

# Hybrids and connections

This chapter considers hybrid methods in which empirical likelihood is combined with other methods. There are a number of problems where a parametric likelihood is known or trusted for part, but not all, of the problem. In those cases, hybrid methods fill in the gaps with empirical likelihood. Similarly, an empirical likelihood can be combined with a Bayesian prior distribution, the bootstrap, and various jackknives. Bootstrap calibration of empirical likelihood is discussed elsewhere (Chapters 3.3 and 5.6.), as is a hybrid between empirical likelihood and permutation tests (Chapter 10.3).

There are also deep connections between empirical likelihood and other non-parametric methods of inference. This is not surprising. Loosely speaking, two methods that are asymptotically correct to some order might be expected to agree to at least that order. Connections to bootstraps, jackknives, and sieves are presented.

## 9.1 Product of parametric and empirical likelihoods

Consider a setting with two samples  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_m$ . Suppose that all  $n + m$  observations are independent, and that we have a trusted parametric model in which  $Y_i \sim g(y_i; \theta)$ , but no such model is available for  $X_i$ . For example,  $Y_i$  may be from an instrument shown by experience to be normally distributed while the  $X_i$  may be from a newer kind of equipment. Similarly, the inter-arrival times  $Y_i$  for a queue may be known to have an exponential distribution, but the service times  $X_i$  may not belong to a known parametric family.

A natural approach is to form a likelihood that is nonparametric in the distribution  $F$  of the  $X_i$  but is parametric in the distribution  $G$  of the  $Y_i$ . That is,

$$L(F, \theta) = \prod_{i=1}^n F(\{X_i\}) \prod_{j=1}^m g(Y_j; \theta),$$

with likelihood ratio function

$$R(F, \theta) = \prod_{i=1}^n n w_i \prod_{j=1}^m \frac{g(Y_j; \theta)}{g(Y_j; \hat{\theta})},$$

where as usual  $F = \sum_{i=1}^n w_i \delta_{X_i}$  for weights  $w_i \geq 0$ ,  $\sum_{i=1}^n w_i = 1$ , and  $\hat{\theta}$  is the parametric MLE of  $\theta$  computed from the  $Y_i$  data.

Suppose that we are interested in a parameter  $\phi$  defined through estimating equations

$$E(h(X, Y, \phi)) = \iint h(x, y, \phi) dG_\theta(y) dF(x) = 0.$$

If  $h(X, Y, \phi)$  only involves  $X$ , or only involves  $Y$ , then ordinary empirical or parametric likelihood, respectively, is available. When both distributions are involved, define

$$\mathcal{R}(\phi) = \max_{F, \theta} R(F, \theta)$$

subject to

$$\sum_{i=1}^n w_i \int h(X_i, y, \phi) dG_\theta(y) = 0.$$

Under mild conditions (see Chapter 9.11), an asymptotic  $\chi^2$  calibration with degrees of freedom equal to the dimension of  $h$  is appropriate.

## 9.2 Parametric conditional likelihood

Consider independent  $(X_i, Y_i)$  pairs, for  $i = 1, \dots, n$ , with  $X_i \in \mathbb{R}^p$  and  $Y_i \in \mathbb{R}^q$ . Now suppose that we have a parametric density or mass function  $g(y|x; \theta)$  for the conditional distribution of  $Y$  given  $X$  but no parametric likelihood for the marginal distribution of  $X$ . Then the hybrid likelihood is

$$L(F, \theta) = \prod_{i=1}^n F(\{X_i\}) g(Y_i | X_i; \theta). \quad (9.1)$$

More generally, with  $m - n \geq 0$  further observations with  $X_i$  but not  $Y_i$  measured, the likelihood is

$$L(F, \theta) = \prod_{i=1}^m F(\{X_i\}) \prod_{i=1}^n g(Y_i | X_i; \theta). \quad (9.2)$$

**Exercise 9.1** considers observations with  $Y_i$  but not  $X_i$  measured.

It is natural to suppose that there is no known relationship between  $\theta$  and  $F$ . If the statistic of interest only involves the  $X$  distribution, or only involves the distribution of  $Y$  given  $X$ , then we may use a marginal empirical likelihood of  $X$  or a parametric conditional likelihood of  $Y$  given  $X$ , respectively. But if the statistic of interest involves both distributions, then there is something to be gained by using (9.1) or (9.2).

Suppose, for example, that given  $X = x$ , the response  $Y$  has the  $N(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2)$  distribution. We might want to know the value  $x_0 = -\beta_1 / (2\beta_2)$ . This  $x_0$  represents a minimum of  $E(Y | X = x)$  if  $\beta_2 > 0$ . It is a maximum if  $\beta_2 < 0$ . The average  $X$  value deviates from this optimum by an amount  $\Delta = E(X) + \beta_1 / (2\beta_2)$ . This  $\Delta$  represents the amount of the correction that needs to

be applied to the average  $X$  value to reach the optimum. It depends on both the parametric and nonparametric parts of the data description.

An asymptotic  $\chi^2$  distribution has been shown for this hybrid likelihood ratio (see Chapter 9.11) for a problem incorporating some side information. That theorem has a scalar parameter of interest  $\mu$  defined through a smooth estimating equation  $E(h(X, \theta, \mu)) = 0$ . There is no reason to suppose that the scalar setting is special here, and so it is reasonable to believe that the hybrid likelihood can be used to generate tests and confidence regions quite generally.

