ with only the constraint $\sum p_j = 1$, that is $\hat{p}$. In other words, the empirical likelihood is maximized by the nonparametric maximum likelihood estimate t.

In (10.3) we maximize over the p_j subject to the constraints imposed by fixing $t(p) = \theta$ and $\sum p_j = 1$, which is effectively a maximization over $n - 2$ quantities when θ is scalar. Remarkably, although the number of parameters over which we maximize is comparable with the sample size, the approximate distributional results from the parametric situation carry over. Let θ_0 be the true value of θ, with T the maximum empirical likelihood estimator. Then under mild conditions on F and in large samples, the empirical likelihood ratio statistic

$$W_{EL}(\theta_0) = 2\{\log L_{EL}(T) - \log L_{EL}(\theta_0)\}$$

has an approximate chi-squared distribution with one degree of freedom. Although the limiting distribution of $W_{EL}(\theta_0)$ is the same as that of $W_p(\theta_0)$ under a correct parametric model, such asymptotic results are typically less useful in the nonparametric setting. This suggests that the bootstrap be used to calibrate empirical likelihood, by using quantiles of bootstrap replicates of $W_{EL}(\theta_0)$, i.e. quantiles of $W_{EL}^{*}(\hat{\theta})$. This idea is outlined below.

Example 10.1 (Air-conditioning data) We consider the empirical likelihood for the mean of the larger set of air-conditioning data in Table 5.6; $n = 24$

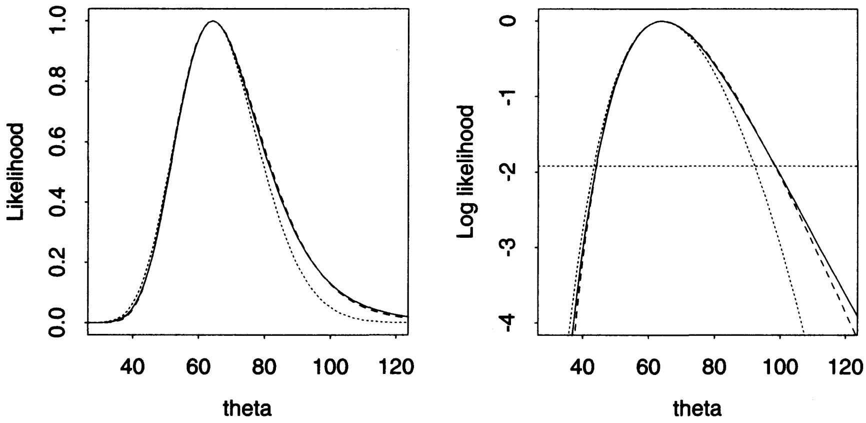

Figure 10.1 Likelihood and log likelihoods for the mean of the air-conditioning data: empirical (dots), exponential (dashes), and gamma profile (solid). Values of θ whose log likelihood lies above the horizontal dotted line in the right panel are contained in an asymptotic 95% confidence set for the true mean.

and $\bar{y} = 64.125$. The mean is $\theta = \int y\,dF(y)$, which equals $\sum_j p_j y_j$ for the multinomial distribution that puts masses p_j on the y_j. For a specified value of θ, finding (10.3) is equivalent to maximizing $\sum \log p_j$ with respect to $p_1, \ldots, p_n$, subject to the constraints that $\sum p_j = 1$ and $\sum p_j y_j = \theta$. Use of Lagrange multipliers gives $p_j \propto \{1 + \eta_\theta(y_j - \theta)\}^{-1}$, where the Lagrange multiplier η_θ is determined by θ and satisfies the equation

$$\sum_{j=1}^{n} \frac{y_j - \theta}{1 + \eta_\theta(y_j - \theta)} = 0. \tag{10.4}$$

Thus the log empirical likelihood, normalized to have maximum zero, is

$$\ell_{EL}(\theta) = -\sum_{j=1}^{n} \log\{1 + \eta_\theta(y_j - \theta)\}. \tag{10.5}$$

This is maximized at the sample average $\theta = \bar{y}$, where $\eta_\theta = 0$ and $p_j = n^{-1}$. It is undefined outside $(\min y_j, \max y_j)$, because no multinomial distribution on the y_j can have mean outside this interval.

Figure 10.1 shows $L_{EL}(\theta)$, which is calculated by successive solution of (10.4) to yield η_θ at values of θ small steps apart. The exponential likelihood and gamma profile likelihood for the mean are also shown. As we should expect, the gamma profile likelihood is always higher than the exponential likelihood, which corresponds to the gamma likelihood but with shape parameter $\kappa = 1$. Both parametric likelihoods are wider than the empirical likelihood. Direct comparison between parametric and empirical likelihoods is misleading, however, since they are based on different models, and here and in later figures

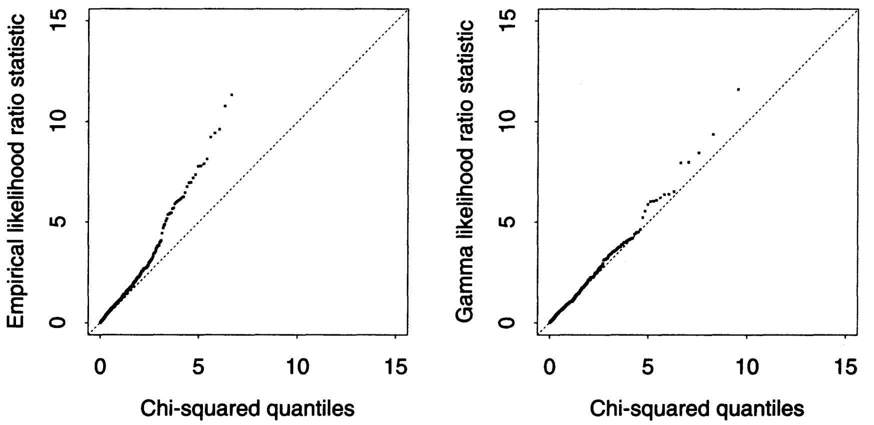

Figure 10.2 Simulated empirical likelihood ratio statistics (left panel) and gamma profile likelihood ratio statistics (right panel) for exponential samples of size 24. The dotted line corresponds to the theoretical χ_1^2 approximation.

we give the gamma likelihood purely as a visual reference. The circumstances in which empirical and parametric likelihoods are close are discussed in Problem 10.3.

The endpoints of an approximate 95% confidence interval for θ are obtained by reading off where $\ell_{E\bar{L}}(\theta) = \frac{1}{2}c_{1,0.95}$, where $c_{d,\alpha}$ is the α quantile of the chi-squared distribution with d degrees of freedom. The interval is $(43.3, 92.3)$, which compares well with the nonparametric BC_a interval of $(42.4, 93.2)$. The likelihood ratio intervals for the exponential and gamma models are $(44.1, 98.4)$ and $(44.0, 98.6)$.

Figure 10.2 shows the empirical likelihood and gamma profile likelihood ratio statistics for 500 exponential samples of size 24. Though good for the parametric statistic, the chi-squared approximation is poor for W_{EL}, whose estimated 95% quantile is 5.92 compared to the χ_1^2 quantile of 3.84. This suggests strongly that the empirical likelihood-based confidence interval given above is too narrow. However, the simulations are only relevant when the data are exponential, in which case we would not be concerned with empirical likelihood.

We can use the bootstrap to estimate quantiles for $W_{EL}(\theta_0)$, by setting $\theta_0 = \bar{y}$ and then calculating $W^*(\theta_0)$ for bootstrap samples from the original data. The resulting Q-Q plot is less extreme than the left panel of Figure 10.2, with a 95% quantile estimate of 4.08 based on 999 bootstrap samples; the corresponding empirical likelihood ratio interval is $(42.8, 93.3)$. With a sample of size 12, 41 of the 999 simulations gave infinite values of $W_{EL}(\theta_0)$ because $\bar{y}$ did not lie within the limits $(\min y_j^*, \max y_j^*)$ of the bootstrap sample. With a sample of size 24, this problem did not arise.