It is interesting to consider the reverse situation where the marginal likelihood is parametric and the conditional likelihood is empirical. Let  $f(x; \theta)$  be the parametric density or mass function of  $X$ . Let  $G_{0,x}(Y)$  be the conditional distribution of  $Y$  given  $X = x$ , and let  $G_x(Y)$  be a candidate. Assuming no ties, the likelihood is

$$L(\theta, G_{X_1}, \dots, G_{X_n}) = \prod_{i=1}^n f(X_i; \theta) G_{X_i}(Y_i).$$

Maximizing this likelihood is degenerate if there are no ties among the  $X_i$ . The NPMLE for  $\theta$  matches the parametric one, but every  $G_{X_i}$  is a point mass at the corresponding value of  $Y_i$ . If there are a small number of distinct  $X_i$  with a large number of occurrences each, then NPMLE's for them are not degenerate. Failing that, a model linking  $G_x$  for different  $x$  values is required. If  $G_x(y) = G(y - h(x))$ , for a known function  $h(x)$ , or more generally  $G_x(y) = G(y - h(x, \gamma))$  for a parameter vector  $\gamma$ , then the likelihood

$$L(\theta, G, \gamma) = \prod_{i=1}^n f(X_i; \theta) G(Y_i - h(X_i, \gamma))$$

is not necessarily degenerate.

### 9.3 Parametric models for data ranges

Another case where parametric and empirical likelihoods mix is where the data  $X_i$  are thought to follow a parametric model  $f(x; \theta)$ , over a subset of their range. For example, the random variable  $X_i$  may be thought to be normally distributed over some central values  $[-M, M]$  but not necessarily in the tails. Or the  $X_i$  may be positive random variables known to have an exponential tail,  $f(x; \theta) \propto \exp(-\theta x)$  for  $x > x_0$ . As in the previous sections, we might get a sample that is partly parametric and partly nonparametric, but in this case it is not known which likelihood applies to an observation until that observation becomes available.

Suppose that  $X \sim f(x; \theta)$  for  $x \in P_0$ , but not necessarily for  $x \notin P_0$ . Introduce the shorthand  $P_0(x)$  for  $1_{x \in P_0}$ . The appropriate likelihood hybrid is then

$$\prod_{i=1}^n [f(X_i; \theta)]^{P_0(X_i)} w_i^{1-P_0(X_i)},$$

where  $w_i \geq 0$ ,  $w_i = 0$  when  $X_i \in P_0$ , taking  $0^0 = 1$ , and we impose the constraint

$$\int_{P_0} dF(x; \theta) + \sum_{i=1}^n w_i (1 - P_0(x_i)) = 1.$$

Again under mild conditions including smoothness in  $\theta$  of the parametric distribution (see Chapter 9.11), the combined likelihood has the asymptotic  $\chi^2$  distribution that one would expect.

## 9.4 Empirical likelihood and Bayes

Let  $\theta$  be a parameter with estimating equation  $E(m(X, \theta)) = 0$ . Suppose that we are willing to specify a prior density  $\pi(\theta)$  for  $\theta$ , but that we have no parametric family for the distribution  $F$  of  $X$ . The opposite situation, where a parametric model is given for  $\theta$  but we are reluctant to specify a prior, is very commonly approached with flat non-informative prior distributions.

We will suppose that  $\int_{\theta \in \Theta} \pi(\theta) d\theta = 1$ . Then a natural procedure is to take the posterior distribution of  $\theta$  to be

$$\mathcal{L}(\theta | X_1, \dots, X_n) = \frac{\pi(\theta) \mathcal{R}(\theta)}{\int_{\theta \in \Theta} \pi(\theta) \mathcal{R}(\theta) d\theta}, \quad (9.3)$$

where  $\mathcal{R}(\theta)$  is the profile empirical likelihood ratio function for  $\theta$ . There is as yet little known about how well this proposal works. Some theory and simulations showing the asymptotic accuracy of posterior probability statements computed from  $\mathcal{L}(\theta | X_1, \dots, X_n)$ , when  $\theta$  is a univariate mean, are described in Chapter 9.11.

The process is simply to multiply the empirical likelihood by a prior distribution, and then renormalize to a proper density function. This appears to avoid putting a prior on the space of all distributions  $F$ , or even on the whole simplex of multinomial weights. The rationale behind the process is as follows: Empirical likelihood is approximately using a least favorable family for  $\theta$ . This is a parametric family of the same dimension as  $\theta$ . The prior distribution on  $\theta$  induces a prior distribution on this same family. Multiplying the prior on the family by the likelihood on the family yields a posterior density on the family.

If nuisance parameters  $\nu$  are defined jointly with  $\theta$  through  $E(m(X, \theta, \nu)) = 0$ , then we place a prior  $\pi(\theta, \nu)$  on  $\theta$  and  $\nu$ . The posterior distribution on  $\theta$  and  $\nu$  is then proportional to  $\pi(\theta, \nu) \mathcal{R}(\theta, \nu)$ , and the posterior distribution for  $\theta$  is obtained by integrating out  $\nu$ .

## 9.5 Bayesian bootstrap

To describe the Bayesian bootstrap, suppose at first that the distribution  $F_0$  attaches probability 1 to the known finite set  $\{z_1, \dots, z_k\}$ . This constraint will be lifted later. The  $z_j$  may be in  $\mathbb{R}^d$  or even in more general spaces. Then there is

a finite dimensional parametric space of candidate distributions  $F$  defined by the parameter vector  $\omega = (\omega_1, \dots, \omega_k)' \in \mathbb{S}_{k-1}$  with  $\omega_j = F(\{z_j\})$ . The unit probability simplex  $\mathbb{S}_{k-1}$  is defined in equation (2.6), replacing  $n$  there by  $k$ .