Vector parameter

In principle, empirical likelihood is straightforward to construct when θ has dimension $d < n-1$. Suppose that $\theta = (\theta_1, \ldots, \theta_d)^T$ is determined implicitly as the root of the simultaneous equations

$$\int u(\theta; y)\, dF(y) = 0, \quad i = 1, \ldots, d,$$

where $u(\theta; y)$ is a $d \times 1$ vector whose ith element is $u_i(\theta; y)$. Then the estimate $\hat{\theta}$ is the solution to the d estimating equations

$$\sum_{j=1}^{n} u(\hat{\theta}; y_j) = 0. \tag{10.6}$$

An extension of the argument in Example 10.1, involving the vector of Lagrange multipliers $\eta_\theta = (\eta_{\theta 1}, \ldots, \eta_{\theta d})^T$, shows that the log empirical likelihood is

$$\ell_{EL}(\theta) = -\sum_{j=1}^{n} \log\left\{1 + \eta_\theta^T u_j(\theta)\right\}, \tag{10.7}$$

where $u_j(\theta) \equiv u(\theta; y_j)$. The value of η_θ is determined by θ through the simultaneous equations

$$\sum_{j=1}^{n} \frac{u_j(\theta)}{1 + \eta_\theta^T u_j(\theta)} = 0. \tag{10.8}$$

The simplest approximate confidence region for the true θ is the set of values such that $W_{EL}(\theta) \le c_{d,1-\alpha}$, but in small samples it will again be preferable to replace the χ_d^2 quantile by its bootstrap estimate.

10.2.2 Empirical exponential family likelihoods

Another data-based multinomial likelihood can be based on an empirical exponential family construction. Suppose that $\hat{\theta}_1, \ldots, \hat{\theta}_d$ are defined as the solutions to the equations (10.6). Then rather than putting probability $n^{-1}\{1+\eta_\theta^T u_j(\theta)\}^{-1}$ on y_j, corresponding to (10.7), we can take probabilities proportional to $\exp\{\xi_\theta^T u_j(\theta)\}$; this is the exponential tilting construction described in Example 4.16 and in Sections 5.3 and 9.4. Here $\xi_\theta = (\xi_{\theta 1}, \ldots, \xi_{\theta d})^T$ is determined by θ through

$$\sum_{j=1}^{n} u_j(\theta) \exp\left\{\xi_\theta^T u_j(\theta)\right\} = 0. \tag{10.9}$$

This is analogous to (10.8), but it may be solved using a program that fits regression models for Poisson responses (Problem 10.4), which is often more convenient to deal with than the optimization problems posed by empirical likelihood. The log likelihood obtained by integrating (10.9) is $\ell_{EEF}(\theta) = \sum \exp\{\xi_\theta^T u_j(\theta)\}$. This can be close to $\ell_{EL}(\theta)$, which suggests that both the corresponding log likelihood ratio statistics share the same rather slow approach to their large-sample distributions.

In addition to likelihood ratio statistics from empirical exponential families and empirical likelihood, many other related statistics can be defined. For example, we can regard ξ_θ as the parameter in a Poisson regression model and construct a quadratic form

$$Q_{EEF}(\theta) = \left\{\sum_{j=1}^{n} u_j(\theta)\right\}^T \left\{\sum_{j=1}^{n} u_j(\theta)u_j(\theta)^T\right\}^{-1} \left\{\sum_{j=1}^{n} u_j(\theta)\right\} \tag{10.10}$$

based on the score statistic that tests the hypothesis $\xi_\theta = 0$. There is a close parallel between $Q_{EEF}(\theta)$ and the quadratic forms used to set confidence regions in Section 5.8, but the nonlinear relationship between θ and $Q_{EEF}(\theta)$ means that the contours of (10.10) need not be elliptical. As discussed there, for example, theory suggests that when the true value of θ is θ_0, $Q_{EEF}(\theta_0)$ has a large-sample χ_d^2 distribution. Thus an approximate $1-\alpha$ confidence region for θ is the set of values of θ for which $Q_{EEF}(\theta)$ does not exceed $c_{d,1-\alpha}$. And as there, it is generally better to use bootstrap estimates of the quantiles of $Q_{EEF}(\theta)$.

Example 10.2 (Laterite data) We consider again setting a confidence region based on the data in Example 5.15. Recall that the quantity of interest is the mean polar axis,

$$a(\theta,\phi) = (\cos\theta\cos\phi, \cos\theta\sin\phi, \sin\theta)^T,$$

which is the axis given by the eigenvector corresponding to the largest eigenvalue of $\mathrm{E}(YY^T)$. The data consist of positions on the lower half-sphere, or equivalently the sample values of $a(\theta,\phi)$, which we denote by y_j, $j = 1,\ldots,n$.

In order to set an empirical likelihood confidence region for the mean polar axis, or equivalently for the spherical polar coordinates (θ,ϕ), we let

$$b(\theta,\phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, -\cos\theta)^T, \qquad c(\theta,\phi) = (-\sin\phi, -\cos\phi, 0)^T$$

denote the unit vectors orthogonal to $a(\theta,\phi)$. Then since the eigenvectors of $\mathrm{E}(YY^T)$ may be taken to be orthogonal, the population values of (θ,ϕ) satisfy simultaneously the equations

$$b(\theta,\phi)^T\mathrm{E}(YY^T)a(\theta,\phi) = 0, \qquad c(\theta,\phi)^T\mathrm{E}(YY^T)a(\theta,\phi) = 0,$$

with sample equivalents

$$b(\theta,\phi)^T\left(n^{-1}\sum_{j=1}^{n} y_j y_j^T\right)a(\theta,\phi) = 0, \qquad c(\theta,\phi)^T\left(n^{-1}\sum_{j=1}^{n} y_j y_j^T\right)a(\theta,\phi) = 0.$$

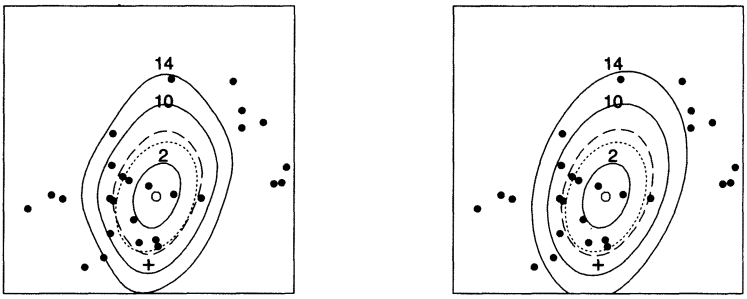

Figure 10.3 Contours of W_{EL} (left) and Q_{EE} (right) for the mean polar axis, in the squai region shown in Figure 5.10. The dashe lines show the 95% confidence regions usir bootstrap quantiles. Tł dotted ellipse is the 95' confidence region base on a studentized statist (Fisher, Lewis and Embleton, 1987, equation 6.9).

In terms of the previous general discussion, we have $d = 2$ and

$$u_j(\theta) \equiv u(\theta, \phi; y_j) = \begin{pmatrix} b(\theta,\phi)^T y_j y_j^T a(\theta,\phi) \\ c(\theta,\phi)^T y_j y_j^T a(\theta,\phi) \end{pmatrix}.$$

The left panel of Figure 10.3 shows the empirical likelihood contours based on (10.7) and (10.8), in the square region shown in Figure 5.10. The corresponding contours for $Q_{EEF}(\theta)$ are shown on the right. The dashed lines show the boundaries of the 95% confidence regions for (θ, ϕ) using bootstrap calibration; these differ little from those based on the asymptotic χ_2^2 distribution. In each panel the dotted ellipse is a 95% confidence region based on a studentized form of the sample mean polar axis, for which the contours are ellipses. The elliptical contours are appreciably tighter than those for the likelihood-based statistics.