The Bayesian bootstrap places a Dirichlet prior on  $\theta$ . The Dirichlet prior is proportional to  $1_{\omega \in \mathbb{S}_{k-1}} \prod_{j=1}^k \omega_j^{m_j}$ . If the sample contains  $n_j$  observations equal to  $z_j$ , then the posterior distribution is also a Dirichlet, and is proportional to  $1_{\omega \in \mathbb{S}_{k-1}} \prod_{j=1}^k \omega_j^{m_j+n_j}$ . The choice  $m_j = -1$  is particularly convenient. If there are any  $z_j$  with  $n_j = 0$ , then the posterior distribution is improper, having an infinite integral. That posterior can be interpreted as placing probability 1 on  $\theta_j = 0$  for every  $j$  with  $n_j = 0$ . Then the posterior distribution is proportional to

$$1_{\omega \in \mathbb{S}_{k-1}} \prod_{j:n_j > 0} \omega_j^{n_j-1} \prod_{j:n_j=0} 1_{\omega_j=0}. \quad (9.4)$$

The unobserved  $z_j$  for which  $n_j = 0$  do not appear in the posterior distribution, and this lifts the constraint that we have to know what they are.

The Bayesian bootstrap samples from the posterior distribution implied by (9.4). To generate a sample from the posterior distribution, draw  $n$  independent  $U(0, 1)$  random variables  $U_i$ , transform them into exponential random variables  $Y_i = -\log(U_i)$ , and then take  $w_i = Y_i / \sum_{j=1}^n Y_j$ . The sampled value of  $\omega_j$  is then  $\sum_{i: X_i=z_j} w_i$ . For a statistic  $T(F)$ , the resampled value is  $T(\sum_{i=1}^n w_i \delta_{X_i})$ . The Bayesian bootstrap sample consists of  $B$  independently sampled values of  $T$ . The posterior probability of a set  $C$  is estimated by the fraction of the  $B$  resampled  $T$  values that happen to be in  $C$ , and a posterior moment is simply the average over  $B$  sampled values of the corresponding power of  $T$ .

The empirical likelihood is proportional to  $\prod_{j=1}^k \omega_j^{n_j}$ . Apart from the way unobserved  $z_j$  are handled, the empirical likelihood is obtained as a posterior distribution for the non-informative Dirichlet prior having  $m_j = 0$ . This is a non-parametric analogue of the familiar fact that the posterior is proportional to the likelihood, when a non-informative prior is used.

## 9.6 Least favorable families and nonparametric tilting

Empirical likelihood works with an  $n$ -dimensional family of distributions supported on the sample points. For data  $X_i \in \mathbb{R}^d$ , we maximize the empirical likelihood subject to a constraint like  $\sum_i w_i X_i = \mu$ . As  $\mu$  varies through a  $d$ -dimensional space, a  $d$ -dimensional subfamily of the multinomial distributions arise as constrained maxima. This family may be indexed by  $\mu$ , or by the Lagrange multiplier  $\lambda$ .

For a statistic  $\theta = T(F) \in \mathbb{R}^p$ , there is usually a reduction to a  $p$ -dimensional family of multinomial distributions. Similar  $p$ -dimensional subfamilies may also be defined through other discrepancies such as empirical entropy or the Euclidean likelihood.

The nonparametric tilting bootstrap draws samples from members of one of

these lower dimensional subfamilies of multinomial distributions. For a scalar parameter  $\theta$  such as the univariate mean, there is a univariate family of distributions. Denote the generic member of this family by  $F_\theta$  and let  $w_i(\theta) = F_\theta\{X_i\}$ . Bootstrap samples can be drawn from these multinomial distributions by taking  $X_i^{*b} = X_{J(i,b)}$ , where for  $i = 1, \dots, n$  and  $b = 1, \dots, b$  the indices  $J(i, b)$  are independent with  $\Pr(J(i, b) = k) = w_k(\theta)$ . Then  $\hat{\theta}^{*b}$  is the value of  $\hat{\theta}$  on the data  $X_1^{*b}, \dots, X_n^{*b}$ . The values

$$\begin{aligned}\theta_L &= \min\{\theta \mid \Pr(\hat{\theta}^* \geq \hat{\theta}; F_\theta) \geq \alpha/2\}, \\ \theta_U &= \max\{\theta \mid \Pr(\hat{\theta}^* \leq \hat{\theta}; F_\theta) \geq \alpha/2\},\end{aligned}$$

are the lower and upper limits, respectively, of the nonparametric tilting approximate  $100(1 - \alpha)\%$  confidence interval.

It can be very laborious to sample from many members of a parametric family, searching for the desired endpoints. Importance sampling can be used to reweight data from one distribution  $F_\theta$  to obtain unbiased expectations under another distribution  $F_{\theta'}$ . The nonparametric tilting bootstrap samples from  $F_{\hat{\theta}}$  with every  $w_i = 1/n$ , and so the importance sampling weight is  $\prod_{i=1}^n n w_{J(i,b)}(\theta)$  in bootstrap sample  $b$ . A further advantage of importance sampling is that the simulations for different values of  $\theta$  are coupled, which makes it more likely that the search for endpoints can be done by looking for the point at which a monotone function is zero. Using the Kullback-Liebler family gives an especially convenient exponential tilting form for the importance sampling weight factors.

A least favorable parametric family is sometimes described as one in which the estimation problem is hardest, sometimes as one in which the estimation problem is as hard as in a parametric problem. Usually difficulty is measured through a discrepancy measure. So if a statistic is defined through  $T(F)$  and  $\theta_0$  is the true value of  $T$ , then family might consist of distributions  $F_\theta$  for which  $T(F_\theta) = \theta$ , and subject to this, a distance measure  $D(F_{\theta_0}, F_\theta)$  is minimized.

Empirical likelihood, nonparametric tilting, and some other methods described in Chapter 9.11 all employ least favorable families. These methods use random, sample based families. The value of a least favorable family is that it has not made the statistical problem artificially easy.

It is possible to construct some parametric families in which inference is outlandishly hard, although still not least favorable. The  $N(0, 1)$  distribution is a member of the family

$$f(x; \theta) = (1 - \epsilon)N(0, 1) + \epsilon N(\theta/\epsilon, 1). \quad (9.5)$$

Here  $f(x; \theta)$  has mean  $\theta$ . For  $\epsilon = 10^{-100}$  and any reasonable sample size, it would be very hard to estimate  $\theta$ . Using multinomial families on the data rules out unreasonably hard cases like (9.5).

## 9.7 Bootstrap likelihood

Suppose that we compute an estimate  $\hat{T} = T(\hat{F})$  of  $\theta = T(F_0)$ , using an IID sample  $X_1, \dots, X_n \in \mathbb{R}^d$ . In a parametric family indexed by  $\theta$ , the probability density function  $f(\hat{T}; \theta)$  could be interpreted as a partial likelihood for  $\theta$ . The qualifier “partial” reflects that  $f$  does not give the joint density of  $X_1, \dots, X_i$ , but just that of the function  $\hat{T}$  computed from them. In Chapter 13.3 a similar density is called a pseudo-likelihood.

The bootstrap likelihood uses two levels of resampling, some density estimation, and some regression smoothing to estimate  $f(\hat{T}; \theta)$  from the data. We will suppose that  $T(F) \in \mathbb{R}$ . For  $r = 1, \dots, R$ , let  $X_1^{*r}, \dots, X_n^{*r}$  be a bootstrap sample of the data, with corresponding  $T$  value  $\hat{T}^{*r}$ . Then for  $r = 1, \dots, R$  and  $s = 1, \dots, S$ , let  $X_1^{*rs}, \dots, X_n^{*rs}$  be a bootstrap sample from  $X_1^{*r}, \dots, X_n^{*r}$  with corresponding  $T$  value  $\hat{T}^{*rs}$ . A preliminary bootstrap likelihood at  $\hat{T}^{*r}$  is then estimated by a kernel density estimate

$$L(\hat{T}^{*r}) = \hat{f}(\hat{T}; \hat{T}^{*r}) = \frac{1}{Sh_2} \sum_{s=1}^S K_2 \left( \frac{\hat{T}^{*rs} - \hat{T}}{h_2} \right),$$

where  $K_2$  is a kernel function (Chapter 5) and  $h_2$  is a bandwidth.

It is possible to have  $L(\hat{T}^{*r}) \neq L(\hat{T}^{*r'})$  even when  $\hat{T}^{*r} = \hat{T}^{*r'}$ . For this reason, and to interpolate, the bootstrap likelihood is defined through further smoothing, such as

$$L_B(\theta) = \hat{f}(\hat{T}; \theta) = \frac{\sum_{r=1}^R K_1 \left( \frac{\hat{T}^{*r} - \theta}{h_1} \right) L(\hat{T}^{*r})}{\sum_{r=1}^R K_1 \left( \frac{\hat{T}^{*r} - \theta}{h_1} \right)},$$

for the kernel  $K_1$  and bandwidth  $h_1$ , or through some other scatterplot smoother.

The bootstrap likelihood has been shown to match the empirical likelihood, but only to first order. Much of the research on bootstrap likelihood aims at reducing the computational burden. See Chapter 9.11.

## 9.8 Bootstrapping from an NPMLE

The usual form of the bootstrap resamples from the empirical distribution  $F_n$ . In IID sampling the empirical distribution is the NPMLE. In settings with side information,  $F_n$  is not the NPMLE, and an attractive alternative is to use empirical likelihood to construct the NPMLE  $\hat{F}$ , and then resample from  $\hat{F}$ . When the side information is specified by  $E(m(X, \theta, \nu)) = 0$  then the NPMLE of Chapter 3.10 is  $\hat{F} = \sum_{i=1}^n w_i \delta_{X_i}$ , where  $w_i \geq 0$ ,  $\sum_{i=1}^n w_i = 1$ ,  $\sum_{i=1}^n w_i m(X_i, \theta, \nu) = 0$ , and  $\prod_{i=1}^n n w_i$  is maximized subject to these constraints. Similarly, when the data were obtained by biased sampling, then bootstrapping from the NPMLE becomes attractive.

Resampling from  $\hat{F}$  is straightforward. Let  $C_i = \sum_{j=1}^i w_j$ , with the understanding that  $C_0 = 0$ . To generate a bootstrap sample draw  $nB$  independent

$U(0, 1)$  random variables  $U_i^b$ , for  $1 \leq i \leq n$  and  $1 \leq b \leq B$ . Turn these into resampled observations where  $X_i^b = X_j$  whenever  $C_{j-1} < U_i^b \leq C_j$ . This produces  $B$  bootstrap data sets  $(X_1^b, \dots, X_n^b)$ , for  $b = 1, \dots, B$ .

The Euclidean likelihood can also be used to define an NPMLE from which to resample. But sampling from the Euclidean likelihood NPMLE is hard to define in cases where some  $w_i < 0$ .