Table 10.1 compares theoretical and bootstrap quantiles for several likelihood-based statistics and the studentized bootstrap statistic, Q, for the full data and for a random subset of size 20. For the full data, the quantiles for Q_{EEF} and W_{EL} are close to those for the large-sample χ_2^2 distribution. For the subset, Q_{EEF} is close to its nominal distribution, but the other statistics seem considerably more variable. Except for Q_{EEF}, it would be misleading to rely on the asymptotic results for the subsample. ■

Theoretical work suggests that W_{EL} should have better properties than statistics such as W_{EEF} or Q_{EEF}, but since simulations do not always confirm this, bootstrap quantiles should generally be used to set the limits of confidence regions from multinomial-based likelihoods.

Table 10.1 Bootstrap p quantiles of likelihood-based statistics for mean polar axis data.

		Full data, $n = 50$				Subset, $n = 20$			
p	χ^2_2	Q	W_{EL}	W_{EEF}	Q_{EEF}	Q	W_{EL}	W_{EEF}	Q_{EEF}
0.80	3.22	3.23	3.40	3.37	3.15	3.67	3.70	3.61	3.15
0.90	4.61	4.77	4.81	5.05	4.69	5.39	5.66	5.36	4.45
0.95	5.99	6.08	6.18	6.94	6.43	7.17	7.99	10.82	7.03

10.3 Bootstrap Likelihood

Basic idea

Suppose for simplicity that our data $y_1, \ldots, y_n$ form a homogeneous random sample for which statistic T takes value t. If the data were governed by a parametric model under which T had the density $f_T(\cdot\,;\theta)$, then a partial likelihood for θ based on T would be $f_T(t;\theta)$ regarded as a function of θ. In the absence of a parametric model, we may estimate the density of T at t, for different values of θ, by means of a nonparametric double bootstrap.

To be specific, suppose that we generate a first-level bootstrap sample $y_1^*, \ldots, y_n^*$ from $y_1, \ldots, y_n$, with corresponding estimator value t^*. This bootstrap sample is now considered as a population whose parameter value is t^*; the empirical distribution of $y_1^*, \ldots, y_n^*$ is the nonparametric analogue of a parametric model with $\theta = t^*$. We then generate M second-level bootstrap samples by sampling from our first-level sample, and calculate the corresponding values of T, namely $t_1^{**}, \ldots, t_M^{**}$. Kernel density estimation based on these second-level values provides an approximate density for T^{**}, and by analogy with parametric partial likelihood we take this density at $t^{**} = t$ to be the value of a nonparametric partial likelihood at $\theta = t^*$. If the density estimate uses kernel $w(\cdot)$ with bandwidth h, then this leads to the *bootstrap likelihood* value at $\theta = t^*$ given by

$$L(t^*) = f_{T^{**}}(t \mid t^*) = \frac{1}{Mh} \sum_{m=1}^{M} w\left(\frac{t_m^{**} - t}{h}\right). \qquad (10.11)$$

On repeating this procedure for R different first-level bootstrap samples, we obtain R approximate likelihood values $L(t_r^*)$, $r = 1, \ldots, R$, from which a smooth likelihood curve $L_B(\theta)$ can be produced by nonparametric smoothing.

Computational improvements

There are various ways to reduce the large amount of computation needed to obtain a smooth curve. One, which was used earlier in Section 3.9.2, is to generate second-level samples from smoothed versions of the first-level samples. As before, probability distributions on the values $y_1, \ldots, y_n$ are denoted by vectors $p = (p_1, \ldots, p_n)$, and parameter values are expressed as $t(p)$; recall that $\hat{p} = (\frac{1}{n}, \ldots, \frac{1}{n})$ and $t = t(\hat{p})$. The rth first-level bootstrap sample gives statistic value t_r^*, and the data value y_j occurs with frequency $f_{rj}^* = np_{rj}^*$, say. In the bootstrap likelihood calculation this bootstrap sample is considered as a population with probability distribution $p_r^* = (p_{r1}^*, \ldots, p_{rn}^*)$ on the data values, and $t_r^* = t(p_r^*)$ is considered as the θ-value for this population.

In order to obtain populations which vary smoothly with θ, we apply kernel smoothing to the p_r^*, as in Section 3.9.2. Thus for target parameter value θ^0 we define the vector $p^*(\theta^0)$ of probabilities

$$p_j^*(\theta^0) \propto \frac{1}{R\varepsilon} \sum_{r=1}^{R} w\left(\frac{\theta^0 - t_r^*}{\varepsilon}\right) p_{rj}^*, \quad j = 1, \ldots, n, \tag{10.12}$$

where typically $w(\cdot)$ is the standard normal density and $\varepsilon = v_L^{1/2}$; as usual v_L is the nonparametric delta method variance estimate for t. The distribution $p^*(\theta^0)$ will have parameter value not θ^0 but $\theta = t\left(p^*(\theta^0)\right)$. With the understanding that θ is defined in this way, we shall for simplicity write $p^*(\theta)$ rather than $p^*(\theta^0)$. For a fixed collection of R first-level samples and bandwidth $\varepsilon > 0$, the probability vectors $p^*(\theta)$ change gradually as θ varies over its range of interest.

Second-level bootstrap sampling now uses vectors $p^*(\theta)$ as sampling distributions on the data values, in place of the p_r^*s. The second-level sample values t^{**} are then used in (10.11) to obtain $L_B(\theta)$. Repeating this calculation for, say, 100 values of θ in the range $t \pm 4v_L^{1/2}$, followed by smooth interpolation, should give a good result.

Experience suggests that the value $\varepsilon = v_L^{1/2}$ is safe to use in (10.12) if the t_r^* are roughly equally spaced, which can be arranged by weighted first-level sampling, as outlined in Problem 10.6.

A way to reduce further the amount of calculation is to use recycling, as described in Section 9.4.4. Rather than generate second-level samples from each $p^*(\theta)$ of interest, one set of M samples can be generated using distribution p on the data values, and the associated values $t_1^{**}, \ldots, t_M^{**}$ calculated. Then, following the general re-weighting method (9.24), the likelihood values are calculated as

$$L_B(\theta) = \frac{1}{Mh} \sum_{m=1}^{M} w\left(\frac{t_m^{**} - t}{h}\right) \prod_{j=1}^{n} \left\{\frac{p_j^*(\theta)}{p_j}\right\}^{f_{jm}^{**}}, \tag{10.13}$$

where f_{jm}^{**} is the frequency of the jth case in the mth second-level bootstrap sample. One simple choice for p is the EDF $\hat{p}$. In special cases it will be possible to replace the second level of sampling by use of the saddlepoint approximation method of Section 9.5. This would give an accurate and smooth approximation to the density of T^{**} for sampling from each $p^*(\theta)$.

Example 10.3 (Air-conditioning data) We apply the ideas outlined above to

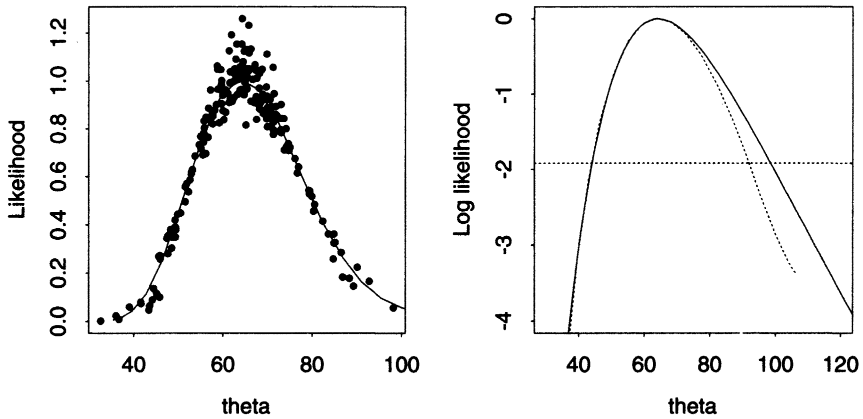

Figure 10.4 Bootstrap likelihood for mean of air-conditioning data. Left panel: bootstrap likelihood values obtained by saddlepoint approximation for 200 random samples, with smooth curve fitted to values obtained by smoothing frequencies from 1000 bootstrap samples. Right panel: gamma profile log likelihood (solid) and bootstrap log likelihood (dots).

the data from Example 10.1. The solid points in the left panel of Figure 10.4 are bootstrap likelihood values for the mean θ for 200 resamples, obtained by saddlepoint approximation. This replaces the kernel density estimate (10.11) and avoids the second level of resampling, but does not remove the variation in estimated likelihood values for different bootstrap samples with similar values of t_r^*. A locally quadratic nonparametric smoother (on the log likelihood scale) could be used to produce a smooth likelihood curve from the values of $L(t_r^*)$, but another approach is better, as we now describe.

The solid line in the left panel of Figure 10.4 interpolates values obtained by applying the saddlepoint approximation using probabilities (10.12) at a few values of θ^0. Here the values of t_r^* are generated at random, and we have taken $\varepsilon = 0.5v_L^{1/2}$; the results depend little on the value of ε.

The log bootstrap likelihood is very close to log empirical likelihood, with 95% confidence interval (43.8, 92.1). ■

Bootstrap likelihood is based purely on resampling and smoothing, which is a potential advantage over empirical likelihood. However, in its simplest form it is more computer-intensive. This precludes bootstrapping to estimate quantiles of bootstrap likelihood ratio statistics, which would involve three levels of nested resampling.

10.4 Likelihood Based on Confidence Sets

In certain circumstances it is possible to view confidence intervals as being approximately posterior probability sets, in the Bayesian sense. This encourages the idea of defining a confidence distribution for θ from the set of confidence limits, and then taking the PDF of this distribution as a likelihood function. That is, if we define the confidence distribution function C by $C(\hat{\theta}_\alpha) = \alpha$, then the associated likelihood would be the "density" $dC(\theta)/d\theta$. Leaving the philosophical arguments aside, we look briefly at where this idea leads in the context of nonparametric bootstrap methods.

10.4.1 Likelihood from pivots

Suppose that $Z(\theta) = z(\theta, \hat{F})$ is a pivot, with CDF $K(z)$ not depending on the true distribution F, and that $z(\theta)$ is a monotone function of θ. Then the confidence distribution based on confidence limits derived from z leads to the likelihood

$$L^\dagger(\theta) = |\dot{z}(\theta)| k\{z(\theta)\}, \qquad (10.14)$$

$\dot{z}(\theta)$ equals $\partial z(\theta)/\partial\theta$.

where $k(z) = dK(z)/dz$. Since k will be unknown in practice, it must be estimated.

In fact this definition of likelihood has a hidden defect. If the identification of confidence distribution with posterior distribution is accurate, as it is to a good approximation in many cases, then the effect of some prior distribution has been ignored in (10.14). But this effect can be removed by a simple device. Consider an imaginary experiment in which a random sample of size $2n$ is obtained, with outcome exactly two copies of the data y that we have. Then the likelihood would be the square of the likelihood $L_Z(\theta \mid y)$ we are trying to calculate. The ratio of the corresponding posterior densities would be simply $L_Z(\theta \mid y)$. This argument suggests that we apply the confidence density (10.14) twice, first with data y to give $L_n^\dagger(\theta)$, say, and second with data (y, y) to give $L_{2n}^\dagger(\theta)$. The ratio $L_{2n}^\dagger(\theta)/L_n^\dagger(\theta)$ will then be a likelihood with the unknown prior effect removed. In an explicit notation, this definition can be written

$$L_Z(\theta) = \frac{L_{2n}^\dagger(\theta)}{L_n^\dagger} = \frac{|\dot{z}_{2n}(\theta)| k_{2n}\{z_{2n}(\theta)\}}{|\dot{z}_n(\theta)| k_n\{z_n(\theta)\}}, \qquad (10.15)$$

where the subscripts indicate sample size. Note that $\hat{F}$ and t are the same for both sample sizes, but quantities such as variance estimates will depend upon sample size. Note also that the implied prior is estimated by $L_n^{\dagger 2}(\theta)/L_{2n}^\dagger(\theta)$.

Example 10.4 (Exponential mean) If data $y_1, \ldots, y_n$ are sampled from an exponential distribution with mean θ, then a suitable choice for $z(\theta, \hat{F})$ is $\bar{y}/\theta$. The gamma distribution for $\bar{Y}$ can be used to check that the original definition (10.14) gives $L^\dagger(\theta) = \theta^{-n-1}\exp(-n\bar{y}/\theta)$, whereas the true likelihood is $\theta^{-n}\exp(-n\bar{y}/\theta)$. The true result is obtained exactly using (10.15). The implied prior is $\pi(\theta) \propto \theta^{-1}$, for $\theta > 0$.

In practice the distribution of Z must be estimated, in general by bootstrap sampling, so the densities k_n and k_{2n} in (10.15) must be estimated. To be specific, consider the particular case of the studentized quantity $z(\theta) = (t-\theta)/v_L^{1/2}$. Apart from a constant multiplier, the definition (10.15) gives

$$L^{\dagger}(\theta) = k_{2n}\left(\frac{t-\theta}{v_{2n,L}^{1/2}}\right) \Big/ k_n\left(\frac{t-\theta}{v_{n,L}^{1/2}}\right), \qquad (10.16)$$

where $v_{n,L} = v_L$ and $v_{2n,L} = \frac{1}{2}v_L$, and we have used the fact that t is the estimate for both sample sizes. The densities k_n and k_{2n} are approximated using bootstrap sample values as follows. First R nonparametric samples of size n are drawn from $\hat{F}$ and corresponding values of $z_n^* = (t_n^* - t)/v_{n,L}^{*1/2}$ calculated. Then R samples of size $2n$ are drawn from $\hat{F}$ and values of $z_{2n}^* = (t_{2n}^* - t)/(v_{2n,L}^*)^{1/2}$ calculated. Next kernel estimates for k_n and k_{2n}, with bandwidths h_n and h_{2n} respectively, are obtained and substituted in (10.16). For example,

$$\hat{k}_n\left(\frac{t-\theta}{v_{n,L}^{1/2}}\right) = \frac{1}{h_n R}\sum_{r=1}^{R} w\left(\frac{t-\theta-v_{n,L}^{1/2}z_{n,r}^*}{h_n v_{n,L}^{1/2}}\right). \qquad (10.17)$$

In practice these values can be computed via spline smoothing from a dense set of values of the kernel density estimates $\hat{k}_n(z)$.

There are difficulties with this method. First, just as with bootstrap likelihood, it is necessary to use a large number of simulations R. A second difficulty is that of ascertaining whether or not the chosen Z is a pivot, or else what prior transformation of T could be used to make Z pivotal; see Section 5.2.2. This is especially true if we extend (10.16) to vector θ, which is theoretically possible. Note that if the studentized bootstrap is applied to a transformation of t rather than t itself, then the factor $|\dot{z}(\theta)|$ in (10.14) can be ignored when applying (10.16).