## 9.9 Jackknives

The jackknife is a leave-one-out method of forming confidence regions for a statistical quantity  $\theta$ . Suppose that the true value is  $\theta_0 = T(F_0) \in \mathbb{R}^p$  and the sample value is  $T(\hat{F})$ . For a candidate distribution  $F = \sum_{i=1}^n w_i \delta_{X_i}$  where  $w_i \geq 0$  and  $\sum_{i=1}^n w_i = 1$ , define  $T(w_1, \dots, w_n) = \theta(F)$ . The NPMLE is  $\hat{\theta} = T(1/n, \dots, 1/n)$ . Define  $T_{-i} = \theta((1 + 1/n)\hat{F} - (1/n)\delta_{X_i})$ , the value of  $\theta$  on a hypothetical sample of the  $n - 1$  observations other than  $X_i$ . Now let

$$T_{-\bullet} = \frac{1}{n} \sum_{i=1}^n T_{-i}, \quad \text{and}$$

$$S = \sum_{i=1}^n (T_{-i} - T_{-\bullet})(T_{-i} - T_{-\bullet})'.$$

The value  $(n - 1)(T_{-\bullet} - T(\hat{F}))$  is often used as an estimate of the bias in  $T(\hat{F})$ , and the corresponding bias-corrected estimate of  $T(F)$  is  $nT(\hat{F}) - (n - 1)T_{-\bullet}$ . Also  $S$  or  $(n - 1)S/n$  can often be used to estimate the variance of  $T(\hat{F})$  or of  $T_{-\bullet}$ . For  $p = 1$ , a simple 95% confidence interval for  $\theta_0$  may then be calculated as  $T(\hat{F}) \pm 1.96\sqrt{S}$ .

The infinitesimal jackknife is based on the linear approximation

$$T(w_1, \dots, w_n) = \hat{\theta} + \sum_{i=1}^n w_i T_i(\hat{F}),$$

where

$$T_i(F) = \lim_{\varepsilon \rightarrow 0} \frac{T((1 - \varepsilon)F + \varepsilon\delta_{X_i}) - T(F)}{\varepsilon}. \quad (9.6)$$

In the infinitesimal jackknife, a variance estimate for  $\hat{\theta} - \theta_0$  can be constructed as

$$\frac{1}{n^2} \sum_{i=1}^n T_i(\hat{F})T_i(\hat{F})'.$$

The coefficient  $n^{-2}$  can be replaced by  $1/(n(n - 1))$  in order to get an unbiased variance estimate for the variance of linear statistics of the form  $T(w_1, \dots, w_n) = \sum_{i=1}^n w_i Q(X_i)$ . One of the algorithms for maximizing empirical likelihood (see

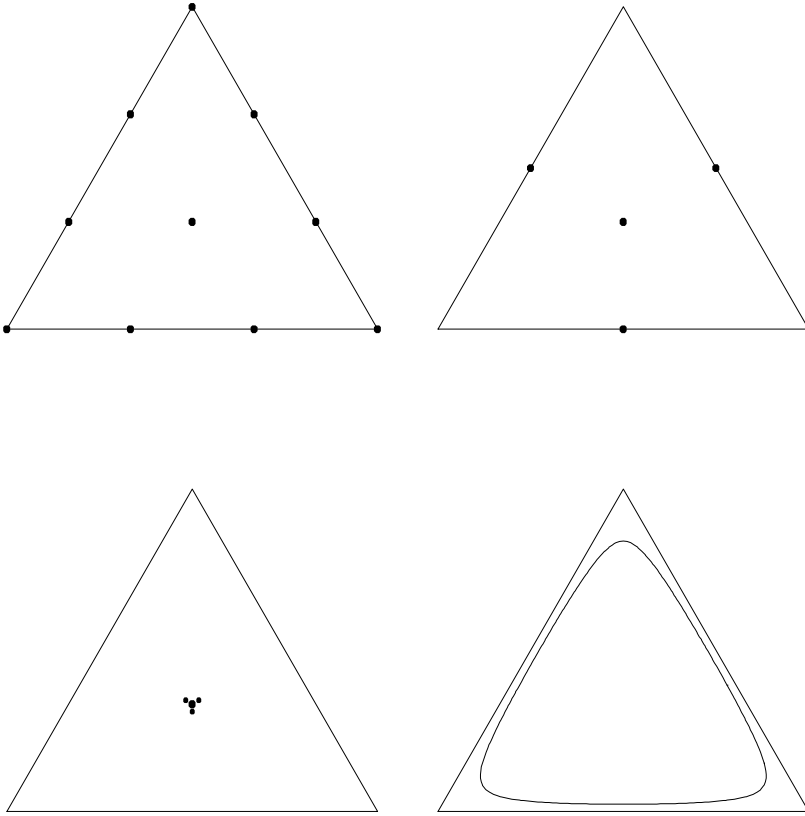


Figure 9.1 The upper left plot shows the reweightings of the data used by the bootstrap, for  $n = 3$ . The upper right and lower left show the jackknife and infinitesimal jackknife reweightings, respectively. The lower right plot depicts a region used by empirical likelihood.

Chapter 12.6) is an empirical likelihood test of whether the  $T_i(\hat{F})$  have mean  $\theta_0 - \hat{\theta}$ . They have sample mean 0.

Figure 9.1 compares the points in the simplex used in nonparametric inferences, for  $n = 3$ . When the infinitesimal jackknife fails it is often because it only uses reweightings that are a negligible distance from the center  $(1/n, \dots, 1/n)$ . It works with a linear approximation to  $T$ , and this may be inadequate if  $T$  is nonlinear enough. The jackknife has similar difficulties because the points it uses are only  $O(1/n)$  away from the center. Even when they do not fail, approxima-

tions based on expansions around the MLE can work badly in some settings. Linearization inference in nonlinear least squares, Greenwood's formula in survival analysis, and Wald tests are often found to be inferior to methods that are not so strongly localized around the MLE.

The bootstrap and empirical likelihood achieve greater generality than the jackknives in part because they consider reweightings at a distance (or average distance) of  $O(n^{-1/2})$  from the center. The difference is illustrated by the median, for which the jackknife and infinitesimal jackknife do not provide consistent variance estimates.

## 9.10 Sieves

In some nonparametric problems, there is an infinite dimensional family of estimators and an MLE does not exist. For example, consider the problem of estimating the density  $f$  from an IID sample of random variables  $X_i \in [0, 1]$ . The density  $f$  is a nonnegative function integrating to 1 over  $[0, 1]$ . The natural likelihood to use is  $L(f) = \prod_{i=1}^n f(X_i)$ , but  $L(f)$  is unbounded over the family of densities, thus there is no NPMLE.

A sieve approach to this problem is to consider a regularized set of densities, such as densities that are piecewise constant on  $[0, 1]$  with the allowed discontinuities at knots  $t_j = j/m$  for  $j = 1, \dots, m - 1$ . Given a value of  $m > 1$ , the NPMLE is a histogram estimator. The NPMLE might not be unique if some  $X_i$  equals some  $t_j$ , but a unique choice can be forced by taking  $f$  to be continuous from the left. The sieve approach to the histogram estimator lets  $m \rightarrow \infty$  at a suitable rate, as  $n \rightarrow \infty$ . Sieves are more general than the histogram estimator, and the controlling parameter  $m$  is not always integer valued. For example, a sieve could be constructed by taking all continuous densities with  $\int_0^1 f(x)^2 dx < m$ .

Empirical likelihood can be thought of as a sieve method. The family of candidate distributions is usually all those that reweight the sample. It may also include distributions that put some weight on a number of non-observed values. Two features of empirical likelihood distinguish it from sieve methods. First, the family of candidate distributions is random, as it depends on the sample. Second, the emphasis is shifted from maximum likelihood to likelihood ratios.

While sieves originated to handle an infinite dimensional parameter set, they may also be applied in problems with an infinite number of estimating equation constraints. Some examples of such problems appear in Chapter 10.

The sieved empirical likelihood (SEL) has been developed for some conditional moment restriction problems. Suppose that  $X$  and  $Z$  are jointly distributed random variables and that the estimating equation

$$E(m(Z, \theta) \mid X = x) = 0 \tag{9.7}$$

holds for all  $x$ . Then, of course,  $E(m(Z, \theta)) = 0$  holds unconditionally, too, but (9.7) is a more stringent condition. For instance, in a regression model we might have  $Z = (X', Y)'$ , and  $m(Z, \theta) = Y - g(X, \theta)$ . Let  $u(X)$  be a vector of func-

tions of  $X$ . Then  $E(u(X)(Y - g(X, \theta))) = 0$  and if  $u$  has enough component functions in it, we get an overdetermined set of unconditional estimating equations for  $\theta$ .

A lower bound is known for the variance of an asymptotically unbiased estimator  $\hat{\theta}$  from  $n$  IID observations:

$$n\text{Var}(\hat{\theta}) \geq V_0 \equiv \left( E(D(X)' \Omega(X)^{-1} D(X)) \right)^{-1}$$

where  $D(x) = E(\partial m(Z, \theta) / \partial \theta \mid X = x)$  and  $\Omega(x) = E(m(Z, \theta)m(Z, \theta)' \mid X = x)$ . Moreover, the solution to estimating equations  $E(u(X)m(Z, \theta)) = 0$  attains this asymptotic variance bound for the (usually unknown) vector  $u_0(x) = D(x)' \Omega(x)^{-1}$ . The functions  $u(X)$  are called instrumental variables and  $u_0$  are the optimal instruments.

A sieved empirical likelihood approach imposes (9.7) at the  $n$  sample values  $x_i$ , using some smoothing. Let  $K_{ij} = K((x_i - x_j)/h)$ , for a kernel function  $K$  and bandwidth  $h$ , and let  $K_{i\bullet} = \sum_{j=1}^n K_{ij}$ . Define the conditional empirical log likelihood function  $L_i(F) = \sum_{j=1}^n K_{ij} \log(w_{ij})$  for  $F(\{Z_j\} \mid X_i) = w_{ij} \geq 0$  and  $\sum_{j=1}^n w_{ij} = 1$ . Maximizing  $L_i$  subject to  $\sum_{j=1}^n w_{ij} m(Z_j, \theta) = 0$  is the same weighted empirical likelihood problem we solved in Chapter 8.7 for finite population sampling designs, with  $K_{i\bullet}$  playing the role of  $D$  there. Let the solution be  $\hat{F}_i$  with weights  $\hat{w}_{ij}(\theta)$ , and now define  $\ell(\theta) = \sum_{i=1}^n \log \hat{w}_{ii}(\theta)$ , and let  $\hat{\theta}$  maximize  $\ell(\theta)$ .

**Theorem 9.1** *Let  $(X_i, Z_i)$  be IID. Under some mild conditions that uniquely identify the true value  $\theta_0$ , conditions that impose smoothness on  $m$ , and conditions on  $K$  and  $h$ ,  $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, V_0)$ . Further, if  $\tilde{\theta}$  maximizes  $\ell(\theta)$  subject to  $s$  conditions  $C(\theta) = 0$  where  $\partial C / \partial \theta$  has rank  $s$ , then  $-2(\ell(\tilde{\theta}) - \ell(\hat{\theta})) \rightarrow \chi^2_{(s)}$  as  $n \rightarrow \infty$ .*