10.4.2 Implied likelihood

In principle any bootstrap confidence limit method can be turned into a likelihood method via the confidence distribution, but it makes sense to restrict attention to the more accurate methods such as the studentized bootstrap used above. Section 5.4 discusses the underlying theory and introduces one other method, the *ABC* method, which is particularly easy to use as basis for a likelihood because no simulation is required.

First, a confidence density is obtained via the quadratic approximation (5.42), with a, b and c as defined for the nonparametric *ABC* method in (5.49). Then, using the argument that led to (10.15), it is possible to show that the induced likelihood function is

$$L_{ABC}(\theta) = \exp\{-\tfrac{1}{2}u^2(\theta)\}, \qquad (10.18)$$

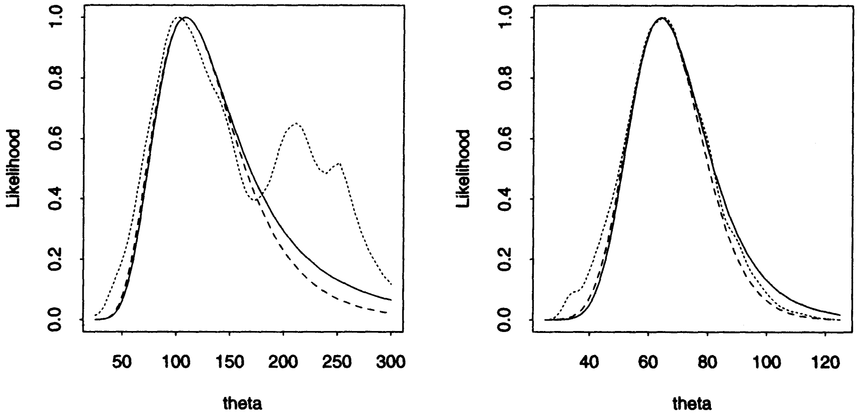

Figure 10.5 Gamma profile likelihood (solid) implied likelihood L_{ABC} (dashes) and pivot-based likelihood (dots) for air-conditioning dataset of size 12 (left panel) and size 24 (right panel). The pivot-based likelihood uses $R = 9999$ simulations and bandwidths 1.0.

where

$$u(\theta) = \frac{2r(\theta)}{1 + 2ar(\theta) + \{1 + 4ar(\theta)\}^{1/2}}, \quad r(\theta) = -\frac{2z(\theta)}{1 + \{1 - 4cz(\theta)\}^{1/2}},$$

with $z(\theta) = (t - \theta)/v_L^{1/2}$ as before. This is called the *implied likelihood.* Based on the discussion in Section 5.4, one would expect results similar to those from applying (10.16).

A further modification is to multiply $L_{ABC}(\theta)$ by $\exp\{(cv_L^{1/2} - b)\theta/v_L\}$, with b the bias estimate defined in (5.49). The effect of this modification is to make the likelihood even more compatible with the Bayesian interpretation, somewhat akin to the adjusted profile likelihood (10.2).

Example 10.5 (Air-conditioning data) Figure 10.5 shows confidence likelihoods for the two sets of air-conditioning data in Table 5.6, samples of size 12 and 24 respectively. The implied likelihoods $L_{ABC}(\theta)$ are similar to the empirical likelihoods for these data. The pivotal likelihood $L_Z(\theta)$, calculated from $R = 9999$ samples with bandwidths equal to 1.0 in (10.17), is clearly quite unstable for the smaller sample size. This also occurred with bootstrap likelihood for these data and seems to be due to the discreteness of the simulations with so small a sample. ■

10.5 Bayesian Bootstraps

All the inferences we have described thus far have been frequentist: we have summarized uncertainty in terms of confidence regions for the unknown parameter θ of interest, based on repeated sampling from a distribution F. A quite different approach is possible if prior information is available regarding F. Suppose that the only possible values of Y are known to be $u_1, \ldots, u_N$, and that these arise with unknown probabilities $p_1, \ldots, p_N$, so that

$$\Pr(Y = u_j \mid p_1, \ldots, p_N) = p_j, \qquad \sum p_j = 1.$$

If our data consist of the random sample $y_1, \ldots, y_n$, and f_j counts how many y_i equal u_j, the probability of the observed data given the values of the p_j is proportional to $\prod_{j=1}^N p_j^{f_j}$. If the prior information regarding the p_j is summarized in the prior density $\pi(p_1, \ldots, p_N)$, the joint posterior density of the p_j given the data is proportional to

$$\pi(p_1, \ldots, p_N) \prod_{j=1}^N p_j^{f_j},$$

and this induces a posterior density for θ. Its calculation is particularly straightforward when π is the Dirichlet density, in which case the prior and posterior densities are respectively proportional to

$$\prod_{j=1}^N p_j^{a_j}, \qquad \prod_{j=1}^N p_j^{a_j+f_j};$$

the posterior density is Dirichlet also. *Bayesian bootstrap* samples and the corresponding values of θ are generated from the joint posterior density for the p_j, as follows.

Algorithm 10.1 (Bayesian bootstrap)

For $r = 1, \ldots, R$,

1 Let $G_1, \ldots, G_N$ be independent gamma variables with shape parameters $a_j + f_j + 1$, and unit scale parameters, and for $j = 1, \ldots, N$ set $P_j^\dagger = G_j/(G_1 + \cdots + G_N)$.

2 Let $\theta_r^\dagger = t(F_r^\dagger)$, where $F_r^\dagger \equiv (P_1^\dagger, \ldots, P_N^\dagger)$

Estimate the posterior density for θ by kernel smoothing of $\theta_1^\dagger, \ldots, \theta_R^\dagger$

In practice with continuous data we have $f_j \equiv 1$. The simplest version of the simulation puts $a_j = -1$, corresponding to an improper prior distribution with support on $y_1, \ldots, y_n$; the G_j are then exponential. Some properties of this procedure are outlined in Problem 10.10.

Example 10.6 (City population data) In the city population data of Example 2.8, for which $n = 10$, the parameter $\theta = t(F)$ and the rth simulated posterior value $\theta^\dagger$ are

$$t(F) = \frac{\int x \, dF(u,x)}{\int u \, dF(u,x)}, \qquad t(F_r^\dagger) = \frac{\sum x_j P_{jr}^\dagger}{\sum u_j P_{jr}^\dagger}.$$

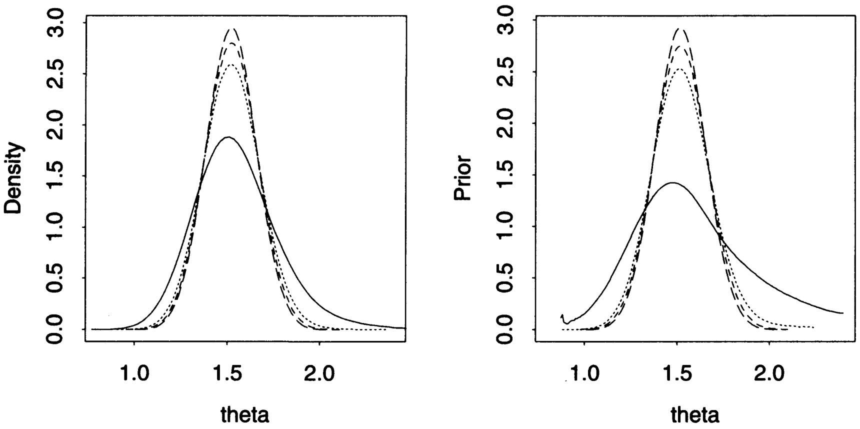

Figure 10.6 Bayesian bootstrap applied to city population data, with $n = 10$. The left panel shows posterior densities for ratio θ estimated from 999 Bayesian bootstrap simulations, with $a = -1, 2, 5, 10$; the densities are more peaked as a increases. The right panel shows the corresponding prior densities for θ.