*Proof.* Fan & Gijbels (1999) and Kitamura, Tripathi & Ahn (2000).  $\square$

More general random sieves have been proposed, particularly for certain regression problems. In linear regression models relating  $Y_i$  to a predictor  $X_i$  and a parameter  $\beta$ , there is a residual  $e_i = Y_i - X_i' \beta$ . In certain models, however, there is no unique  $e_i$  that can be computed from a given  $X_i$ ,  $Y_i$  and  $\beta$ . A simple example without any  $X_i$ , is the following mixed effects model

$$Y_{ij} = \theta_j + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, k$$

for scalar parameters  $\theta_j$  and independent errors  $\alpha_i, \epsilon_{ij}$ , with mean 0,  $\text{Var}(\alpha_i) = \sigma_1^2$ , and  $\text{Var}(\epsilon_{ij}) = \sigma_2^2$ . Given  $\theta = (\theta_1, \dots, \theta_k)$  and  $Y_i = (Y_{i1}, \dots, Y_{ik})$  we cannot compute  $e_i = (\alpha_i, \epsilon_{i1}, \dots, \epsilon_{ik})$  though we do know that  $\alpha_i + \epsilon_{ij} = Y_{ij} - \theta_j$ .

For a random sieve model, we can identify a set  $B_i(\theta)$  known to contain  $e_i$ . For the random effect example

$$B_i(\theta) = \{(\alpha_i, \epsilon_{i1}, \dots, \epsilon_{ik}) \mid \alpha_i + \epsilon_{ij} = Y_{ij} - \theta_j, j = 1, \dots, k\}.$$

If  $B_i(\theta)$  is not a finite set of points, the random sieve method selects a set of points  $E_{ij} \in B_i(\theta)$ ,  $j = 1, \dots, n_i$ . For larger  $n$ , the numbers  $n_i$  increase to give better coverage of  $B_i$ . Consider a family of distributions, with the generic member  $F$  putting probability  $p_{ij} \geq 0$  on the observation  $(X_i, Y_i, E_{ij})$  for  $i = 1, \dots, n$  and  $j = 1, \dots, n_i$  with  $p_i = \sum_{j=1}^{n_i} p_{ij}$  and  $\sum_{i=1}^n p_i = 1$ . The sieved likelihood is the maximum of  $\prod_{i=1}^n p_i$  over members of the family satisfying a set of constraints. For random effects the constraints might express that  $E(\alpha_i) = 0$ ,  $\text{Var}(\alpha_i) = \sigma_1^2$ , and that for  $j = 1, \dots, k$ ,  $E(\epsilon_{ij}) = E(\epsilon_{ij} - \sigma_2^2) = E(\alpha_i \epsilon_{ij}) = 0$ . Methods for choosing the  $E_{ij}$  are in articles cited in Chapter 9.11.

## 9.11 Bibliographic notes

### *Parametric-empirical hybrids*

Qin (1994) presents the semi-empirical likelihood, in which an empirical likelihood is used for one sample and a parametric one is used for the other. He considers the example where  $X_i$  and  $Y_j$  are scalars and interest centers on  $E(X) - E(Y)$ . Qin (2000) considers multiplying a parametric conditional likelihood by an empirical marginal one, and presents a theorem from Qin (1992) establishing a  $\chi^2_{(1)}$  limit for a statistic defined through a smooth estimating function. Qin (2000) simulates some examples that Imbens & Lancaster (1994) treated by generalized method of moments, and remarks that the likelihood formulation makes it easier to incorporate observations where  $X_i$  but not  $Y_i$  was measured. Those examples consider samples augmented by side information from the census.

Qin & Wong (1996) present another semi-empirical likelihood in which the data are parametric or not, depending on their values. They consider the case where the parametric model holds if  $x \leq T_0$ , and establish a  $\chi^2$  calibration for a univariate  $\theta$  that enters the likelihood smoothly. Moeschberger & Klein (1985) consider a parametric model for a tail subject to censoring combined with a non-parametric model to the left of that tail.

Lazar (2000) studies the product of a prior density on the univariate mean and an empirical likelihood for that mean. She shows that the posterior distribution is asymptotically normal, as one would expect from an asymptotically quadratic log likelihood. As a result, we can expect in general that the computed posterior probabilities of intervals are asymptotically justified. The arguments parallel those of Monahan & Boos (1992) for parametric likelihoods. In simulations, accuracy can be measured by finding the distribution under sampling of the posterior probability attached to the set  $(-\infty, \mu]$  where  $\mu$  is the true mean. This posterior probability should have nearly a uniform distribution. Figure 1 of Lazar (2000) shows an example in which the accuracy is good for  $n = 50$  but perhaps not for  $n = 10$ . Figure 2 shows an example where the accuracy appears good with  $n = 20$ .

### *Bootstrap connections*

The Bayesian bootstrap was proposed by Rubin (1981). Newton & Raftery (1994) illustrate its use on a number of frequentist inference problems. The bootstrap likelihood was proposed by Davison, Hinkley & Worton (1992). Some computational improvements are given in Davison, Hinkley & Worton (1995) and Pawitan (2000).

Efron (1981) introduces the nonparametric/exponential tilting bootstrap. Further results for it appear in DiCiccio & Romano (1990), as described below under least favorable families. Efron emphasizes the Kullback-Liebler family but also presents a version using a nonparametric likelihood discrepancy, giving rise to the same family of distributions in the simplex that empirical likelihood uses.

Zhang (1999) resamples from an NPMLE constructed to incorporate side information in the form of estimating equations. He produces confidence bands and tests for the distribution function of a scalar random variable, and confidence intervals for the mean and variance of a scalar. Zhang (1997) constructs confidence bands for the quantile function by bootstrapping from an NPMLE. Hall & Presnell (1999*b*) term this the *b*-bootstrap, and they show its wide applicability and describe the asymptotic behavior.