The left panel of Figure 10.6 shows kernel density estimates of the posterior density of θ based on $R = 999$ simulations with all the a_j equal to $a = -1, 2, 5,$ and 10. The increasingly strong prior information results in posterior densities that are more and more sharply peaked.

The right panel shows the implied priors on θ, obtained using the data doubling device from Section 10.4. The priors seem highly informative, even when $a = -1$.

The primary use of the Bayesian bootstrap is likely to be for imputation when data are missing, rather than in inference for θ *per se*. There are theoretical advantages to such *weighted bootstraps*, in which the probabilities $P^{\dagger}$ vary smoothly, but as yet they have been little used in applications.

10.6 Bibliographic Notes

Likelihood inference is the core of parametric statistics. Many elementary textbooks contain some discussion of large-sample likelihood asymptotics, while adjusted likelihoods and higher-order approximations are described by Barndorff-Nielsen and Cox (1994).

Empirical likelihood was defined for single samples by Owen (1988) and extended to wider classes of models in a series of papers (Owen, 1990, 1991). Qin and Lawless (1994) make theoretical connections to estimating equations, while Hall and La Scala (1990) discuss some practical issues in using empirical likelihoods. More general models to which empirical likelihood has been applied include density estimation (Hall and Owen, 1993; Chen 1996), length-biased data (Qin, 1993), truncated data (Li, 1995), and time series (Monti, 1997). Applications to directional data are discussed by Fisher *et al.* (1996). Owen (1992a) reports simulations that compare the behaviour of the empirical likelihood ratio statistic with bootstrap methods for samples of size up to 20, with overall conclusions in line with those of Section 5.7: the studentized bootstrap performs best, in particular giving more accurate confidence intervals for the mean than the empirical likelihood ratio statistic, for a variety of underlying populations.

Related theoretical developments are due to DiCiccio, Hall and Romano (1991), DiCiccio and Romano (1989), and Chen and Hall (1993). From a theoretical viewpoint it is noteworthy that the empirical likelihood ratio statistic can be Bartlett-adjusted, though Corcoran, Davison and Spady (1996) question the practical relevance of this. Hall (1990) makes theoretical comparisons between empirical likelihood and likelihood based on studentized pivots.

Empirical likelihood has roots in certain problems in survival analysis, notably using the product-limit estimator to set confidence intervals for a survival probability. Related methods are discussed by Murphy (1995). See also Mykland (1995), who introduces the idea of dual likelihood, which treats the Lagrange multiplier in (10.7) as a parameter. Except in large samples, it seems likely that our caveats about asymptotic results apply here also.

Empirical exponential families have been discussed in Section 10.10 of Efron (1982) and DiCiccio and Romano (1990), among others; see also Corcoran, Davison and Spady (1996), who make comparisons with empirical likelihood statistics. Jing and Wood (1996) show that empirical exponential family likelihood is not Bartlett adjustable. A univariate version of the statistic Q_{EEF} in Section 10.2.2 is discussed by Lloyd (1994) in the context of M-estimation.

Bootstrap likelihood was introduced by Davison, Hinkley and Worton (1992), who discuss its relationship to empirical likelihood, while a later paper (Davison, Hinkley and Worton, 1995) describes computational improvements.

Early work on the use of confidence distributions to define nonparametric likelihoods was done by Hall (1987), Boos and Monahan (1986), and Ogbonmwan and Wynn (1986). The use of confidence distributions in Section 10.4 rests in part on the similarity of confidence distributions to Bayesian posterior distributions. For related theory see Welch and Peers (1963), Stein (1985) and Berger and Bernardo (1992). Efron (1993) discusses the likelihood derived from *ABC* confidence limits, shows a strong connection with profile likelihood and related likelihoods, and gives several applications; see also Chapter 24 of Efron and Tibshirani (1993).

The Bayesian bootstrap was introduced by Rubin (1981), and subsequently used by Rubin and Schenker (1986) and Rubin (1987) for multiple imputation in missing data problems. Banks (1988) has described some variants of the Bayesian bootstrap, while Newton and Raftery (1994) describe a variant which they name the weighted likelihood bootstrap. A comprehensive theoretical discussion of weighted bootstraps is given in Barbe and Bertail (1995).

10.7 Problems

1 Consider empirical likelihood for a parameter $\theta = t(F)$ defined by an estimating equation $\int u(t;y)\,dF(y) = 0$, based on a random sample $y_1,\dots,y_n$.
(a) Use Lagrange multipliers to maximize $\sum \log p_j$ subject to the conditions $\sum p_j = 1$ and $\sum p_j u(t;y_j) = 0$, and hence show that the log empirical likelihood is given by (10.7) with $d = 1$. Verify that the empirical likelihood is maximized at the sample EDF, when $\theta = t(\hat{F})$.
(b) Suppose that $u(t;y) = y - t$ and $n = 2$, with $y_1 < y_2$. Show that η_θ can be written as $(\theta - \bar{y})/\{(\theta - y_1)(y_2 - \theta)\}$, and sketch it as a function of θ.
(Section 10.2.1)

2 Suppose that $x_1,\dots,x_n$ and $y_1,\dots,y_m$ are independent random samples from distributions with means μ and $\mu + \delta$. Obtain the empirical likelihood ratio statistic for δ.
(Section 10.2.1)

3 (a) In (10.5), suppose that $\theta = \bar{y} + n^{-1/2}\sigma\varepsilon$, where $\sigma^2 = \text{var}(y_j)$ and ε has an asymptotic standard normal distribution. Show that $\eta_\theta \doteq -n^{-1/2}\varepsilon/\sigma^2$, and deduce that near $\bar{y}$, $\ell_{EL}(\theta) \doteq -\frac{n}{2}(\bar{y} - \theta)^2/\sigma^2$.
(b) Now suppose that a single observation from F has log density $\ell(\theta) = \log f(y;\theta)$ and corresponding Fisher information $i(\theta) = \text{E}\{-\ddot{\ell}(\theta)\}$. Use the fact that the MLE $\hat{\theta}$ satisfies the equation $\dot{\ell}(\hat{\theta}) = 0$ to show that near $\hat{\theta}$ the parametric log likelihood is roughly $\ell(\theta) \doteq -\frac{n}{2} i(\theta)(\hat{\theta} - \theta)^2$
(c) By considering the double exponential density $\frac{1}{2}\exp(-|y - \theta|)$, $-\infty < y < \infty$, and an exponential family density with mean θ, $a(y)\exp\{yb(\theta) - c(\theta)\}$, show that it may or may not be true that $\ell_{EL}(\theta) \doteq \ell(\theta)$.
(Section 10.2.1; DiCiccio, Hall and Romano, 1989)

4 Let θ be a scalar parameter defined by an estimating equation $\int u(\theta;y)\,dF(y) = 0$. Suppose that we wish to make likelihood inference for θ based on a random sample $y_1,\dots,y_n$, using the empirical exponential family

$$\pi_j(\theta) = \Pr(Y = y_j) = \frac{e^{\xi_\theta u(\theta;y_j)}}{\sum_{k=1}^n e^{\xi_\theta u(\theta;y_k)}}, \quad j = 1,\dots,n,$$

where ξ_θ is determined by

$$\sum_{j=1}^n \pi_j(\theta) u(\theta;y_j) = 0. \qquad (10.19)$$

(a) Let $Z_1,\dots,Z_n$ be independent Poisson variables with means $\exp(\xi u_j)$, where $u_j \equiv u(\theta;y_j)$; we treat θ as fixed. Write down the likelihood equation for ξ, and show that when the observed values of the Z_j all equal zero, it is equivalent to (10.19). Hence outline how software that fits generalized linear models may be used to find ξ_θ.
(b) Show that the formulation in terms of Poisson variables suggests that the empirical exponential family likelihood ratio statistic is the Poisson deviance $W_{EEF}(\theta_0)$,

while the multinomial form gives $W'_{EEF}(\theta_0)$, where

$$\begin{aligned} W_{EEF}(\theta_0) &= 2\sum\{1-\exp(\xi_\theta u_j)\}, \\ W'_{EEF}(\theta_0) &= 2\left[n\log\left\{n^{-1}\sum e^{\xi_\theta u_j}\right\} - \xi_\theta \sum u_j\right]. \end{aligned}$$

(c) Plot the log likelihood functions corresponding to W_{EEF} and W'_{EEF} for the data in Example 10.1; take $u_j = y_j - \theta$. Perform a small simulation study to compare the behaviour of W_{EEF} and W'_{EEF} when the underlying data are samples of size 24 from the exponential distribution.
(Section 10.2.2)

5 Suppose that $a = (\sin\theta, \cos\theta)^T$ is the mean direction of a distribution on the unit circle, and consider setting a nonparametric confidence set for a based on a random sample of angles $\theta_1, \ldots, \theta_n$; set $y_j = (\sin\theta_j, \cos\theta_j)^T$.
(a) Show that $\hat{a}$ is determined by the equation $\sum y_j^T b = 0$, where $b = (\cos\theta, -\sin\theta)^T$. Hence explain how to construct confidence sets based on statistics from empirical likelihood and from empirical exponential families.
(b) Extend the argument to data taking values on the unit sphere, with mean direction $a = (\cos\theta\cos\phi, \cos\theta\sin\phi, \sin\theta)^T$.
(c) See Practical 10.2.
(Section 10.2.2; Fisher *et al.*, 1996)

6 Suppose that t has empirical influence values l_j, and set

$$p_j^*(\theta^0) = \frac{e^{\xi l_j}}{\sum_{i=1}^n e^{\xi l_i}}, \qquad (10.20)$$

where $\xi = v^{1/2}(\theta^0 - t)$ and $v = n^{-2}\sum l_j^2$.
(a) Show that $t(\hat{F}_\xi) \doteq \theta_0$, where $\hat{F}_\xi$ denotes the CDF corresponding to (10.20). Hence describe how to space out the values t_r^* in the first-level resampling for a bootstrap likelihood.
(b) Rather than use the tilted probabilities (10.12) to construct a bootstrap likelihood by simulation, suppose that we use those in (10.20). For a linear statistic, show that the cumulant-generating function of T^{**} in sampling from (10.20) is $\lambda t + n\{K(\xi + n^{-1}\lambda) - K(\xi)\}$, where $K(\xi) = \log(\sum e^{\xi l_j})$. Deduce that the saddlepoint approximation to $f_{T^{**}|T^*}(t \mid \theta^0)$ is proportional to $\exp\{-nK(\xi)\}$, where $\theta^0 = K'(\xi)$. Hence show that for the sample average, the log likelihood at $\theta^0 = \sum y_j e^{\xi y_j} / \sum e^{\xi y_j}$ is $n\{\xi t - \log(\sum e^{\xi y_j})\}$.
(c) Extend (b) to the situation where t is defined as the solution to a monotonic estimating equation.
(Section 10.3; Davison, Hinkley and Worton, 1992)

7 Consider the choice of h for the raw bootstrap likelihood values (10.11), when $w(\cdot)$ is the standard normal density. As is often roughly true, suppose that $T^* \sim N(t, v)$, and that conditional on $T^* = t^*$, $T^{**} \sim N(t^*, v)$.
(a) Show that the mean and variance of the product of $v^{1/2}$ with (10.11) are I_1 and $M^{-1}(I_2 - I_1^2)$, where

$$I_k = (2\pi)^{-k/2}\gamma^{-(k-1)}(\gamma^2 + k)^{-1/2}\exp\left\{-\frac{k\delta^2}{2(\gamma^2 + k)}\right\},$$

where $\gamma = hv^{-1/2}$ and $\delta = v^{-1/2}(t^* - t)$. Hence verify some of the values in the following table:

	$\delta = 0$			$\delta = 1$			$\delta = 2$		
γ	0	1	2	0	1	2	0	1	2
Density $\times 10^{-2}$	39.9	39.9	39.9	24.2	24.2	24.2	5.4	5.4	5.4
Bias $\times 10^{-2}$	−0.8	−2.9	−5.7	0	−0.1	−0.5	0.3	1.2	2.5
$M\times$ variance $\times 10^{-2}$	40.4	13.4	5.6	28.3	11.2	5.7	7.5	3.8	2.6

(b) If γ is small, show that the variance of (10.11) is roughly proportional to the square of its mean, and deduce that the variance is approximately constant on the log scale.
(c) Extend the calculations in (a) to (10.13).
(Section 10.3; Davison, Hinkley and Worton, 1992)

8 Let y represent data from a parametric model $f(y;\theta)$, and suppose that θ is estimated by $t(y)$. Assuming that simulation error may be ignored, under what circumstances would the bootstrap likelihood generated by parametric simulation from f equal the parametric likelihood? Illustrate your answer with the $N(\theta, 1)$ distribution, taking t to be (i) the sample average, (ii) the sample median.
(Section 10.3)

9 Suppose that we wish to construct an implied likelihood for a correlation coefficient θ based on its sample value T by treating $Z = \frac{1}{2}\log\{(1+T)/(1-T)\}$ as normal with mean $g(\theta) = \frac{1}{2}\log\{(1+\theta)/(1-\theta)\}$ and variance n^{-1}. Show that the implied likelihood and implied prior are proportional to

$$\exp\left[-\tfrac{n}{2}\{g(t)-g(\theta)\}^2\right], \qquad (1-\theta)^{-2}, \qquad |\theta| < 1.$$

Is the prior here proper?
(Section 10.4)

10 The Dirichlet density with parameters $(\xi_1, \ldots, \xi_n)$ is

$$\frac{\Gamma(\xi_1 + \cdots + \xi_n)}{\Gamma(\xi_1)\cdots\Gamma(\xi_n)} p_1^{\xi_1 - 1}\cdots p_n^{\xi_n - 1}, \qquad p_j > 0, \quad \sum p_j = 1, \quad \xi_j > 0.$$

Show that the P_j have joint moments

$$\mathrm{E}(P_j) = \frac{\xi_j}{s}, \quad \mathrm{cov}(P_j, P_k) = \frac{\xi_j(\delta_{jk}s - \xi_k)}{s^2(t+1)},$$

where $\delta_{jk} = 1$ if $j = k$ and zero otherwise, and $s = \xi_1 + \cdots + \xi_n$.
(a) Let $y_1, \ldots, y_n$ be a random sample, and consider bootstrapping its average. Show that under the Bayesian bootstrap with $a_j \equiv a$,

$$\mathrm{E}^\dagger(P_j^\dagger) = n^{-1}, \quad \mathrm{cov}^\dagger(P_j^\dagger, P_k^\dagger) = \frac{\delta_{jk} - 1/n}{n(2n + an + 1)}. \qquad (10.21)$$