Ren (2001) defines the leveraged bootstrap. The leveraged bootstrap takes IID samples of size  $m$  from an NPMLE. The bootstrap samples are IID, even though the original data was subject to interval censoring. Careful calibration makes up for the discrepancy.

Chuang & Lai (2000) describe a hybrid method in which they construct a parametric family on the simplex of observation weights and use the bootstrap in that family. They find good results this way for some hard inferential problems such as the analysis of group sequential trials, explosive time series, and Galton-Watson processes.

### *Least favorable families*

The notion of a least favorable family is due to Stein (1956). He used it to reduce nonparametric problems to parametric ones that were at least as hard. Efron (1981) shows a least favorable family property of the one-dimensional multinomial sub-family. When  $\theta$  is the mean, the Fisher information in the sub-family is the inverse of the sample variance, holding for  $F_{\bar{X}}$  and even for other members  $F_{\theta}$ .

DiCiccio & Romano (1990) show that one can construct a  $p$ -dimensional sub-family of multinomial distributions using any of various discrepancy measures. Once one has such a family, one can bootstrap within it, use the likelihood function in it, or use another distance function in it. They show that one-sided coverage errors are  $O(1/n)$  for the nonparametric tilting bootstrap. They also show that the reduced family is least favorable for more general statistics than the mean. Hesterberg (1999) investigates exponential tilting bootstrap confidence intervals for nonlinear statistics, using importance sampling. He compares likelihood- and entropy-based intervals.

Empirical likelihood corresponds to using the likelihood ratio function in the least favorable family defined by the likelihood discrepancy. Lee & Young (1999) recommend defining a  $p$ -dimensional parametric family through the exponential discrepancy, and following up with a likelihood ratio test in that family. Corcoran et al. (1995) recommend using the exponential discrepancy (empirical entropy) to form a  $p$ -dimensional family, and then using the Wald test (a sandwich estimator) within that family. They report some simulations in which a  $\chi^2$  calibration is quite accurate for this combination. Davison & Hinkley (1997, Chapter 10) recommend a Rao test within a family formed by exponential discrepancies.

Where Efron (1981) considers sampling from various members of a data-determined least favorable family, the  $b$ -bootstrap and related papers above pick a single best-fitting member of that family and resample from it.

### *Other connections*

The jackknife was introduced by Quenouille (1949) for correcting bias in estimates. Tukey (1958) showed that it can be used to construct variance estimates. The infinitesimal jackknife is due to Jaeckel (1972). Hesterberg (1995a) proposes a “butcher knife” formed by taking divided differences in (9.6) with  $\epsilon = O(n^{-1/2})$ . By varying the sample more than  $O(1/n)$  this results in a more widely applicable jackknife. See also Shao & Tu (1995) for a discussion of jackknives that delete  $d$  of  $n$  observations.

Saddlepoint approximations allow one to construct approximate sampling densities for statistics. Monti & Ronchetti (1993) provide a formula that allows one to translate between empirical likelihoods and saddlepoint densities for statistics defined through estimating equations. Reid (1988) provides a survey of saddlepoint methods.

The connection between empirical likelihood and random sieves is given by Shen, Shi & Wong (1999), who also consider a random effects model like the one in Chapter 9.10 but incorporating predictors  $X_{ij}$ . Shen et al. (1999) describe methods for selecting a finite set of points within each  $B_i(\theta)$ .

The results in [Theorem 9.1](#) are based on independent work by Fan & Gijbels (1999) and Kitamura et al. (2000). The combination of smoothing and empirical likelihood studied there was proposed by LeBlanc & Crowley (1995).

Fan, Zhang & Zhang (2001) describe a sieve-based approach to problems such as nonparametric regression, including testing whether a smooth function, or indeed an additive model, might be linear. They obtain more general Wilks’s type results with  $-r \log R$  approximately  $\chi^2_{\nu(n)}$  with non-integer numbers of degrees of freedom  $\nu(n) \rightarrow \infty$  as  $n \rightarrow \infty$  and  $r$  not necessarily equal to 2. Their formulation has a parametric conditional likelihood with a parameter  $\theta(x)$  estimated nonparametrically as a smooth function of  $x$ . Fan & Zhang (2000) make an extension to an empirical conditional likelihood for the nonparametric  $\theta(x)$ .

## 9.12 Exercises

**Exercise 9.1** Suppose that there is no parametric model for the marginal distribution of  $X$  but there is a parametric density or mass function  $f(y|x; \theta)$  for the conditional distribution of  $Y$  given  $X$ . Write a hybrid likelihood assuming that  $(X_i, Y_i)$  are observed for  $i = 1, \dots, n_1$ ,  $X_i$  only is observed for  $i = n_1 + 1, \dots, n_1 + n_2$ , and  $Y_i$  only is observed for  $i = n_1 + n_2 + 1, \dots, n_1 + n_2 + n_3$ . Assume that missingness of  $X_i$  or  $Y_i$  is non-informative in the sense of coarsening at random.

**Exercise 9.2** The standard 95% confidence interval for the univariate sample mean is  $\bar{X} \pm 1.96s$  where  $s^2 = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ . What is the ratio of the length of the leave-one-out jackknife version of the 95% confidence interval to the length of the standard 95% confidence interval?

**Exercise 9.3** The jackknife is known to fail for the median, and it seems to be because only a small number of  $T_{-i}$  values are possible for a given sample. But the jackknife provides a reliable confidence interval for  $\Pr(X \leq Q)$ , so long as  $0 < \Pr(X \leq Q) < 1$ . Describe how to employ the jackknife to construct a confidence set for the median. If necessary assume that the observations are IID from a distribution having a unique median and a positive density function on an open interval containing that median.

**Exercise 9.4** Derive a bias estimate based in the infinitesimal jackknife.