Hence show that the posterior mean and variance of $\theta^\dagger = \sum y_j P_j^\dagger$ are $\bar{y}$ and $(2n + an + 1)^{-1}m_2$, where $m_2 = n^{-1}\sum(y_j - \bar{y})^2$.
(b) Now consider the average $\bar{Y}^\dagger$ of bootstrap samples generated as follows. We generate a distribution $F^\dagger \equiv (P_1^\dagger, \ldots, P_n^\dagger)$ on $y_1, \ldots, y_n$ under the Bayesian bootstrap,

and then make $Y_1^\dagger, \ldots, Y_n^\dagger$ by independent multinomial sampling from $F^\dagger$. Show that

$$\mathrm{E}^\dagger(\bar{Y}^\dagger) = \bar{y}, \quad \mathrm{var}(\bar{Y}^\dagger) = \frac{n(a+3)}{(2n+an+1)} \frac{m_2}{n}.$$

Are the properties of this as $n \to \infty$ and $a \to \infty$ what you would expect? How does this compare with samples generated by the usual nonparametric bootstrap? (Section 10.5)

10.8 Practicals

1 We compare the empirical likelihoods and 95% confidence intervals for the mean of the data in Table 3.1, (a) pooling the eight series:

```
attach(gravity)
grav.EL <- EL.profile(g,tmin=70,tmax=85,n.t=51)
plot(grav.EL[,1],exp(grav.EL[,2]),type="l",xlab="mu",
     ylab="empirical likelihood")
lik.CI(grav.EL,lim=-0.5*qchisq(0.95,1))
```

and (b) treating the series as arising from separate distributions with the same mean and plotting eight individual likelihoods:

```
gravs.EL <- EL.profile(g[series==1],n.t=21)
plot(gravs.EL[,1],exp(gravs.EL[,2]),type="n",xlab="mu",
     ylab="empirical likelihood",xlim=range(g))
lines(gravs.EL[,1],exp(gravs.EL[,2]),lty=2)
for (s in 2:8)
{ gravs.EL <- EL.profile(g[series==s],n.t=21)
  lines(gravs.EL[,1],exp(gravs.EL[,2]),lty=2) }
```

Now we combine the individual likelihoods into a single likelihood by multiplying them together; we renormalize so that the product has maximum one.

```
lims <- matrix(NA,8,2)
for (s in 1:8) { x <- g[series==s]; lims[s,] <- range(x) }
mu.min <- max(lims[,1]);  mu.max <- min(lims[,2])
gravs.EL <- EL.profile(g[series==1],
                  tmin=mu.min,tmax=mu.max,n.t=21)
gravs.EL.L  <- gravs.EL[,2]
gravs.EL.mu <- gravs.EL[,1]
for (s in 2:8)
gravs.EL.L <- gravs.EL.L + EL.profile(g[series==s],
                      tmin=mu.min,tmax=mu.max,n.t=21)[,2]
gravs.EL.L <- gravs.EL.L - max(gravs.EL.L)
lines(gravs.EL.mu,exp(gravs.EL.L),lwd=2)
lik.CI(cbind(gravs.EL.mu,gravs.EL.L),lim=-0.5*qchisq(0.95,1))
```

Compare the intervals with those in Example 3.2. Does the result for (b) suggest a limitation of multinomial likelihoods in general?
Compare the empirical likelihoods with the profile likelihood (10.1) and the adjusted profile likelihood (10.2), obtained when the series are treated as independent normal samples with different variances but the same mean.
(Section 10.2.1)

2 Dataframe `islay` contains 18 measurements (in degrees east of north) of palaeocurrent azimuths from the Jura Quartzite on the Scottish island of Islay. We aim to use multinomial-based likelihoods to set 95% confidence intervals for the mean direction $a(\theta) = (\sin\theta, \cos\theta)^T$ of the distribution underlying the data; the vector $b(\theta) = (\cos\theta, -\sin\theta)^T$ is orthogonal to a. Let $y_j = (\sin\theta_j, \cos\theta_j)^T$ denote the vectors corresponding to the observed angles $\theta_1, \ldots, \theta_n$. Then the mean direction $\hat{\theta}$ is the angle subtended at the origin by $\sum y_j / \| \sum y_j \|$.
For the original estimate, plots of the data, log likelihoods and confidence intervals:

```
attach(islay)
th <- ifelse(theta>180,theta-360,theta)
a.t <- function(th) c(sin(th*pi/180), cos(th*pi/180))
b.t <- function(th) c(cos(th*pi/180), -sin(th*pi/180))
y <- t(apply(matrix(theta, 18,1), 1, a.t))
thetahat <- function(y)
{ m <- apply(y,2,sum)
  m <- m/sqrt(m[1]^2+m[2]^2)
  180*atan(m[1]/m[2])/pi }
thetahat(y)
u.t <- function(y, th) crossprod(b.t(th), t(y))
islay.EL <- EL.profile(y, tmin=-100, tmax=120, n.t=40, u=u.t)
plot(islay.EL[,1],islay.EL[,2],type="l",xlab="theta",
     ylab="log empirical likelihood",ylim=c(-25,0))
points(th,rep(-25,18)); abline(h=-3.84/2,lty=2)
lik.CI(islay.EL,lim=-0.5*qchisq(0.95,1))
islay.EEF <- EEF.profile(y, tmin=-100, tmax=120, n.t=40, u=u.t)
lines(islay.EEF[,1],islay.EEF[,2],lty=3)
lik.CI(islay.EEF,lim=-0.5*qchisq(0.95,1))
```

Discuss the shapes of the log likelihoods.
To obtain 0.95 quantiles of the bootstrap distributions of W_{EL} and W_{EEF}:

```
islay.fun <- function(y, i, angle)
{   u <- as.vector(u.t(y[i,], angle))
    z <- rep(0,length(u))
    EEF.fit <- glm(z~u-1,poisson)
    W.EEF <- 2*sum(1-fitted(EEF.fit))
    EL.loglik <- function(lambda) - sum(log(1 + lambda * u))
    EL.score <- function(lambda) - sum(u/(1 + lambda * u))
    assign("u",u,frame=1)
    EL.out <- nlmin(EL.loglik,0.001)
    W.EL <- -2*EL.loglik(EL.out$x)
    c(thetahat(y[i,]), W.EL, W.EEF, EL.out$converged) }
islay.boot <- boot(y,islay.fun,R=999,angle=thetahat(y))
islay.boot$R <- sum(islay.boot$t[,4])
islay.boot$t <- islay.boot$t[islay.boot$t[,4]==1,]
apply(islay.boot$t[,2:3],2,quantile,0.95)
```

How do the bootstrap-calibrated confidence intervals compare with those based on the χ_1^2 distribution, and with the basic bootstrap intervals using the $\hat{\theta}^*$?
(Sections 10.2.1, 10.2.2; Hand *et al.*, 1994, p. 198)

3 We compare posterior densities for the mean of the air-conditioning data using (a) the Bayesian bootstrap with $a_j \equiv -1$:

```
air1 <- data.frame(hours=aircondit$hours,G=1)
air.bayes.gen <- function(d, a)
{ out <- d
  out$G <- rgamma(nrow(d),shape=a+2)
  out }
air.bayes.fun <- function(d) sum(d$hours*d$G)/sum(d$G)
air.bayesian <- boot(air1, air.bayes.fun, R=999, sim="parametric",
                     ran.gen=air.bayes.gen,mle=-1)
plot(density(air.bayesian$t,n=100,width=25),type="l",
     xlab="theta",ylab="density",ylim=c(0,0.02))
```

and (b) an exponential model with mean θ for the data, with prior according to which θ^{-1} has a gamma distribution with index κ and scale λ^{-1}:

```
kappa <- 0; lambda <- 0
kappa.post <- kappa + length(air1$hours)
lambda.post <- lambda + sum(air1$hours)
theta <- 30:300
lines(theta,
      lambda.post/theta^2*dgamma(lambda.post/theta,kappa.post),
      lty=2)
```

Repeat this with different values of a in the Bayesian bootstrap and κ, λ in the parametric case, and discuss your results.
(Section 10.5)